Evolution and diversity of hominid genomes

Peter H. Sudmant

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Evan E Eichler, Chair
Philip Green
Joshua M Akey

Program Authorized to Offer Degree:
Genome Sciences

©Copyright 2013
Peter H. Sudmant

University of Washington

Abstract

Evolution and diversity of hominid genomes

Peter H. Sudmant

Chair of the Supervisory Committee:
Professor Evan E. Eichler
Department of Genome Sciences

Copy number variants (CNVs) contribute to a substantial amount of diversity both between and among hominid species. In this thesis I set out to catalog this variation across the hominid lineage and assess its impact on the evolution of great ape species and its contribution to extant diversity. I develop methods to detect and genotype CNVs using high-throughput sequencing and demonstrate the first population scale analysis of the diversity of segmental duplications and duplicated gene families in a panel of 159 humans. I identify 56 highly population stratified genes and 23 genes that have undergone duplication specifically in the human lineage. I go on to demonstrate a resource of 4.1 million singly unique nucleotide (SUN) identifiers, which can be used to infer paralog-specific copy number and explore duplicated gene families. I next describe the sequencing of a panel of 79 wild and captive-born great apes to better understand the genetic diversity of extant great ape populations. I identify 86.5 million single nucleotide polymorphisms (SNPs) and use this resource to construct a model of great ape demography and population diversity. I outline the topology of relationships between great ape populations and the relative levels of heterozygosity among these populations. Finally, I identify 23,770 CNVs in these great ape individuals and characterize diversity both within and between species. Among these CNVs I identify the first example of a genomic disorder in a chimpanzee—a single individual with features resembling Smith-Magenis syndrome. I calculate the exact rates of duplication and deletion along each of the great ape lineages and show that an excess of deletions occurred along the chimpanzee-bonobo ancestral lineage. This excess coincides with a severe drop in the effective population size of the chimpanzee-bonobo ancestor. Analysis of recently bottlenecked great ape populations also shows an excess

burden of duplications and deletions among these individuals. These results suggest that changes in effective population size may be responsible for historical shifts in the rates of duplication and deletion in the hominid lineage.

TABLE OF CONTENTS

List of Figures 5
List of Tables 6
Acknowledgements 7
1 - Introduction 9
1.1 Naturalists of the genome 9
1.2 The *hominidae* family 9
1.3 Structure and sequence of the hominid genome 10
1.4 Great ape diversity and demography 12
1.5 Human origins and worldwide patterns of diversity 13
1.6 Archaic hominid genomics 15
1.7 Methods to assay diversity 17
1.8 Research goals 18
1.9 Chapter organization 19
2 - Diversity of Human Copy Number Variation and Multicopy Genes 21
2.1 Summary 21
2.2 Introduction 21
2.3 Results 22
2.3.1 CNV landscape diversity 24
2.3.2 Gene copy number diversity and evolution 26
2.3.3 A singly unique nucleotide (SUN) identifier map 29
2.3.4 Paralog-specific copy number 30
2.3.5 Paralog-specific gene family diversity 31
2.4 Discussion 33
2.5 Notes 34
3 - Great Ape Genetic Diversity and Population History 36
3.1 Summary 36
3.2 Results 37
3.2.1 Patterns of diversity 39
3.2.2 Evolutionary history and demography 41
3.2.3 Genes 44
3.3 Discussion 45
3.4 Methods 46
3.5 Notes 47
4 - Evolution and Diversity of Copy Number Variation in the Great Ape Lineage 48

4.1 Summary 48
4.2 Introduction 49
4.3 Results 50
4.3.1 Patterns and diversity 50
4.3.2 Genes 55
4.3.3 Rates and CNV load 57
4.3.4 A putative chimpanzee genomic disorder 59
4.4 Discussion 62
4.5 Methods 64
4.6 Notes 65
5 - Summary and Future Directions 66
5.1 Assessing hominid genome diversity: insights and challenges 66
5.2 Expression trajectories of duplicated genes 68
5.3 Further insights from ancient genomes 72
5.4 Exploring human diversity 76
References 78
Appendix A - Supplementary Material for Chapter 2 91
Appendix B - Supplementary Material for Chapter 3 197
Appendix C - Supplementary Material for Chapter 4 318

List of Figures

Figure 1.1. Diagram of the HSA 18 – PTR 17 pericentric inversion. 11
Figure 1.2. Human migrations out of Africa with approximate times. 14
Figure 1.3. Comparison between paired-end, read-depth and split-read signatures of deletions and duplications. 18
Figure 2.1. Landscape of human copy number variation. 23
Figure 2.2. Validation and application. 26
Figure 2.3. Human gene family copy number diversity and evolution. 28
Figure 2.4. Paralog-specific copy number resolution and genotyping. 30
Figure 2.5. Paralog-specific gene family copy number variation. 32
Figure 3.1. Samples, heterozygosity and genetic diversity. 39
Figure 3.2. Inferred population history. 42
Figure 3.3. PSMC analysis. 43
Figure 4.1. Duplication and deletion landscape. 52
Figure 4.2. Hominid deletion phylogeny. 54
Figure 4.3. Genic variation. 56
Figure 4.4. CNV rates and polymorphism. 58
Figure 4.5. A chimpanzee genomic disorder. 62
Figure 5.2.1. Transcript abundance versus copy number. 68
Figure 5.2.2. The fates of duplicated genes. 70
Figure 5.2.3. Expression of *GOLGA* paralogs. 71
Figure 5.3.1. *BOLA2* copy number in 2641 hominids. 72
Figure 5.3.2. PSMC of the human lineage. 74
Figure 5.4.1. The C-team sequencing project. 76

List of Tables

Table 1.1. Currently extant great ape species and subspecies……..10
Table 3.1. Genetic variation summary by species and subspecies……..38
Table 4.1. Summary of all fixed and segregating duplications and deletions identified in great apes……51
Table 4.2. Common clinical features of Smith-Magenis syndrome and related features of Western chimpanzee Susie-A with a corresponding 17p11.2 deletion……61
Table 5.3.1. Summary of the number of duplicated base pairs along individual lineages……73
Table 5.3.2. Lineage-specific segmental duplications along each of the terminal branches of the Human-Neanderthal-Denisova lineage……74

Acknowledgements

I am indebted to many wonderful teachers, mentors, friends, colleagues, and family members for their support, friendship, and collaboration throughout my life. I mention some of them here knowing that I have inevitably forgotten some and for that I apologize. First and foremost, I am eternally grateful to my family for their love and friendship, support and inspiration: my mom Caroline Heshedahl, my dad Walter Sudmant, my two fantastic brothers Andrew and Adam, and of course, our past and present ridiculous dogs, Lucy and Joni. My love of learning and exploring was inspired and encouraged by my elementary school teacher Mrs. Cook and my high school teachers Mr. Buium, Mrs. McCreight and Mr. Kamen among others, all of whom I am so pleased to have known and been able to learn from. I am thankful also to have had so many good friends to support me throughout university and since: Heidi Gumpert, Kate Prince and Tiffany Smith, in particular, I am so lucky to know you. Throughout my life my three best friends have been Graeme Ross, Mark Eugster and Martin Clark, and I always look forward to spending more time with them. In Seattle I've been lucky enough to meet another magnificent group of people: Aaron Brooks, Christie Galitsky, Sheila Teves and Andy Rivers and Pretzel. I am particularly thankful to have such wonderful classmates: Jacob Kitzman, Keisha Carlson, Melody Palmer, Leslie Emery and Jarrett Egertson; you have been immensely fun to do a PhD with and are all fantastic people to know. Throughout Genome Sciences a number of marvelous people have made my time in Seattle and beyond so happy, including my good friends Graham McVicker and Rene George, and of course, Caitlin Connelly. The lab as well has provided no shortage of inspirational people and good friends. I'd also like to recognize past and present postdocs—Katie Campbell, Santhosh Girirajan, Fereydoun Hormozdiari, Beth Dumont and Tomas Marques-Bonet— as well as fellow students Jeff Kidd, Andy Itsara, Xander Nuttle and, of course, my comrade through thick and thin, Nik Krumm. I would be remiss to not mention John Huddleston, Maika Malig, Carl Baker and Tonia Brown. Tonia has tirelessly helped so many of us with so much and still continues to be amazingly good natured after all we have put her through. And of

course, the Genome Sciences IT staff have earned a special place in the hearts of many of those in our department and I am no exception. My committee members Phil Green, Josh Akey, Jay Shendure and Walter Ruzzo have all been excellent mentors to me and I am thankful for their guidance and advice. Finally, I am tremendously grateful to have had such a supportive and caring mentor in Evan Eichler. I continuously marvel at Evan's passion and work ethic, which has in turn inspired me to work hard and pursue questions that excite and challenge me. I feel very lucky to have been able to work alongside Evan and his colleagues over the course of my PhD.

1 - Introduction

1.1 Naturalists of the genome

The naturalists Charles Darwin and Alfred Wallace are credited with developing the theory of natural selection and thus the foundations upon which the study of all genetics rests (Darwin and Wallace 1958). Each of these eminent scientists was inspired in turn by voyages to some of the richest havens of biodiversity on the planet—Darwin to the Galapagos Islands (Darwin 1909) and Wallace to Southeast Asia and the 'Malay Archipelago' (Wallace 1869). The separate travels of these two exposed each to a scale of floral and faunal variety that displayed spectacularly before them the apparent effects of evolution. In much the same way, more than a century later, the study of genetics can be thought of as an exercise in marveling at the diversity of the natural world and trying to link these observed phenotypes with the genotypes that confer them. Indeed, to some degree the plight of naturalists to catalogue the worlds many plants and animals closely parallels the task of sequencing and annotating genomes, cataloguing the molecular extent of natural diversity on the planet.

1.2 The *hominidae* family

This thesis focuses in particular on exploring and cataloguing the genetic diversity of humans and our closest relatives, chimpanzees, bonobos, gorillas and orangutans. Collectively these species make up the *hominidae* family, or the great apes, and consist of four major genera, *Homo, Pan, Gorilla* and *Pongo.* Among these four genera are seven living species of hominid, which can be additionally subdivided into 14 subpopulations (**Table 1.1**). The *hominidae* family additionally extends to all of the extinct immediate ancestors of great apes, such as Neanderthals and Denisovans.

Table 1.1. Currently extant great ape species and subspecies (Caldecott and Miles 2005).

Species	Subspecies	Common name
Homo sapiens	-	Human
Pan troglodytes	*troglodytes*	Central Chimpanzee
	schweinfurthii	Eastern Chimpanzee
	verus	Western Chimpanzee
	ellioti	Nigeria-Cameroon Chimpanzee
Pan paniscus	-	Bonobo
Gorilla gorilla	*gorilla*	Eastern Lowland Gorilla
	diehli	Cross River Gorilla
Gorilla beringei	*graueri*	Eastern Lowland Gorilla
	beringei	Mountain Gorilla
Pongo abelii	-	Sumatran Orangutan
Pongo pygmaeus	*pygmaeus*	Northwest Bornean Orangutan
	morio	Northeast Bornean Orangutan
	wurmbii	Central Bornean Orangutan

1.3 Structure and sequence of the hominid genome

Despite the great phenotypic diversity among great apes, the genomes of these species are far more similar than they are different. Early cytological comparisons among great ape metaphase chromosomes revealed strikingly few large chromosomal differences (Yunis and Prakash 1982). Amongst humans, chimpanzees and gorillas, 18 of 23 chromosomes were found to be cytologically essentially identical, while the remaining 5 showed only slight variation among species (Yunis and Prakash 1982). The only uniquely human large chromosomal variations are two pericentric inversions on chromosomes 1 and 18 (**Figure 1.1**) and the fusion of the ancestral 2A and 2B chromosomes (Yunis and Prakash 1982; Fan et al. 2002). Similarly, studies of the relationship between chimpanzee and human proteins have determined the average human polypeptide to be >99% identical to its chimpanzee ortholog (King and Wilson 1975), a striking observation that led to the hypothesis that mutations affecting regulatory sequence may have been more consequential in recent evolution. Targeted sequencing of neutral sequence determined the relative divergence between humans and our closest relatives, chimpanzees, to be similarly small, ~1.24% (Chen and Li 2001). Divergence between humans and gorillas is ~1.62% and between humans and orangutans is ~3.08% (Chen and Li 2001).

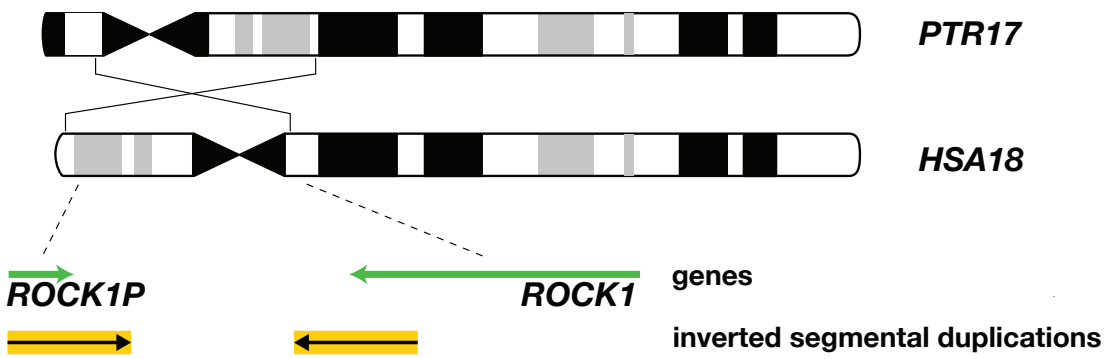


Figure 1.1. Diagram of the HSA 18 – PTR 17 pericentric inversion.

Accompanying the inversion, duplicative transposition of 19 kbp of sequence from the ancestral q-arm to the p-arm resulted in a pair of human-specific inverted segmental duplications overlapping the *ROCK1* gene (Goidts et al. 2005; Dennehey et al. 2004)**.**

A more complete picture of the structure and content of great ape genomes began to take form with the sequencing of the human genome (International Human Genome Sequencing Consortium 2001). In particular, the high-quality BAC-based assembly of the human genome (Cheung et al. 2001) helped reveal the presence and structure of ~150 mega base pairs (Mbp) of recently segmentally duplicated sequence >1 kbp long with >90% identity (Bailey and Eichler 2006). Structural variants, such as inversions, deletions and duplications, are enriched for flanking segmental duplications (Feuk et al. 2006; Sharp et al. 2006; Stefansson et al. 2005; Zody et al. 2008), which can mediate rearrangements through non-allelic homologous recombination (NAHR). As such, segmental duplication flanked sequences are hotspots for CNV-mediated diseases (Iafrate et al. 2004; Sharp et al. 2008; Kumar et al. 2008; Weiss et al. 2008; Stefansson et al. 2008).

Segmental duplications also associate with structural variants and large-scale rearrangements between great ape species (Fan et al. 2002; Yue et al. 2005; Kehrer-Sawatzki et al. 2005b; 2005a; 2005c; Locke et al. 2003a; Shimada et al. 2005; Goidts et al. 2005; Dennehey et al. 2004; Stankiewicz et al. 2004; Newman et al. 2005) underscoring their importance in shaping hominid genome evolution. Analysis of the chimpanzee draft genome (Chimpanzee Sequencing and Analysis Consortium 2005) revealed that segmental duplications contributed more sequence differences between humans and chimpanzees than SNPs, 2.7% versus 1.2%, respectively (Cheng et al. 2005). The advent of additional primate genome sequences, including

the rhesus macaque, which provided an out-group to the great ape lineage, orangutan and later gorilla (Locke et al. 2003b; Scally et al. 2012) revealed a striking sevenfold excess in the rate of segmental duplication along the African great ape lineage (Marques-Bonet et al. 2009; Ventura et al. 2011). More recently, it appears lineage-specific segmental duplication activity has acquiesced, though less so along the gorilla lineage, which appears to have accrued more segmental duplication than either the human or chimpanzee lineages (Ventura et al. 2011).

1.4 Great ape diversity and demography

The draft assemblies of the orangutan (Locke et al. 2011), gorilla (Scally et al. 2012) and bonobo (Prüfer et al. 2012) genomes, like the chimpanzee, (Chimpanzee Sequencing and Analysis Consortium 2005) were shotgun sequenced and are of much poorer quality than the human genome, precluding a comprehensive delineation of repetitive elements and structural features. These genomes, however, have offered numerous insights into primate evolution, speciation and diversity. The draft of the chimpanzee genome characterized 1.66 million SNPs and estimated a Western chimpanzee per-base-pair heterozygosity of $9.5X10^{-4}$ compared to a Central chimpanzee heterozygosity of $17.6X10^{-4}$, illustrating the striking differences in diversity among chimpanzee populations. Low-coverage sequencing of three Western lowland and an Eastern lowland gorillas as part of the gorilla genome project (Scally et al. 2012) demonstrated the low diversity of Eastern lowland populations ($7.6X10^{-4}$ hets/bp) compared to a relatively high heterozygosity ($17.8X10^{-4}$ - $18.9X10^{-4}$ hets/bp) in Western lowland gorillas. Five each Sumatran and Bornean orangutans revealed increased diversity among Sumatran orangutans compared to Bornean ($\theta_{W=}1.62$ and $\theta_{W=}1.21$, respectively) (Locke et al. 2011). The study of copy number variation and diversity among these nonhuman primate populations has been based largely on array comparative genomic hybridization (CGH). Though these studies have identified patterns of increased polymorphism among segmental duplications and both shared and lineage-specific CNVs, they generally have provided limited size resolution and do not have the ability to distinguish discrete copy numbers (Locke et al. 2003b; Gazave et al. 2011; Fortna et al. 2004; Perry et al. 2008; Dumas et al. 2007). Importantly, all of the aforementioned studies focused on captive great ape individuals and, thus, are

perhaps not representative of the true extent of wild great ape diversity. Additionally, sequencing experiments performed only a low to moderate amount of coverage per individual.

Despite the relatively small number of complete high-coverage primate genomes, however, demographic inferences point to a range of diverse population histories and mechanisms of speciation among the great apes sampled. The speciation of the two orangutan species and two gorilla species ~0.4 mya (million years ago) and ~0.5 mya, respectively, each were accompanied by low-level gene flow between the diverging populations. In the case of gorillas, the bidirectional exchange of genetic material is estimated to be ~0.2 individuals per generation (Scally et al. 2012). These patterns contrast strikingly with the history of bonobos, where little gene flow is observed after their divergence from chimpanzees ~2 mya, consistent with the formation of the Congo River ~1.5-2.5 mya, which would have formed a natural barrier to hybridization (Prüfer et al. 2012). Chimpanzee demography has been more difficult to reconstruct due to the lack of available sequence data from a diverse subset of individuals. However, using a collection of 73 loci from the literature Hey constructed a model of chimpanzee and bonobo demography including the Eastern, Western and Central chimpanzees (denoted here as B, E, W, C, respectively). These analyses suggested a (B,(W,(E,C))) relationship among populations with a complex history of gene flow (Hey 2010). Nigeria-Cameroon chimpanzees were only identified as a subspecies in 1997 (Gonder et al. 1997) and subsequently have been shown to be a genetically distinct population from both Western and Eastern/Central chimpanzee populations (Gonder et al. 2011).

1.5 Human origins and worldwide patterns of diversity

Human diversity and evolutionary origins have been the subject of much study, tracing the emergence of anatomically modern humans in Africa 150-200 kya (thousand years ago) and their dispersion across the world ~50 kya (Stoneking and Krause 2011) (**Figure 1.2**). This history is reflected in the diversity of our genomes (Stoneking and Krause 2011). Contemporary African populations, for instance, exhibit more genetic diversity than non-African populations, as a result of much of the sequence diversity of non-African populations being a subset of that found in

Africa. Many of the specifics of early human migrations and demographics, however, remain elusive (Stoneking and Krause 2011) despite great strides in cataloguing and genotyping worldwide genetic diversity (The International HapMap Consortium 2005; Frazer et al. 2007; Consortium et al. 2010; Li et al. 2008). More recent projects have utilized next-generation sequencing technologies to catalog diversity. The 1000 Genomes Project performed low-coverage sequencing of more than 1000 individuals from 14 diverse populations (1000 Genomes Project Consortium 2010; 1000 Genomes Project Consortium et al. 2012), capturing >98% of human SNPs at >1% allele frequency. Though the efforts of the 1000 Genomes Project have resulted in an important resource for the study of human populations, the low sequencing coverage of the vast majority of individuals is problematic for many analyses, especially those focused on low frequency variants.

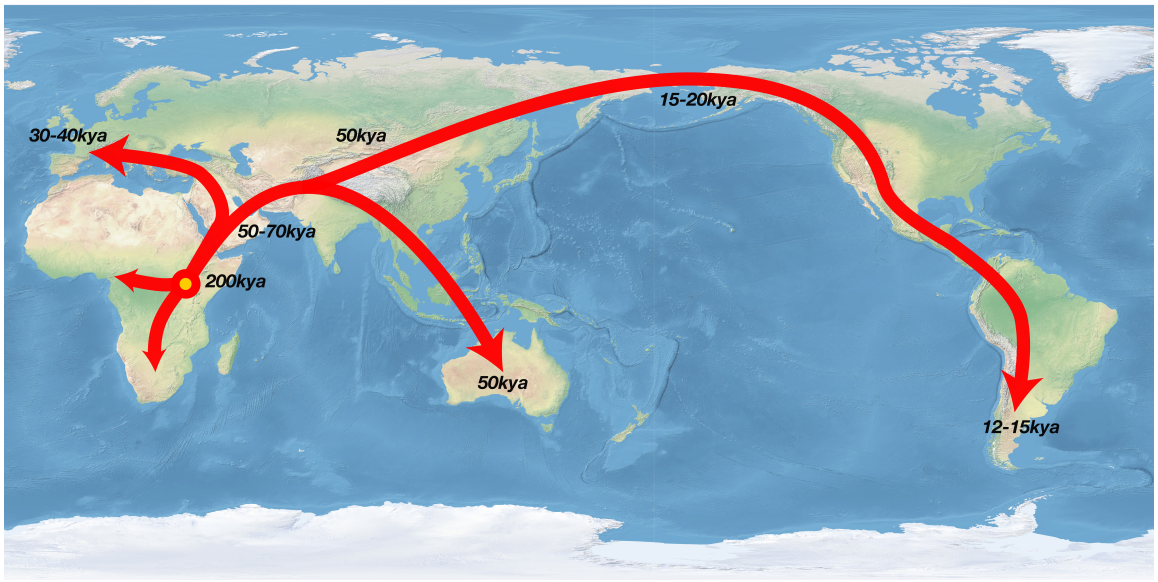


Figure 1.2. Human migrations out of Africa with approximate times.

Structural variants and CNVs are also challenging to identify from low-coverage sequencing, as multiple supporting reads are often required to indicate the presence of a variant in paired-end analyses and read-depth-based methods require increasingly higher coverage to detect smaller variants and overcome the effects of noise. Despite these challenges, using data from the initial 1000 Genomes Pilot (1000 Genomes Project Consortium 2010), Mills et al. identified 22,025 deletions and 6,000 insertions, tandem duplications, and additional structural variants (Mills et al. 2011). In Phase I of the project, targeting 1,092 individuals, 14,000 larger

deletions were computationally phased with SNPs and short indels (1000 Genomes Project Consortium et al. 2012). Previous to the era of full-genome next-generation sequencing, hybridization and clone-end sequencing-based methods each contributed substantially to the study of human copy number variation (Tuzun et al. 2005). Using both SNP microarrays and array CGH, Redon et al. identified 1,447 copy number variable loci, yet poor resolution led to the erroneous conclusion that 12% of the genome was copy number variable (Redon et al. 2006). This flaw was identified by Sanger-based fosmid end sequencing of eight individuals identifying with higher resolution 1,695 structural variants in addition to 525 novel insertions not represented in the reference genome (Kidd et al. 2008). Sequencing of a subset (n=261) of these variants provided the first large-scale analysis of breakpoints and CNV formation mechanisms. Targeted array CGH of the novel insertions identified by Kidd et al. revealed most to be copy number polymorphic with some showing stratification between populations (Campbell et al. 2011). A follow up study of clone-end sequencing demonstrated next-generation sequencing methods could be applied to end sequencing to reduce the cost of such a project (Korbel et al. 2007). Clone-based end sequencing methods are particularly powerful in their ability to identify novel sequence and to detect copy number neutral events such as inversions, though, in some cases, even the relatively long insert size of fosmids (~40 kbp) can be too short to detect inversions between large blocks of segmental duplication. More recently, utilizing high-density array CGH, Conrad et al. identified 11,700 CNVs down to a size threshold of 443 base pairs and genotyped a subset of these CNVs in a panel of 450 individuals (Conrad et al. 2010). In addition to providing more insights into mechanisms of structural variant formation, this resource has proven to be one of the most accurate and comprehensive genome-wide datasets of structural variation.

1.6 Archaic hominid genomics

Though Neanderthals were contemporary with early modern humans in Europe and Western Asia, they vanished ~30 kya. DNA sequence from archaic hominids, nevertheless, can provide a unique perspective into the speciation of humans and the potential molecular differences between our own species and our extinct relatives. Early attempts to sequence Neanderthal DNA were limited to

mitochondrial sequence (Briggs et al. 2009; Ovchinnikov et al. 2000) or targeted amplification of the nuclear genome (Lalueza-Fox et al. 2007; Krause et al. 2007). Shotgun- and cloning-based methods were able to sequence as much as 1 Mbp of Neanderthal DNA, although plagued by modern human contamination (Green et al. 2006; Noonan et al. 2006). These early studies revealed the sequence diversity of the mitochondrial genome to be outside of the range of human variation, yet were underpowered to assess the nuclear genome and thus could not rule out the possibility of introgression among humans and Neanderthals. A draft sequence of the Neanderthal genome derived from the bone material of three individuals (~5.3 Gbp, 1.9X sequence coverage), however, provided strong evidence favoring the presence of gene flow from Neanderthals into the ancestors of non-African modern humans previous to the divergence of Eurasian population groups from each other (Green et al. 2010). This draft sequence also revealed a partial catalog of coding changes between Neanderthals and humans in addition to a number of candidates of selection in early modern humans.

In addition to humans and Neanderthals, the fossil record suggests at least two other archaic hominids exited the African continent: *Homo erectus* ~2 mya and *Homo heidelbergensis* 0.3-0.5 mya (Lalueza-Fox and Gilbert 2011; Jobling et al. 2013). In 2010 Krause and colleagues strikingly identified a new unknown hominid specimen from the complete sequencing of the mitochondrial genome of a bone found in the Denisova cave in the Altai mountains of Russia (Krause et al. 2010). A draft genome of the Denisova sequenced to 1.9X (5.2 Gbp) revealed these ancient hominids were indeed distinct from humans and Neanderthals and had again interbred with early modern humans, contributing 4-6% of their genetic material to modern-day Melanesian populations (Reich et al. 2010). More recent advances in archaic DNA preparation have yielded a complete Denisova genome (~31X coverage) enabling the construction of an almost complete catalog of single nucleotide changes and many structural and copy number changes as well (Meyer et al. 2012).

1.7 Methods to assay diversity

Techniques to assess genome diversity depend on the scale and type of variation being targeted; for genome-wide analyses, two major classes of technology dominate: hybridization-based methods and sequencing-based methods.

Hybridization-based methods exploit the property of single-stranded DNA to hydrogen-bond to its complement. SNP microarrays are thus spotted with DNA probes targeted to known SNPs and either the relative binding to allele-specific probes or an extension reaction then indicate the presence of the A-allele or the B-allele or both (LaFramboise 2009). The drawback of this technology is that only a known panel of SNPs can be assayed. In addition to genotyping SNPs, the intensity of the hybridization compared to a reference panel can be used to detect CNVs (LaFramboise 2009; Alkan et al. 2011; McCarroll et al. 2006; 2008). Another commonly used hybridization technology is array CGH. In this method two differentially labeled samples are hybridized to an array and the ratio of binding is compared, revealing differences in copy number (Pinkel et al. 1998; Snijders et al. 2001; Alkan et al. 2011). As mentioned, while these techniques have been extremely powerful in delineating CNVs, they often suffer from size constraints and are poor at genotyping exact copy numbers. Additionally, array-based methodologies are ignorant of DNA sequence not present in the reference from which they were designed and cannot distinguish copy number neutral variants.

Sequencing-based methods include the process of sequencing and assembling entire genomes (International Human Genome Sequencing Consortium 2001), yet they more commonly use a reference genome as a scaffold upon which to map sequenced DNA and elucidate variation. Alignment of sequenced reads to the reference allows small variants to be identified from the presence of multiple supporting reads completely encompassing the variant. Larger structural variants are identified using three different computational strategies: read-depth, paired-end and split-read methods. Read-depth methods rely on the uniformity of reads sampled across the genome and identify regions with an excess or depletion of reads (Bailey et al. 2002; Alkan et al. 2009). Paired-end methodologies rely on the input library of DNA fragments to have a predictable size distribution. Both the orientation of and

distance between the sequenced ends of the fragment can then be used to infer the presence of a structural variant (Tuzun et al. 2005; Kidd et al. 2008; Korbel et al. 2007; Hormozdiari et al. 2009; 2010). Finally, split-read-based methods identify reads intersecting structural variant breakpoints. Each end of the read will map back to distinct positions in the reference (or into novel sequence) with the junction indicating the breakpoint (Ye et al. 2009; Karakoc et al. 2012) (**Figure 1.3**). In recent years, sequencing-based analyses of genome diversity have rapidly matured due to the availability and decreasing costs of next-generation sequencing technologies (Shendure and Ji 2008). While early technologies were limited to extremely short reads of ~35 bp (Bentley et al. 2008; Wang et al. 2008b), steady improvements have increased read-lengths to 150-200 bp (Quail et al. 2012). Average read-lengths in excess of 3000 bp, on average, can be obtained from newer technologies (Eid et al. 2009) allowing assembly of even complex regions of the genome; however, there remains room for improvements in the accuracy and scale of these methods.

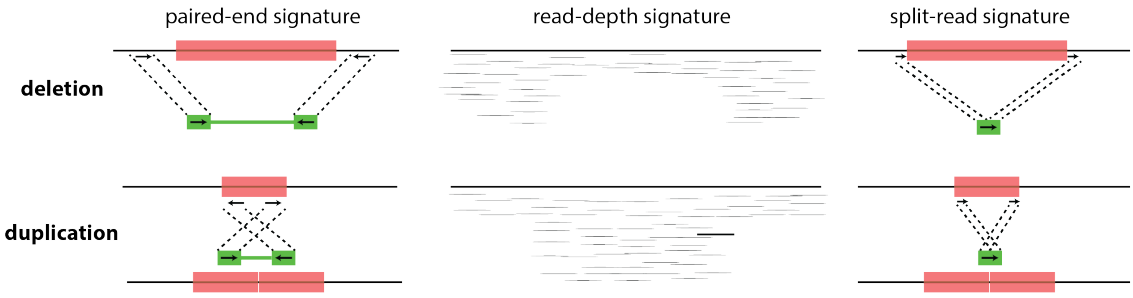


Figure 1.3. Comparison between paired-end, read-depth and split-read signatures of deletions and duplications.

1.8 Research goals

The goal of this dissertation is to develop methods to assess the diversity of segmental duplications and CNVs using next-generation sequencing technologies and then to explore the extent of variation in complex regions of the genome throughout hominid evolution. Methods will be developed specifically to assay the total copy number of duplicated loci and the paralog-specific copy number. These methods will then be used to assess human segmental duplication variation and duplication diversity among and between great ape lineages. The rates and scales of this form of diversity will then be compared to other estimates of diversity and

forms of variation to determine how segmental duplication activity has varied over the last 14 million years and how this has affected the genomes of hominids.

1.9 Chapter organization

In this thesis I outline methods to assess copy number variation in segmental duplications from next-generation sequencing data. I then go on to employ these methods to characterize duplication diversity among humans and nonhuman primates. Chapter two describes a method to genotype the total copy number of segmental duplications using high-throughput Illumina shotgun sequence and appling it to 159 human genomes. This chapter describes the method to be highly accurate and identifies duplicated sequence absent from the human genome. Chapter two then goes on to identify genes stratified in copy number between human populations and genes duplicated specifically in humans but not in nonhuman primates. This chapter also describes the construction of a resource of 4.1 million singly unique nucleotide (SUN) positions, which can be used to distinguish between copies and perform paralog-specific copy number genotyping. Finally, this chapter documents the usage of the SUN map to highlight copy number diversity in gene families and to identify signatures of non-allelic gene conversion in duplicated genes.

Chapter three focuses on an effort to characterize the genetic diversity of wild great ape populations from complete high-coverage genome sequencing. The sequencing of 79 wild and captive-born great apes is described along with the identification of ~88.3 million fixed substitutions and ~86.5 SNPs segregating in great ape populations. This data is then used to demonstrate the range in single nucleotide diversity between ape populations and the distinct clustering of individual populations. This chapter demonstrates that Western and Nigeria-Cameroon chimpanzees form a clade and Eastern and Central chimpanzees form a clade. Effective population sizes throughout history are estimated and used to describe the demographic history of great ape populations. Also described are 1,972 loss-of-function mutations and the rates of accumulation of these variants along the individual great ape lineages are found to be consistent.

Chapter four describes the identification of duplicated, deleted and copy number polymorphic regions among 97 deeply sequenced great apes from seven different species. This chapter identifies 340 fixed gene deletion events and 405 fixed gene duplications occurring along individual great ape lineages. Along the human lineage over the last 16 million years, 13.5 Mbp of sequence has been lost. The diversity of copy number variation is compared to the diversity of single nucleotide variation among the different lineages and is correlated (r^2=0.5). The phylogeny of the great apes is constructed from structural variants alone and the rates of duplication and deletion are computed as a function of the number of substitutions along each lineage demonstrating duplications outpace deletions by 2.8-fold. Recently, bottlenecked populations are shown to have a higher load of large CNVs segregating in their genomes. Similarly, this chapter describes a twofold increase in the rate of deletion along the chimpanzee-bonobo ancestral lineage corresponding to a sharp drop in the ancestral effective population size.

2 - Diversity of Human Copy Number Variation and Multicopy Genes

This chapter has been published: Sudmant PH*, Kitzman JO*, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J; 1000 Genomes Project, Eichler EE. *Science*. 2010 Oct 29;330(6004):641-6.
* These authors contributed equally to this work

2.1 Summary

Copy number variants (CNVs) affect both disease and normal phenotypic variation but those lying within heavily duplicated, highly identical sequence have been difficult to assay. By analyzing short-read mapping depth for 159 human genomes we demonstrate accurate estimation of absolute copy number for duplications as small as 1.9 kbp, ranging from 0-48 copies. We identified 4.1 million singly unique nucleotide (SUN) positions informative in distinguishing specific copies, and use them to genotype the copy and content of specific paralogs within highly duplicated gene families. These data identify human-specific expansions in genes associated with brain development, reveal extensive population genetic diversity, and detect signatures consistent with gene conversion in the human species. Our approach makes ~1000 genes accessible to genetic studies of disease association.

2.2 Introduction

Copy number variable genes in humans tend to map to duplicated sequences (aka segmental duplications) (Bailey et al. 2002; Iafrate et al. 2004; Conrad et al. 2010; Kidd et al. 2008), but neither the copy number nor the locus identity of these genes can be accurately assessed with existing hybridization-based experimental methods. This bias stems from the inability to discriminate subtle differences over the full range of copy number by array comparative genomic hybridization (CGH), and a lack of informative probes for these regions in single-nucleotide polymorphism (SNP) genotyping platforms (Cooper et al. 2008). This bias is more pronounced in cases where more copies of a duplicated gene are present, because multicopy integer states are difficult to resolve proportionally, resulting in a lack of understanding of the true extent of human copy number variation. In addition, the

most dynamic and variable genes are frequently excluded from genome-wide association studies (Consortium 2010). We have developed a method to survey whole-genome shotgun sequence data to accurately assay specific duplicated genes and gene families for copy number from data generated primarily from the 1000 Genomes Project.

2.3 Results

Read-depth can be used to accurately predict the copy number of duplicated genes within high-coverage human genomes (Alkan et al. 2009). We applied this method to perform a CNV analysis of 159 human genomes sequenced using the Illumina platform, including 15 high-coverage human genomes (12.3-43X) of diverse ethnicity and 141 low-coverage genomes (1.5-7X) from the 1000 Genomes Project representing three populations (see Appendix A). For each genome, reads were mapped to a repeat-masked human reference genome (Build 36) with the mrsFAST aligner, which returns all possible mapping locations of a read. Read-depth profiles were then constructed and corrected for biases introduced in library construction. Copy number prediction was performed by regression against a standard curve of regions of known copy. Using this approach, we estimated the absolute copy number genome-wide for windows of 3,000 non-repetitive bases each, and generated heatmaps displaying copy number for all 159 human genomes (**Figure 2.1A,B**).

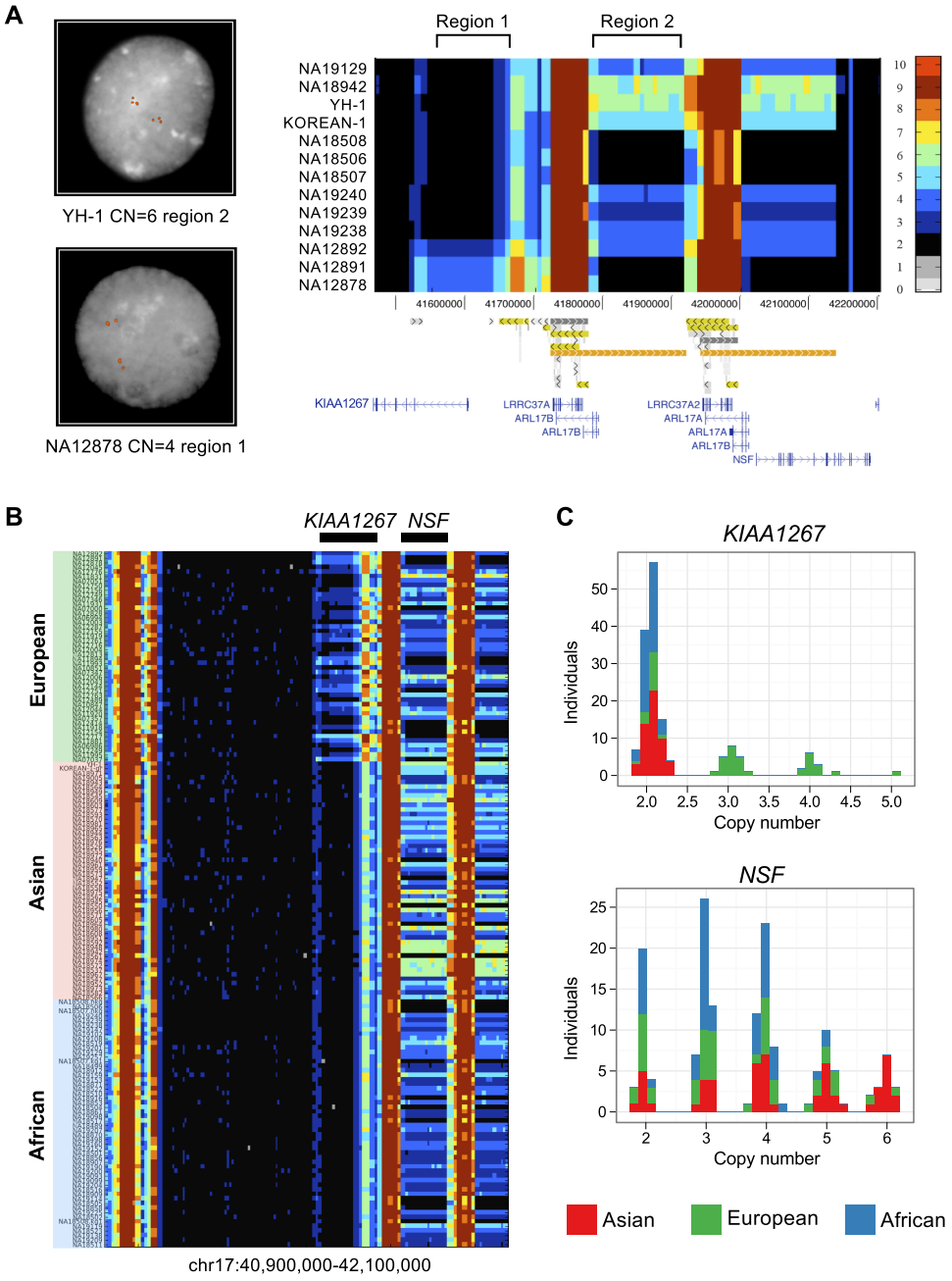


Figure 2.1. Landscape of human copy number variation.

(A) CNV heatmap of a 734 kbp duplicated region flanking the 17q21.31 MAPT locus in 13 individuals (11 sequenced to high coverage). Read-depth-based copy number (CN) estimations (3 kbp windows) are indicated by color (scale provided to the right). FISH at two separate loci validates these absolute CN predictions across five individuals (8). (B) Copy number landscape of the 17q21.31 locus across three different populations (n=159 genomes) showing marked population stratification. A European-enriched duplication overlaps the gene KIAA1267 and is present on two haplotypes—a long (205 kbp) and short form (155 kbp). A 210 kbp duplication of the NSF gene ranges from two to six copies with increased copy number in Asians. For validation with array CGH see Fig. S31. c) Copy number frequency histograms of the KIAA1267 and NSF duplications based on median read-depth predict discrete copies. Duplications of the KIAA1267 locus are specific to Europeans at a frequency of 72%. 25% of Asians have six copies of NSF.

Since many of these genomes (Table S1) were sequenced at low coverage, we tested the accuracy of our copy number predictions using three orthogonal methods of experimental validation (**Figure 2.1A**, Appendix A). We assessed our ability to detect simple events (gains, losses, and homozygous losses) from 2,270 events (median size 39.6kb, smallest event 1.6kb) across 109 common individuals. This demonstrated a concordance of 94-100%, dependent upon the size and type of the event (Figure S20). To assess our ability to genotype multicopy number states, we performed 59 fluorescent in situ hybridization (FISH) experiments across 21 larger loci ranging in copy from 2-48. 93% (55/59) of our copy number estimates were within +/-1 of the number of FISH signals observed in interphase nuclei (e.g., **Figure 2.1A**). Since FISH-based estimates are problematic when copy number exceeds 10, we designed a series of quantitative PCR (qPCR) assays to assess dynamic range response. These showed high correlation with our read-depth-based estimates, with 7/9 assays having r>0.92 (Appendix A). Read-depth- and qPCR-based copy number estimates for the 1.9 kbp gene CCL3L1 across 150 individuals were concordant (r=0.95), with both methods capturing the population-specific distributions of copy number at this locus (Figure S26).

2.3.1 CNV landscape diversity

We constructed genome-wide copy number maps across 159 genomes at 3 kbp resolution (**Figure 2.1 A, B**) to assay the full extent of large-scale copy number variation among human populations. We identified 952 large copy number variants greater than 50 kbp across 159 individuals. As expected, events of increasing size occur with progressively lower frequency. We noted that the majority of large events (55%, 522/952) overlapped segmental duplications. Out of all events, 47% (452/952) were common, being observed in more than eight individuals (>5% of genomes). Association with segmental duplications also strongly influenced CNV frequency. Events entirely outside segmental duplications were generally rare, with 71% (390/546) detected in three or fewer individuals (<2% of genomes). By contrast, 91% (461/506) of most CNVs overlapping segmental duplications were observed in more than three individuals, consistent with recurrent variation in these dynamic regions.

We identified 22 regions longer than 100 kbp that showed evidence of significant copy number differences between Asians, Europeans, and Africans (average Vst of

>0.2, Table S9) (Redon et al. 2006). These included duplications flanking the 17q21.31 MAPT locus, which has been associated with positive selection, rapid evolutionary turnover and neurological disease (Stefansson et al. 2005). Analyses of the region across populations revealed that although the inversion of this locus and its associated duplication are largely restricted to European-Mediterranean populations with 20% allele frequency, duplications overlapping the KIAA1267 gene were more common and definitive of populations of European origin (found in 33/46 individuals at 50% allele frequency) but were nonexistent in all other ethnic groups examined (**Figure 2.1 B,C**). Based on the extent of phased copy number change, we predicted and confirmed a short (155 kbp) duplication occurring at 20% allele frequency and a long version (205 kbp) at 30% allele frequency in Europeans. Distal to this locus, we identified a larger, more copy number polymorphic region (ranging from 2-6 copies). This more distal population stratified duplication encompassed the first 13 exons of the N-ethylmaleimide-sensitive (NSF) gene and shows increased copy number among Asians with the six-copy state occurring at ~25% frequency (13/54 individuals) (**Figure 2.1 B,C**). Interestingly, this gene is preferentially expressed in the human nervous system; reduction in its expression is associated with schizophrenia (Mirnics et al. 2000), and disruptions of its Drosophila ortholog lead to defective synaptic transmission (Pallanck et al. 1995).

Additionally, we identified regions that are duplicated in most human individuals but previously were incorrectly classified as diploid in the reference genome including 10 regions of >100 kbp each. We identified 173 segmentally duplicated regions for which the majority of genomes have a copy number greater than that of the reference genome. We thus have established a copy number baseline in humans, allowing more accurate genotyping from SNP microarrays. These data may be used in conjunction with single-channel intensity data from previous array CGH experiments to develop a calibration curve specifically for each region of the genome, allowing multi-allelic loci to be more robustly genotyped on array-based platforms (**Figure 2.2A**). This CNV landscape diversity map may also be used to select an ideal reference genome (from the 1000 Genomes catalog) to maximize discriminatory power in array CGH studies which will open more complex, copy number polymorphic regions of the genome to further experimental characterization.

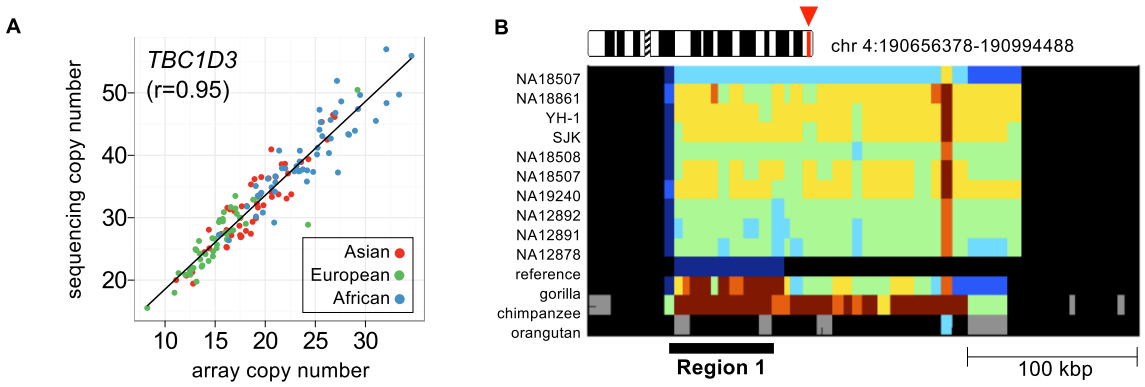


Figure 2.2. Validation and application.

(A) Single channel array CGH data is highly correlated (r=0.95) with read-depth-based genotypes for the highly duplicated TBC1D3 gene (copy number range 5-53). Note the reduced copy number of this gene family among Europeans. (B) Heatmap of a 340 kbp region proximal to the fascioscapulohumeral muscular dystrophy (FSHD) region on chromosome 4 identifies a polymorphic segmental duplication ranging from 5-8 copies. In the human reference genome (build 36) this segment is annotated as a single copy (i.e., unique) but all humans carry duplications mapping to chromosomes 4, 13, 14 and 21.

2.3.2 Gene copy number diversity and evolution

To assess the impact of copy number variation specifically on the coding portion of the genome, we genotyped the absolute copy of 25,832 individual RefSeq gene models (UCSC Genome Browser). As expected, our read-depth assessment predicts that 99.3% of human gene models outside of segmental duplications show a median copy number of two. Limiting our set to all genes greater than 10 kbp, we find that 91% of human genes are fixed as diploid in all 159 humans examined. Of the copy number variable gene families, including those within segmental duplications, we found that 80% vary between 0-5 copies, suggesting that extreme variation is limited to only a few gene families. We identified the 56 most variable gene families in humans (variance > 3.0 among combined populations) ranging in median copy from 5 to ~368 (Figure S33, Table S7). These genes were dramatically enriched for segmental duplications (OR=311.3, P<2.2x10-16, Fisher's Exact). In 19 of these cases, no individual exhibited a copy number less than or equal to that of the reference genome, suggesting that some of these gene duplicates are underrepresented in the reference. We found 44 'hidden' duplicated gene families (Figure S33, Table S6) including the rapidly evolving and high-copy ANKRD (~6 missing haploid copies), NBPF (>9 missing copies) and NPIP (~5 missing copies) gene families. The missing members of these gene families should be targeted for sequence finishing in order to more accurately capture the architecture and diversity of the human genome.

Since significant differences in allele frequency can be a signature of selection, we searched for genes with the most extreme differences in copy among the YRI, CEU and ASN populations using the Vst statistic (Redon et al. 2006) and identified 64 gene families (Vst>0.2; **Figure 2.3A**, Table S8). These genes mapped almost exclusively to segmental duplications (OR=72.0, P<2.2x10-16) and ranged in copy from 2-368 (77%, 49/64 with <12 copies). However, we observed no significant correlation between Vst and copy number (P=0.199) indicating that the differences between populations were not an artifactual product of higher copy numbers. In general, the African YRI showed greater variance in gene copy number compared to either Europeans or Asians (P<1x10-4 and P<1.2x10-9 after multiple-testing correction; Welch's one-tailed t), including genes with known stratification between populations, such as LILRA3 (Hirayasu et al. 2008) and UGTB17 (Xue et al. 2008). One of the most stratified genes identified was CCL3L1 (Figure S26), for which the importance of accurate multi-allelic copy number genotyping has been recently highlighted due to conflicting reports of association with HIV susceptibility (Urban et al. 2009). In addition, we discovered stratification among several previously uncharacterized gene families, many of which show reduced copy in Europeans and Asians compared to Africans (**Figure 2.3A**, Appendix A).

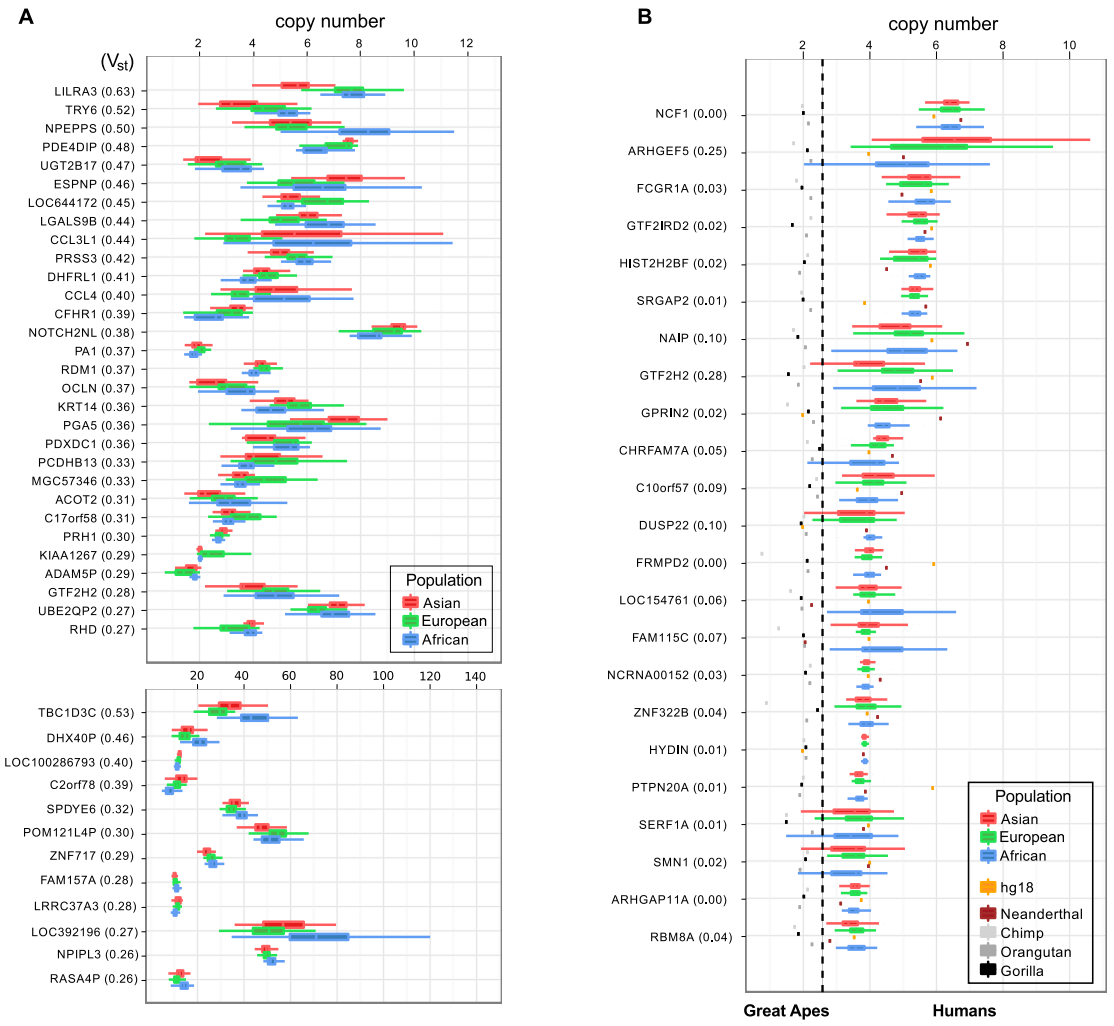


Figure 2.3. Human gene family copy number diversity and evolution.

(A) The most copy number stratified genes in the human genome on the basis of Vst analysis of European, African and Asian populations. (B) Human-specific gene family expansions.

We characterized the evolutionary context of human gene copy number variation by comparatively analyzing short-read-depth data from a gorilla, a chimpanzee and a Bornean orangutan. We identified 53 gene families with increased copy number within the human lineage, 23 of which were diploid in each of the great apes, and 8 of which appear to be fixed in humans (**Figure 2.3B**, Fig. S38; Tables S10-S11). Consistent with an origin after the human-ape divergence, the human-specific duplications and expansions displayed higher sequence identity than most duplicated genes (97.0% and 98.7%, respectively; P<3.5x10-5, Welch's one-tailed t). Human-specific duplications include the genes GPRIN2 and SRGAP2, which have been implicated in neurite outgrowth and branching (Chen 1999; Charrier et al. 2012); the brain-specific HYDIN2 gene, associated with micro and macrocephaly (Brunetti-Pierri et al. 2008); DRD5, a dopamine D5 receptor; the GTF2I

transcription factors whose deletion has been associated with visual-spatial and sociability deficits among Williams-Beuren syndrome patients (Dai et al. 2009; Edelmann et al. 2006); duplication of the SMN1 genes, at which copy increases ameliorate the severity of spinal muscular atrophy (SMA) deletions; and, duplication of the CHRNA7 locus on 15q13.2 recently implicated in cases of intellectual disability and epilepsy (Sharp et al. 2008; Shinawi et al. 2009). We note that the fixed duplications we identified at HYDIN and GPRIN2 are not annotated in the reference and thus represent 'hidden' duplications (Figure S40).

2.3.3 A singly unique nucleotide (SUN) identifier map

Although our read-depth-based genotyping provides absolute rather than relative copy number measurements and improved dynamic range relative to other platforms, it still lacks specificity within highly homologous segmental duplications. To discriminate among these paralogous sequences, we focused on the positions in the reference genome at which they diverge (**Figure 2.4A**). We identified 4.07x106 singly unique nucleotide (SUN) positions within high-identity duplicated regions (Table S2). SUNs are a distinct class of paralogous sequence variants that uniquely tag a specific paralog. By definition, the density of these markers diminishes with increasing duplication identity, yet we estimated that ~70% of duplications contain sufficient SUN density to allow paralog-specific genotyping. Nearby SNPs or rare variants could disrupt short-read mapping at SUN positions, confounding copy number estimation of the tagged paralog. We examined the extent of such effects among SUNs in 12 unrelated individuals sequenced to high coverage (mean=23.4X). Reads from these individuals were mapped to SUN positions, stringently requiring perfect 36 bp matches so as to exclude contamination between paralogous sequences. Essentially all autosomal SUNs (99.7%) were present in at least one individual with most (84.8%) present in all 12 individuals. Among these, we noted a strong enrichment for SNPs (274,245 SNPs from dbSNP v130, OR=1.6, $P<2.2\times10^{-16}$), suggesting that some SUNs may be misannotated as allelic variants (Bailey et al. 2002; Estivill et al. 2002).

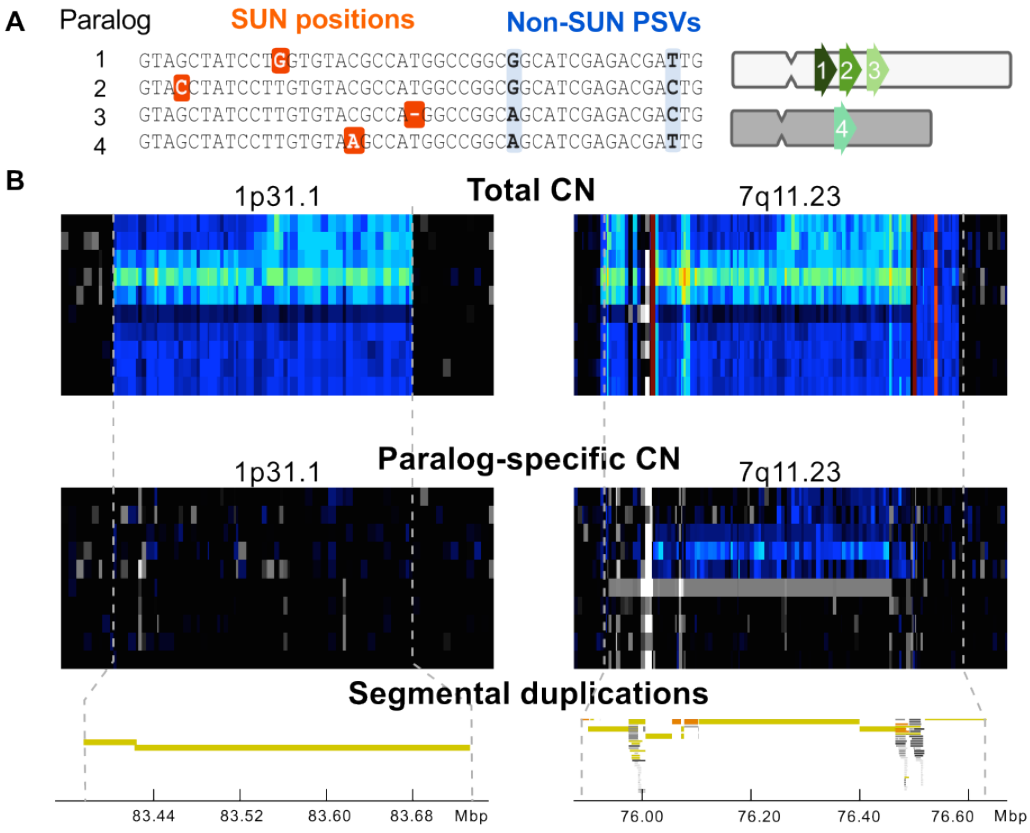


Figure 2.4. Paralog-specific copy number resolution and genotyping.

(A) Schematic showing singly unique nucleotide (SUN) identifiers among four high-identity duplications. SUNs (orange) uniquely distinguish one duplicated copy from all others, in contrast to paralogous sequence variants (PSVs, blue), which may be shared among copies. (B) Resolving duplication 'mirror' effects with paralog-specific genotyping. Total read-depth and array CGH fail to distinguish the origin of copy number variation between two high identity (98.5%) segmental duplications mapping to chromosome 1p13.1 and 7q11.23. SUN read-depth mapping, however, predicts that copy number variation is restricted to 7q11.23 and not 1p13.1. FISH on these samples confirms copy number gains and losses and their localization to 7q (Fig. S51).

2.3.4 Paralog-specific copy number

We utilized our SUN database to complement our estimates of total copy number by developing genome-wide maps of paralog-specific copy number (psCN). At locations uniquely identified by an overlapping SUN, we counted the number of reads perfectly matching and, as before, inferred copy number using a linear model trained on regions of known copy. We validated the accuracy of our psCN estimates by specifically analyzing 383 high-confidence deletion intervals that had been detected and fully resolved by capillary sequencing on the same individuals (Kidd et al. 2008). We found that 99.4% (308/310) of deletions within unique regions and 93.2% (68/73) of deletions within duplicated regions were accurately predicted, underscoring the specificity of this approach.

Paralog-specific copy number genotyping revealed CNVs within duplicated gene families. For example, at the complement factor H (CFH) locus, we detected deletions in close agreement with their known boundaries (+/- ~6 kbp on average, Figure S60) (Kidd et al. 2008). Genotyping the resulting psCN at these intervals across all 159 samples reveals overall deletion allele frequencies for CFHR3/1 and CFHR1/4 of 29% and 4%, respectively, with elevated frequency among Africans at both deletions (Hageman et al. 2006). We also identify rare genotypes, such as a reciprocal, single copy amplification of CFHR1/4 in a single African individual (ABT) and a deletion with novel breakpoints in a single Asian individual (NA18563).

Copy number analysis from read mapping depth can potentially lead to misassignment of CNV events to the wrong copy due to cross-mapping between highly identical sequences. Using SUNs to resolve paralogous copies, we regenotyped 406 large (>50 kbp) CNV events overlapping segmental duplications, previously called on the basis of total read mapping depth comparisons. We find that 60% (245/406) of these regions show no signature of variation using SUNs and hence represent variation at one of the homologous loci. This 'mirror' effect arises from cross-mapping between highly identical sequences but can be resolved by leveraging the SUN tags (**Figure 2.4B**). Similarly, we found that a recently reported set of copy number variants identified by arrayCGH (Conrad et al. 2010) contained a significant fraction of calls (1547/4412, 35.1%) within duplicated sequences that show no evidence of paralog-specific CNV.

2.3.5 Paralog-specific gene family diversity

A prerequisite to understanding the function of highly duplicated gene families is the ability to genotype them for both their copy and content. Using our psCN approach, we reassessed 990 human genes partially or completely contained within segmental duplications (**Figure 2.5A**). This analysis allowed us to distinguish two distinct classes of paralogs: 49.2% of duplicate genes which appear largely copy-invariant within the human species, and the remainder, which show extensive variation in copy number with some being biased toward gain or loss.

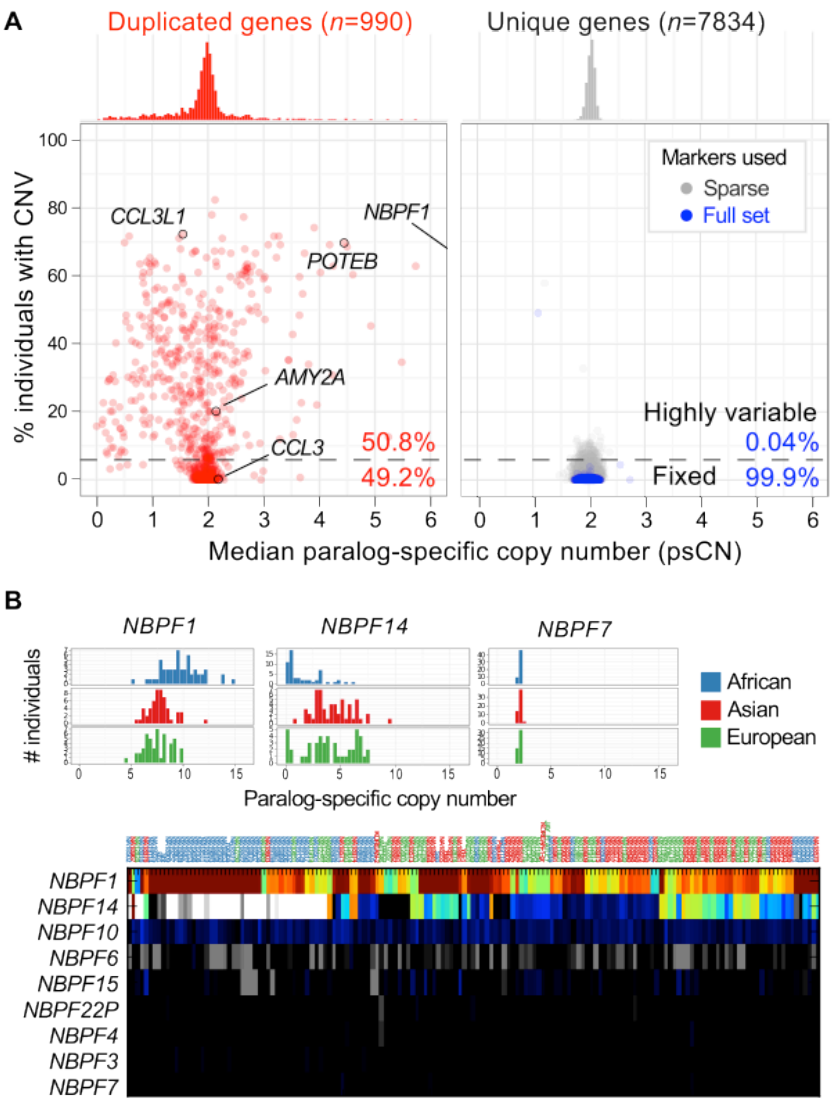


Figure 2.5. Paralog-specific gene family copy number variation.

(A) Paralog-specific copy number estimates of 990 duplicated genes shows that most are on average diploid within the human species (median psCN = 2 +/- 0.5), and nearly half show little variation in copy. 49.2% of genes deviate from the median copy only rarely (<=5% of individuals). By contrast, genes outside of segmental duplications and other known regions of copy number variation are nearly devoid of common CNV (blue), even when genotyping with randomly subsampled positions (gray) to mimic the restricted density of SUN markers within duplicated genes. (B). Population stratification and paralog-specific copy variability of a human expanded gene family of unknown function, NBPF (neuroblastoma breakpoint gene family). Certain paralogs (e.g., NBPF1) are highly amplified, extremely variable, and population stratified while others are nearly fixed and diploid (e.g., NBPF7).

A particularly peculiar set of variable genes are those mapping to 'core duplicons' (Jiang et al. 2007) which have undergone recent bursts of expansion within the human lineage, including NPIP, NBPF (**Figure 2.5B**), and others (Marques-Bonet et al. 2009). We confirmed these genes' human-specific expansion but also observed them to be stratified among human populations (Figure S38 and **Figure 2.3A**). We also observed more localized psCN patterns, with small patches showing reciprocal

gain and loss but remaining unchanged in total copy, a signature consistent with interlocus gene conversion (i.e., unidirectional transfer of genetic information between duplicate copies) [reviewed in (Innan 2003)]. We validated a small number of these events and confirmed this signature at the Rh blood group antigen genes RHD and RHCE (Figure S71, Appendix A) (Innan 2003). A scoring metric found 78 regions with a score exceeding that of RHD/RHCE with a significant excess of this signature relative to shuffled controls (P<2.2x10-16, one-sided Kolmogorov-Smirnov). Consistent with known examples and mechanisms of gene conversion, this was nearly exclusive to high-identity duplications (>95% ID) and was preferential to nearby tandem duplications (<=1 Mbp).

2.4 Discussion

We have leveraged next-generation sequence data to explore some of the most complex genetic variation in the human species and show that we can reliably predict absolute copy number without bias. More importantly, this data allowed us to assess the copy and content of specific duplicated genes. Nevertheless, several limitations remain. Some genes (~30%) show too little paralogous variation or are too small to be reliably genotyped; in other cases the paucity of markers combined with the low sequence coverage lead to uncertainty in copy number accuracy (our analysis suggests that at least 8X genome-wide sequence coverage is required to achieve copy number accuracy above 97%). Finally, we found 28 large regions with such extraordinary complexity that it is difficult to interpret the underlying pattern of genetic variation (e.g., Figure S34).

Through our analysis we identified that duplicated regions are more likely to be stratified between human populations when compared to copy number variation within unique regions of the genome. For example, 59 (92%) of the top 64 stratified gene families overlap segmental duplications (P<2.2x10-16). Remarkably, many of these highly polymorphic genes map to duplications that promote recurrent rearrangements associated with intellectual disability, autism, schizophrenia and epilepsy. We hypothesize that the extreme polymorphism may contribute to genomic instability associated with disease and predispose certain populations to different chromosomal rearrangements (Antonacci et al. 2009).

We have also defined the ~49% of gene duplicates that are largely invariant in copy among humans. Although this is based only on an assessment of 159 genomes from select populations, the fact that this fraction of genes remain copy number invariant in a milieu of recurrent unequal crossover suggests functional importance. Interestingly, among these we find a number of genes involved in neurological development and disease. We note that many of these duplicated genes are themselves incomplete and may represent non-processed pseudogenes which may modulate the expression of the ancestral gene. The characterization of the most recently duplicated genes should facilitate identification of those that acquired new functions (neofunctionalization) versus those that have become pseudogenes or have partitioned their function among duplicate copies (Force et al. 1999).

The ability to distinguish where copy number variation maps and where it does not within regions of high sequence identity facilitates breakpoint delineation within duplications, allowing us to refine the intervals associated with structural variants such as the recurrent, schizophrenia-associated deletion on 15q11.2 (Fig. S70) (Stefansson et al. 2008). We anticipate that overlaying our SUN map with functional genomics data (e.g., ChIP-seq, RNA-seq) may begin to resolve the epigenetic and expression landscape of duplicated regions and the ~1000 genes therein that have been largely inaccessible to genetic study of disease (Manolio et al. 2009).

2.5 Notes

Acknowledgements

We thank M. Ventura and M. Ross for sharing unpublished data, C. Lee for technical assistance, C. Campbell, J. Kidd, P. Green, and M. Hoopman for discussion, and T. Brown for manuscript preparation assistance. This work was supported by an NSERC Fellowship (P.H.S.), an NSF Graduate Research Fellowship (J.O.K.), and NIH grant HG004120 to E.E.E. E.E.E. is an Investigator of the Howard Hughes Medical Institute. E.E.E is on the scientific advisory board for Pacific Biosciences. A.T., N.S., and L.B. are employees of Agilent Technologies.

Data resources

Sequence and array data are deposited at the NCBI under accessions SRP002878, SRP003500, and GSE24334.

Additional methods

A complete description of the material and methods is provided in the Supplementary Note, which is provided as Appendix A to this document.

3 - Great Ape Genetic Diversity and Population History

This chapter has been published: Prado-Martinez J*, Sudmant PH*, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. *Nature*. 2013 Jul 25;499(7459):471-5. Epub 2013 Jul 3.
* These authors contributed equally to this work

3.1 Summary

Most great ape genetic variation remains uncharacterized (Gonder et al. 1997; 2011); however, its study is critical for understanding population history (Hobolth et al. 2007; Hey 2010; Becquet and Przeworski 2007), recombination (Coop and Przeworski 2006), selection (Bustamante et al. 2005), and susceptibility to disease (Hahn et al. 2000; Keele et al. 2006). Here, we sequence to high coverage a total of 79 wild- and captive-born individuals representing all six great ape species and seven subspecies and report ~86.5 million single nucleotide polymorphisms. Our analysis provides support for genetically distinct populations within each species, novel signals of gene flow, and the split of common chimpanzees into two distinct groups: Nigeria-Cameroon/Western and Central/Eastern populations. We find extensive inbreeding in almost all wild populations with Eastern gorillas being the most extreme. Inferred effective population sizes have varied radically over time in different lineages and this appears to have a profound effect on the genetic diversity at or close to genes in almost all species. We comprehensively discover and assign 1,982 loss-of-function variants throughout the human and great ape lineages, determining that the rate of gene loss has not been different in the human branch compared to other internal branches in the great ape phylogeny. This comprehensive catalog of great ape genome diversity provides a framework for understanding evolution and a resource for more effective management of wild and captive great ape populations.

3.2 Results

We sequenced great ape genomes to a mean of 25-fold coverage per individual (Table 3.1, Appendix B, Table S1) sampling natural diversity by selecting captive individuals of known wild-born origin as well as individuals from protected areas in Africa (**Figure 3.1a**). We also included nine human genomes—three African and six non-African individuals (Meyer et al. 2012). Variants were called using the software package GATK (DePristo et al. 2011) (Methods), applying several quality filters, including conservative allele balance filters, and requiring that genomes showed <2% contamination between samples (Methods and Appendix B). In order to assess the quality of single nucleotide variant (SNV) calls, we performed three sets of independent validation experiments with concordance rates ranging from 86%-99% depending on allele frequency, the great ape population analyzed, and the species reference genome used (Appendix B, Table S2). In total, we discovered ~88.3 million fixed substitutions and ~86.5 million segregating sites of high quality (Table 3.1, Table S3) providing the most comprehensive catalog of great ape genetic diversity to date. From these variants we also constructed a list of potentially ancestry-informative markers (AIMs) for each of the surveyed populations, although a larger sampling of some subspecies is still required (Appendix B).

Genus	*Scientific name Species/subspecies*	*Common name*	*N*	*Mean Coverage*	*Fixed Sites to Human reference*	*No. of SNVs*[a]	*Mean SNVs per Individual*[a]	*No. of Singletons*[b]	*Ancestry Informative Markers (AIMs)*[c]	*Ne (Kyrs)*[d]
Homo		Non-African	6	18.3	386,974	5,887,443	2,639,546	1,379,448	12,316	9.7 - 19.5
		African	3	20.9	632,253	6,309,453	3,203,178	2,448,454	12,316	13.9 - 27.9
	Homo sapiens	Humans	9	19.2	224,660	9,172,573	3,061,604	3,827,902	-	13.1 - 16.2
Pan	*Pan troglodytes ellioti*	Nigerian-Cameroon	10	16.7	25,017,403	12,605,585	4,816,435	2,695,109	2,213	18.5 - 37.0
	Pan troglodytes schweinfurthii	Eastern	6	28.7	25,126,506	11,264,879	4,843,530	2,228,396	1,265	19.7 - 39.5
	Pan troglodytes troglodytes	Central	4	23.8	25,080,750	11,820,858	4,983,933	3,948,347	619	24.4 - 48.7
	Pan troglodytes verus	Western*	4	27.3	26,832,247	4,729,933	2,411,501	1,481,079	145,548	9.8 - 19.5
	Pan troglodytes	Common Chimpanzees	24	22.5	24,087,088	27,153,659	5,693,903	10,352,931	149,645	30.9 - 61.8
Pan	*Pan paniscus*	Bonobos	13	27.5	27,068,299	8,950,002	2,738,755	3,159,889	-	11.9 - 23.8
Gorilla	*Gorilla beringei graueri*	Eastern lowland	3	22.8	34,537,496	3,866,117	2,578,328	484,482	317,028	12.2 - 24.3
	Gorilla gorilla diehli	Cross river	1	17.6	35,553,861	2,585,360	2,585,360	165,482	35,693	14.9 - 29.8
	Gorilla gorilla gorilla	Western lowland	23	17.8	31,602,620	17,314,403	6,410,662	2,797,388	19,902	26.8 - 53.5
		Gorillas	27	18.3	31,376,203	19,177,989	6,492,831	3,447,352	372,623	28.4 - 56.9
Pongo	*Pongo abellii*	Sumatran	5	28.7	62,880,923	14,543,573	7,263,256	5,681,303	1,132,808	27.5 - 55.0
	Pongo pygmaeus	Bornean	5	25.8	64,249,235	10,321,213	5,763,354	3,555,596	1,132,808	19.5 - 39.0
		Orangutans	10	27.3	60,661,869	24,309,920	9,338,148	6,409,648	-	42.3 - 84.6
		All	83	23	83,954,672	83,580,213	-	-	-	-

Table 3.1 | Genetic variation summary by species and subspecies. *Summary statistics for each species and subspecies.*

a.*Polymorphic variants found in each species/subspecies after substracting fixed sites.*

b.*Singletons and doubletons calculated combining all the samples within the species.*

c.*Variants only found in a single group within each species.*

d.*Calculated from Θw. μ = 1e-9 - 0.5e-9 mut·bp-1·yr-1 and g = 25 for Homo and Pan, 19 for Gorilla and 26 for Pongo.*

***Hybrid sample Donald and 4 related gorillas were excluded.*

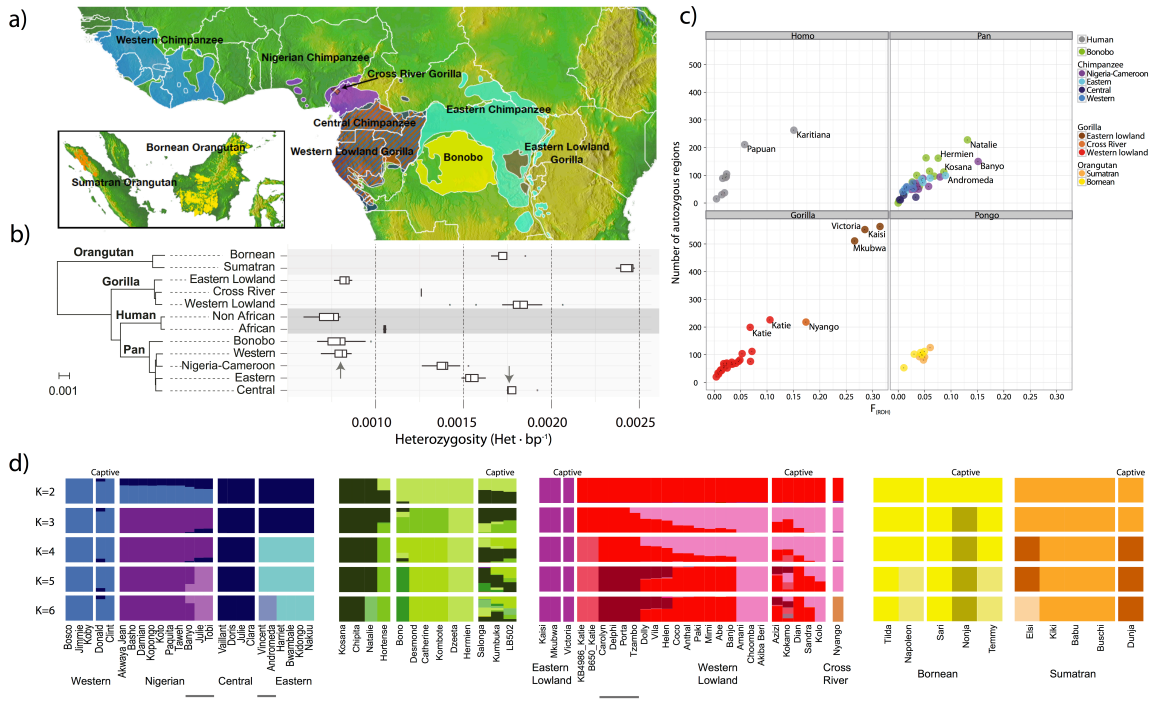


Figure 3.1. Samples, heterozygosity and genetic diversity.

a. Geographical distribution of great ape populations across Indonesia and Africa sequenced in this study. The formation of the islands of Borneo and Sumatra resulted in the speciation of the two corresponding orangutan populations. The Sanaga River forms a natural boundary between Nigeria-Cameroon and Central chimpanzee populations while the Congo River separates the bonobo population from the Central and Eastern chimpanzees. Eastern lowland and Western lowland gorillas are both separated by a large geographical distance. b. Heterozygosity estimates of each of the individual species and subspecies are superimposed onto a neighbor-joining tree from genome-wide genetic distance estimates. Arrows indicate heterozygosities previously reported30 for Western and Central chimpanzee populations c. Runs of homozygosity among great apes. The relationship between the coefficient of inbreeding (FROH) and the number of autozygous >1 Mbp segments is shown. Bonobos and Eastern lowland gorillas show an excess of inbreeding compared to the other great apes, suggesting small population sizes or a fragmented population. d. Genetic structure based on clustering of great apes. All individuals (columns) are grouped into different clusters (K=2 to K=6, rows) colored by species and according to their common genetic structure. Most captive individuals, labeled on top, show a complex admixture from different wild populations. A signature of admixture, for example, is clearly observed in the known hybrid Donald, a second-generation captive where we predict 15% admixture of Central chimpanzee on a Western background consistent with its pedigree. A gray line at the bottom denotes new groups at K=6 in agreement with the location of origin or ancestral admixture.

3.2.1 Patterns of diversity

We initially explored the genetic relationships between individuals by constructing neighbor-joining phylogenetic trees from both autosomal and mitochondrial genomes (Appendix B). The autosomal tree identified separate monophyletic groupings for each species/subspecies designation (Suppl. Figure 8.5.1) and supports a split of extant chimpanzees into two groups. Nigeria-Cameroon and Western chimpanzees form a monophyletic clade (>97% of all autosomal trees) while Central and Eastern chimpanzees form a second group (72% of all autosomal trees).

Genome-wide patterns of heterozygosity (**Figure 3.1b**) reveal a threefold range in single nucleotide polymorphism (SNP) diversity. Non-African humans, Eastern lowland gorillas, bonobos, and Western chimpanzees show the lowest genetic diversity (~0.8 x 10-3 heterozygotes/bp). In contrast, Central chimpanzees, Western lowland gorillas, and both orangutan species show the greatest (1.6-2.4 x 10-3 heterozygotes/bp). These differences are also reflected by measures of inbreeding from runs of homozygosity (Pemberton et al. 2012) (**Figure 3.1c**, Appendix B). Bonobos and Western lowland gorillas, for example, have similar distributions of tracts of homozygosity as human populations that have experienced strong genetic bottlenecks (Karitiana and Papuan). Eastern lowland gorillas appear to represent the most inbred population, with evidence that they have been subjected to both recent and ancient inbreeding.

To examine the level of genetic differentiation between individuals we performed a principal component analysis (PCA) of SNP genotypes (Appendix B). Chimpanzees were stratified between subspecies with PC1 separating Western and Nigeria-Cameroon chimpanzees from the Eastern and Central chimpanzees and PC2 separating Western and Nigeria-Cameroon chimpanzees. In gorillas, PC1 clearly separates Eastern and Western gorillas while the Western lowland gorillas are distributed along a gradient of PC2, with individuals from the Congo and Western Cameroon positioning in opposite directions along the axis. The isolated Cross River gorilla is genetically more similar to Cameroon Western lowland gorillas and can be clearly differentiated with PC3 (Suppl. Figure 8.2.9).

We explored the level of shared ancestry among individuals within each group (Tang et al. 2005) using an admixture model (FRAPPE). In chimpanzees, the four known subspecies are clearly distinguished when fitting the model using four ancestry components (K=4) (**Figure 3.1d**). Additional substructure is identified among the Eastern chimpanzees Vincent and Andromeda (K=6), who hail from the most Eastern sample site (Gombe National Park, Tanzania). As in Gonder et al. (Gonder et al. 2011) we have identified three Nigeria-Cameroon samples (Julie, Tobi and Banyo, K=3-5) with components of Central chimpanzee ancestry. However, taking Central chimpanzees and the remaining Nigeria-Cameroon chimpanzees as ancestral populations shows no evidence of gene flow by either the F3 statistic or

HapMix. This suggests these three samples are not the result of a recent admixture and may represent a genetically distinct population (Appendix B).

In gorillas, following the separation of Eastern and Western lowland species (K=2), an increasing number of components further subdivide Western lowland populations distinguishing Congolese and Cameroonian gorillas—a pattern consistent with the structure observed in the PCA analysis (Suppl. Figure 8.2.9). One striking observation is the extent of admixed ancestry predicted for captive individuals when compared to wild-born. Our analysis suggests that most captive individuals included in this study are admixed from two or more genetically distinct wild-born populations leading to an erosion of phylogeographic signal. This finding is consistent with microsatellite analyses of captive gorillas (Nsubuga et al. 2009) and the fact that great ape breeding programs have not been managed at the subspecies level.

3.2.2 Evolutionary history and demography

As great apes have been evolving on separate lineages since the middle Miocene, we attempted to reconstruct the history of these various species and subspecies by applying methods sensitive to branching processes, changes in effective population size (Ne), and gene flow occurring at different time scales. Using a combination of speciation times inferred from a haploid pairwise sequential Markovian coalescent (PSMC) analysis (Li and Durbin 2011), a coalescent hidden Markov model (CoalHMM) (Hobolth et al. 2007), and incomplete lineage sorting approaches, we were able to estimate the most ancient split times and effective population sizes among the great ape species. By combining these estimates with an approximate Bayesian computation (ABC) (Beaumont et al. 2002) analysis applied to the more complex chimpanzee phylogeny, we constructed a composite model of great ape population history over the last ~15 million years (**Figure 3.2**). This model presents a complete overview of great ape divergence and speciation events in the context of historical effective population sizes.

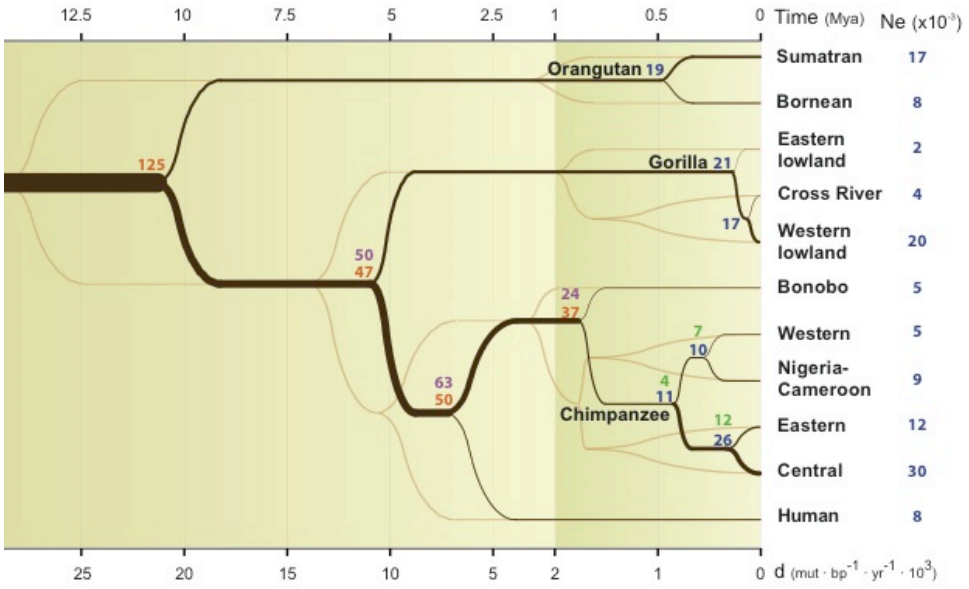


Figure 3.2. Inferred population history.

Population splits and effective population sizes (Ne) during great ape evolution. Split times (dark brown) and divergence times (light brown) are plotted as a function of divergence (d) on the bottom and time on top. Time is estimated using a single mutation rate (μ) of 1·10-9 mut/(bp·year). The ancestral and current effective population sizes are also estimated using this mutation rate. The results from several methods used to estimate Ne, (COALHMM, ILS COALHMM, PSMC and ABC are colored in orange, purple, blue and green respectively). The chimpanzee split times are estimated using the ABC method. The x-axis is rescaled for divergences larger than 2·10-3 to provide more resolution in recent splits. All the values used in this figure can be found in Table S5.

PSMC analyses of historical Ne (**Figure 3.3**) suggests that the ancestral Pan lineage had the largest effective population size of all lineages >3 million years ago (mya), after which the ancestral bonobo-chimpanzee population experienced a dramatic decline. Both PSMC and ABC analyses support a model of subsequent increase in chimpanzee Ne starting ~1 mya, prior to their divergence into separate subspecies. Following an Eastern chimpanzee increase in Ne (~500 thousand years ago, kya), the Central chimpanzees reached their zenith ~200-300 kya followed by the Western chimpanzee ~150 kya. Although the PSMC profiles of the two subspecies within each of the major chimpanzee clades (Eastern/Central and Nigeria-Cameroon/Western) closely shadow each other between 100 kya and 1 mya, the Western chimpanzee PSMC profile is notable for its initial separation from that of the other chimpanzees, followed by its sudden rise and decline (Appendix B, **Figure 3.3**). The different gorilla species also show variable demographic histories over the past ~200 kya. Eastern lowland gorillas have the smallest historical Ne, consistent with smaller present-day populations and a history of inbreeding (**Figure 3.1c**). A comparison of effective population sizes with the ratio of nonsynonymous to

synonymous substitutions finds that selection has acted more efficiently in populations with higher Ne, consistent with neutral theory (Appendix B).

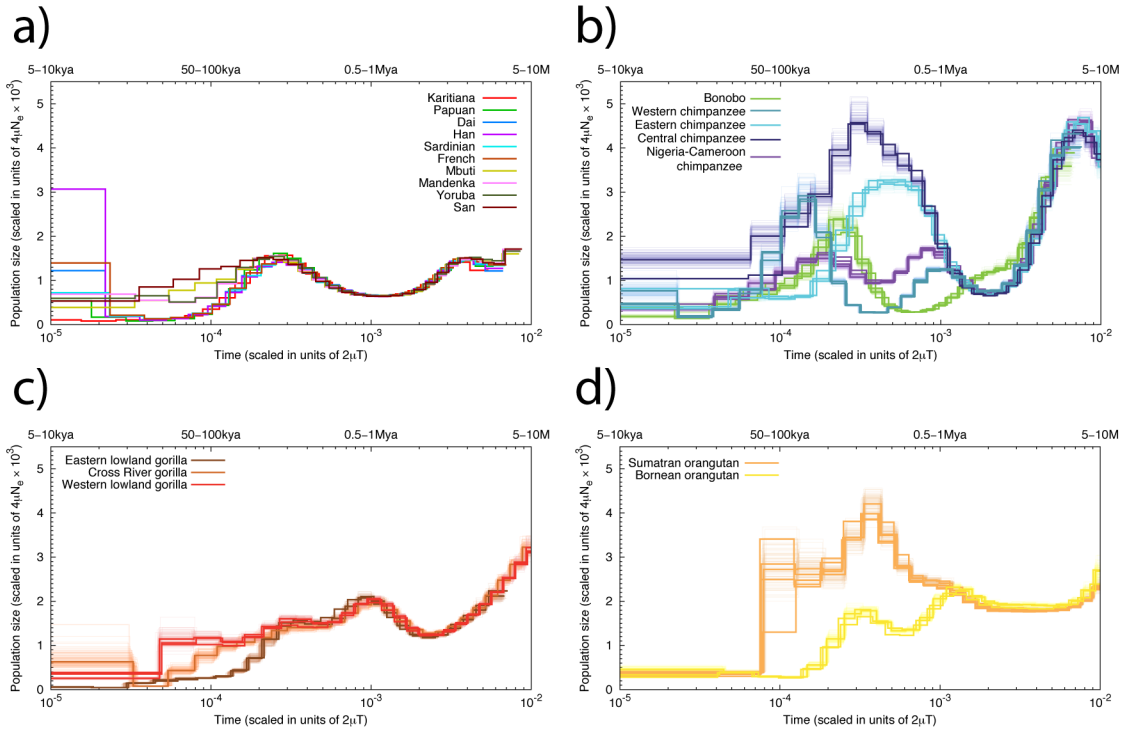


Figure 3.3. PSMC analysis.

Inferred historical population sizes by PSMC. The lower x-axis gives time measured by pairwise sequence divergence and the y-axis gives the effective population size measured by the scaled mutation rate. The upper x-axis indicates scaling in years, assuming a mutation rate ranging from 10-9 to 5·10-10 per site per year. The top left panel shows the inference for modern human populations. In the rest of the three panels, thin light lines of the same color correspond to PSMC inferences on 100 rounds of bootstrapped sequences.

Although the phylogeny of bonobos and Western, Central and Eastern common chimpanzees has been well established based on genetic data (Morin et al. 1994), there is still uncertainty regarding their relationship to Nigeria-Cameroon chimpanzees (Gonder et al. 2011; Bjork et al. 2011). Regional neighbor-joining trees and a maximum-likelihood tree estimated from allele frequencies both show that Nigeria-Cameroon and Western chimpanzees form a clade. A complex demographic history has been previously reported for chimpanzees with evidence of asymmetrical gene flow among different subspecies. For instance, Hey (Hey 2010) identified migration from Western into Eastern chimpanzees, two subspecies that are currently geographically isolated. We find support for this using the D-statistic, a model-free approach that tests whether unequal levels of allele sharing between an out-group and two populations that have more recently diverged (D(H,W;E,C)>16SD). However, no previous genome-wide analysis that has examined

gene flow included chimpanzees from the Nigeria-Cameroon subspecies and a comparison of them with Eastern chimpanzees results in a highly significant D-statistic (D(H,E;W,N)>25SD). Furthermore, TreeMix, a model-based approach that identifies gene flow events to explain allele frequency patterns not captured by a simple branching phylogeny, infers a signal of gene flow between Nigeria-Cameroon and Eastern chimpanzees (p=2x10300). A more detailed treatment of gene flow applying different models and methods may be found in the Appendix B.

Genetic diversity is depressed at or close to genes in almost all species (Suppl. Fig 11.1) with the effect less pronounced in subspecies with lower estimated Ne, consistent with population genetic theory. When we compare the relative level of X chromosome and autosomal (X/A) diversity across great apes as a function of genetic distance from genes, the Eastern lowland gorillas and Bornean orangutans are outliers, with substantially reduced X/A diversity compared to the neutral expectation of 0.75, regardless of the distance to genes. This pattern is consistent with a recent reduction in effective population size (Pool and Nielsen 2007), clearly visible in the PSMC analysis for both species (**Figure 3.3**). However, bonobos also demonstrate a relatively constant level of X/A diversity regardless of distance from genes, with values very much in line with neutral expectations. All other subspecies demonstrate a pattern consistent with previous studies in humans[21] where X/A diversity is lower than 0.75 close to genes and higher farther away from genes.

3.2.3 Genes

It has been hypothesized that loss of gene function may represent a common evolutionary mechanism to facilitate adaptation to changes in an environment (Olson and Varki 2003). There has been speculation that the success of humans may have, in part, been catalyzed by an excess of beneficial loss-of-function mutations (Olson 1999). We, thus, characterized the distribution of fixed loss-of-function mutations among different species of great apes identifying nonsense and frameshift mutations resulting from SNVs (n=806) and indels (n=1080) in addition to gene deletion events (n=96)(Table S4). We assigned these events to the phylogeny and determined that the number of fixed loss-of-function mutations scales proportionally to the estimated branch lengths (R2=0.987 SNVs, R2=0.998 indels).

In addition, we found no evidence of distortion on the terminal branches of the tree compared to point mutations based on a maximum likelihood analysis (Appendix B). Thus, the human branch in particular showed no excess of fixed loss-of-function mutations even after accounting for human-specific pseudogenes (Wang et al. 2006) (Appendix B).

Our analysis provides one of the first genome-wide views of the major patterns of evolutionary diversification among great apes. We have generated the most comprehensive catalog of SNPs for chimpanzees (17.8 million), bonobos (5.8 million), gorillas (18.8 million), and orangutans (17.9 million) (Table 1) to date and identified several thousand AIMs, which provides a useful resource for future analyses of ape populations. Humans, Western chimpanzees, and Eastern gorillas all show a remarkable dearth of genetic diversity when compared to other great apes. It is striking, for example, that sequencing of 79 great ape genomes identifies more than double the number of SNPs obtained from the recent sequencing of more than a thousand diverse humans (1000 Genomes Project Consortium 2010)—a reflection of the unique out-of-Africa origin and nested phylogeny of our species.

3.3 Discussion

We provide strong genetic support for distinct populations and subpopulations of great apes with evidence of additional substructure. The common chimpanzee shows the greatest population stratification when compared to all other lineages with multiple lines of evidence supporting two major groups: the Western and Nigeria-Cameroon and the Central and Eastern chimpanzees. The PSMC analysis indicates a temporal order to changes in ancestral effective population sizes over the last two million years, previous to which the Pan genus suffered a dramatic population collapse. Eastern chimpanzee populations reached their maximum size first, followed by the Central and Western chimpanzee. The Nigerian chimpanzee population size appears much more constant.

Despite their rich evolutionary history, great apes have experienced drastic declines in suitable habitat in recent years (Junker et al. 2012), along with declines in local

population sizes of up to 75% (Campbell et al. 2008). These observations highlight the urgency to sample from wild ape populations to more fully understand reservoirs of genetic diversity across the range of each species and to illuminate how basic demographic processes have affected it. The ~80 million SNPs we identified in this study may now be used to characterize patterns of genetic differentiation among great apes in sanctuaries and zoos and, thus, are of great importance for the conservation of these endangered species with regard to their original range. These efforts will greatly enhance conservation planning and management of apes by providing important information on how to maintain genetic diversity in wild populations for future generations.

3.4 Methods

We sequenced to a mean coverage of 25X (Illumina HiSeq 2000) a total of 79 great ape individuals, representing 10 subspecies and four genera of great apes from a variety of populations across the African continent and Southeast Asia. SNPs were called using GATK (DePristo et al. 2011) after BWA (Li and Durbin 2010) mapping to the human genome (NCBI Build 36) using relaxed mapping parameters. Samples combined by species were realigned around putative indels. SNP calling was then performed on the combined individuals for each species. For indels, we used the GATK Unified Genotyper to produce an initial set of indel candidates applying several quality filters and removing variants overlapping segmental duplications and tandem repeats. We also removed groups of indels clustering within 10 bp to eliminate possible artifacts in problematic regions. Conservative allelic imbalance filters were used to eliminate false heterozygotes that may affect demographic analyses, some of which are sensitive to low levels of contamination. We estimate that the application of this filter resulted in a 14% false negative rate for heterozygotes. Our multispecies study design facilitated this assessment of contamination, which may remain undetected in studies focused on assessing diversity within a single species. The amount of cross-species contamination was estimated from the amount of non-endogenous mitochondrial sequence present in an individual. Because we wished to compare patterns of variation between and within species, we report all variants with respect to coordinates of the human genome reference. For FRAPPE analyses, we used MAF0.06 (human, orangutan, and bonobo) and 0.05 (chimpanzee and gorilla) to remove singletons. For most of the

analyses, we only used autosomal markers, except in the X/A analysis. To determine the amount of inbreeding, we calculated the heterozygosity genome-wide in windows of 1 Mbp with 200 kbp sliding windows. We then clustered together the neighboring regions to account for runs of homozygosity. For the PSMC analyses, we called the consensus bases using samtools (Li et al. 2009).

3.5 Notes

Data resources

Underlying raw sequence data is available through the SRA (PRJNA189439/SRP018689). Data generated in this work are available from http://biologiaevolutiva.org/greatape/.

Additional methods

A complete description of the material and methods is provided in the supplementary note, which is provided as Appendix B to this document.

4 - Evolution and Diversity of Copy Number Variation in the Great Ape Lineage

This chapter has been published: Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J; Great Ape Genome Project, Marques-Bonet T, Eichler EE. *Genome Res.* 2013 Jul 23. [Epub ahead of print]

4.1 Summary

Copy number variation (CNV) contributes to the genetic basis of disease and has significantly restructured the genomes of humans and great apes. The diversity and rate of this process, however, has not been extensively explored among the great ape lineages. We analyzed 97 deeply sequenced great ape and human genomes and estimate that 16% (469 Mbp) of the hominid genome has been affected by recent copy number changes. We identify a comprehensive set of fixed gene deletion (n=340) and duplication (n=405) events as well as more than 13.5 Mbp of genomic sequence that has been specifically lost on the human lineage over the last 16 million years of evolution. We compared the diversity and rates of copy number and single nucleotide variation across different time points of the hominid phylogeny. We find that CNV diversity partially correlates with single nucleotide polymorphism diversity (r^2=0.5) and recapitulates the phylogeny of apes with few exceptions. Duplications significantly outpace deletions (2.8-fold), especially along ancestral African great ape branches. The load of segregating duplications remains significantly higher in bonobos, Western chimpanzees, and Sumatran orangutans—populations that have experienced recent genetic bottlenecks (P=0.0014, 0.02 and 0.0088, respectively). We find that the rate of fixed deletion has been more clocklike with the exception of the chimpanzee lineage where we observe a twofold increase in the chimpanzee-bonobo ancestor (P=4.79 X 10^{-9}) and evidence of increased deletion load among Western chimpanzees (P=0.002). The latter includes the first evidence of a genomic disorder in a chimpanzee with features resembling Smith-Magenis syndrome mediated by a chimpanzee-specific increase in segmental duplication complexity. We hypothesize that demographic effects, such as bottlenecks, have contributed to larger and more gene-rich segments being deleted in the chimpanzee lineage and that this effect, more generally, may account for episodic bursts in CNV during hominid evolution.

4.2 Introduction

Sequence and assembly of great ape reference genomes has consistently revealed that copy number variation (CNV) affects more base pairs than single nucleotide variation (SNV) (Chimpanzee Sequencing and Analysis Consortium 2005; Locke et al. 2011; Cheng et al. 2005). Segmental duplications, in particular, have disproportionately affected the African great ape (human, chimpanzee and gorilla) lineages where they appear to have accumulated at an accelerated rate (Marques-Bonet et al. 2009; Cheng et al. 2005). This has led to speculation that differences in fixation and copy number polymorphism may have contributed to the phenotypic "plasticity" and species-specific differences between humans and great apes(Varki et al. 2008; Olson 1999). While there is some evidence that fixed deletions and duplications contribute to morphological differences between humans and great apes (McLean et al. 2011; Dennis et al. 2012; Charrier et al. 2012), a comprehensive assessment of these differences at the level of the genome has not yet been performed. Previous studies of CNV have been predominated by array comparative genomic hybridization (CGH) experiments (Locke et al. 2003b; Perry et al. 2008; Gazave et al. 2011; Dumas et al. 2007; Fortna et al. 2004), which provide limited size resolution, are imprecise in absolute copy number differences, and are biased by probes derived from the human reference genome. Comparisons of reference genomes have been complicated by assessments of a single individual and distinguishing CNVs from assembly errors (Chimpanzee Sequencing and Analysis Consortium 2005; Locke et al. 2011; Ventura et al. 2011; Prüfer et al. 2012). Here, we compare the evolution and diversity of deletions, duplications, and SNVs in 97 great ape individuals sequenced to high coverage (median ~25X) (Prado-Martinez et al. 2013). The set includes multiple individuals from the four great ape genera, including Bornean and Sumatran orangutans, each of the four recognized chimpanzee subspecies, bonobos, and both Eastern and Western gorillas in addition to ten diverse humans and a high-coverage archaic Denisovan individual. This dataset provides unprecedented genome-wide resolution to interrogate multiple forms of genetic variation and a unique opportunity to directly compare mutational processes and patterns of diversity in great apes.

4.3 Results

4.3.1 Patterns and diversity

We constructed maps of deletions and segmental duplications by measuring sequence read-depth in 500 bp unmasked windows across the genome (Sudmant et al. 2010). We employed a scale-space filtering algorithm to identify deletion and duplication breakpoints (Appendix C Section 3, **Figure 4.1a,b**). In addition to the breakpoints of deletions and duplications, read-depth genotyping allows us to determine the absolute copy number of loci at an individual genome level. We partitioned CNVs into three categories: fixed (i.e., the deletion or duplication was seen as a homozygous event in most individuals), copy number polymorphic, and private (observed only once) (see Appendix C for definitions). Fixed lineage-specific (events occurring on edges between nodes in the species tree) segmental duplications are non-randomly distributed ($P<0.0002$, permutation test) with more than 20% mapping within 5 kbp of shared ancestral duplications (Appendix C section 7)—a phenomenon we previously described as duplication shadowing (Marques-Bonet et al. 2009; Cheng et al. 2005). Deletions, in contrast, are randomly distributed across great ape genomes with respect to one another ($P>0.2$, permutation test).

Table 4.1: Summary of all fixed and segregrating duplications and deletions identified in 97 great apes.

lineage	substitutions along branch (bp)	copy corrected fixed duplicated basepairs	fixed deleted basepairs	fixed duplicated sites	fixed deleted sites (>2kb)	Copy number polymorphisms (>2kb*)		θ CNVs / genome
						private†	segregating‡	
Bornean/Sumatran Orangutan Ancestor	36,672,566	27,854,532	11,292,486	885	1,111	-	-	-
Bornean Orangutan (n=9)	4,645,727	487,342	411,190	48	36	357	310	97.5
Sumatran Orangutan (n=8)	4,444,985	361,732	210,551	38	10	501	405	122.1
Human/Denisova Ancestor	15,006,845	5,715,010	4,303,500	282	919	-		-
Human (n=10)	2,113,519	280,445	52,505	19	12	383	181	49.4
Denisova (n=1)	1,909,910	142,093	931,589	6	113			-
Chimpanzee/Bonobo Ancestor	11,212,835	6,050,953	7,943,919	121	665	-	-	-
Bonobo (n=14)	5,075,887	860,900	1,644,331	67	149	408	265	77
Chimpanzee ancestor	2,322,863	802,219	451,846	30	64	-	-	-
Nigerian/Western chimpanzee	372,805	13,830	21,208	2	4	-	-	-
Central/Eastern chimpanzee	172,063	1,790	8,730	1	1	-	-	-
Western chimpanzee (n=5)	2,466,250	31,250	283,012	15	34	303	106	37.5
Nigerian-Cameroon chimpanzee (n=10)	2,322,863	1,302	35,715	1	7	241	197	55.5
Eastern chimpanzee (n=6)	2,494,927	13,994	10,633	2	3	602	206	68.2
Central chimpanzee (n=2)	2,380,218	20,823	12,655	10	5	813	-	-
Human/Chimpanzee/Bonobo Ancestor	3,613,343	4,227,312	1,232,172	269	213	-	-	-
Gorilla Ancestor	16,747,559	29,413,779	7,701,191	278	770	-	-	-
Western/Cross River Gorilla (n=29)	487,514	339,371	178,376	35	9	532	418	98.3
Eastern Gorilla (n=3)	2,953,764	446,285	533,317	89	72	178	-	-
Cross River Gorilla (n=1)	2,265,508	111,722	1,194,717	31	146	-	-	-
African Great Ape Ancestor	21,284,312	55,560,468	8,729,420	1,744	1,185	-	-	-
Ancestral to Great Apes	-	192,822,491	-	7,863	-	-	-	-
Total (n=97)	***140,966,263***	**325,559,643**	**47,183,063**	**11,836**	**5,528**	**4,318**	**2,088**	-

CNPs encompassing 2kb of non-repeat-masked sequence are reported, the median size of CNPs reported is 8061 bp

† 56.9Mbp in total of private CNPs and ‡ 39.3 Mbp of segregating CNPs

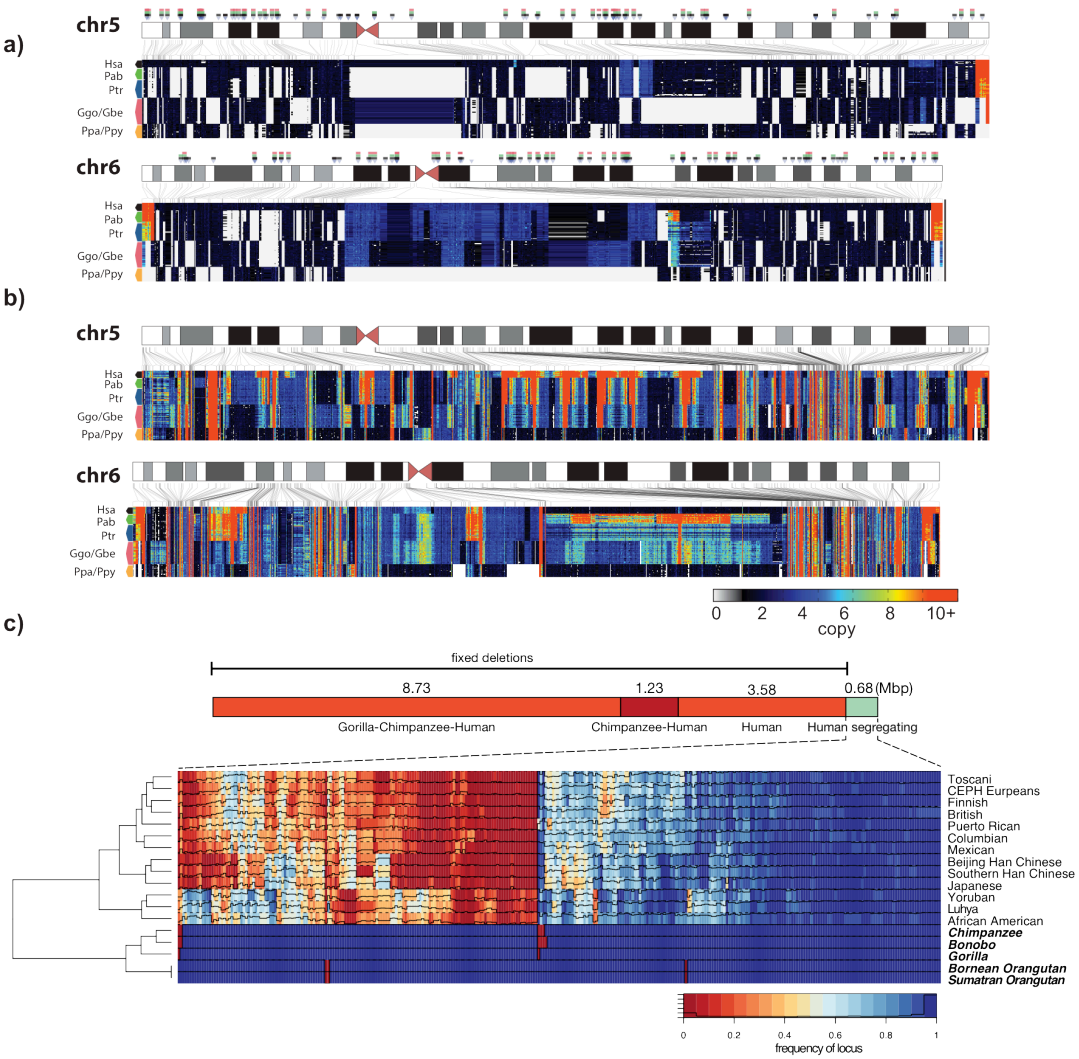


Figure 4.1. Duplication and deletion landscape.

a. Ideograms of human autosomes 5 and 6 overlaid with copy number heat maps of the deletion landscape of great apes across 7 species and 11 distinct populations. Each row represents one of 97 individuals sorted by species; each column shows the estimated copy number in each of these individuals for deleted loci in 500 bp unmasked windows. Arrows above the chromosome ideogram indicate deletions identified along the lineages leading to the human species, the African great ape, chimpanzee-human, and human lineages, respectively. b. Ideograms of human autosomes 5 and 6 overlaid with copy number heat maps of the duplication landscape of great apes. c. Breakdown of the number of base pairs lost along the lineage leading to humans identified by screening sequence absent from the human reference genome yet present in the orangutan, gorilla, or chimpanzee reference genomes against the 97 great apes sequenced in this study. A total of 13.54 Mbp have been lost in these lineages since the divergence of African great apes and orangutans. We find an additional 680 kbp (316 loci) of sequence absent from the human reference genome (4.8% of the total) is fixed in all nonhuman great apes and segregating humans. For these loci a hierarchically clustered heat map is shown. Colors indicate the frequency of sequences absent in the human reference genome assessed in 624 diverse individuals from 13 different populations sequenced to low coverage by the 1000 Genomes Project and found to be segregating ≥5% frequency in at least one population. The hierarchical clustering recapitulates all the relationships between the individual human populations and the different great ape species assessed in this study. We identify 53.8 kbp of sequence segregating exclusively in African populations compared to only 1.4 kbp of sequence segregating specifically in Europeans.

We parsimoniously assigned fixed events to ancestral branches based on comparisons between populations. In total, we identify 469 Mbp of CNVs (**Table 4.1**). This set includes 11,836 fixed duplicated loci (325 Mbp; median length of 3,778 bp), 5,528 fixed deletions (47 Mbp; median size=4,227 bp), and 6,406 private and segregating copy number variants (96.2 Mbp) (**Table 4.1** and Appendix C section 3). In order to assess the accuracy of these calls, we performed 104 fluorescent *in situ* hybridization (FISH) experiments confirming 102 of the loci tested (98.1%). We also designed three custom duplication and deletion array comparative genomic hybridization (CGH) microarrays confirming 85% of CNPs (1,294/1,520 of events >2 kbp), 96.9% (3,660/3,776) of fixed duplications, and 98.6% (3,966/4,021) of fixed deletions (Appendix C section 4). As part of our assessment of deletions, we also screened sequence absent from the human reference genome yet present in one or more of the great ape reference genomes (Appendix C section 6). Since these "missing sequences" may represent artifacts or polymorphisms, we additionally estimated the frequency of each segment in 624 diverse humans from 13 different populations (1000 Genomes Project Consortium et al. 2012). We assigned 13.54 Mbp of human deletions unambiguously to specific time intervals during the evolution of our species. Notably, ~5% of these deleted sequences are still segregating in the human population, consistent with known population relationships among extant humans (**Figure 4.1c**).

Since fixed deletions are less likely than duplications to be subjected to recurrent mutation events, we assessed whether they might serve as reliable genetic markers for phylogenetic reconstruction of ape populations. The resulting neighbor-joining tree of deletion genotypes (**Figure 4.2a**) accurately recapitulates the ape phylogeny, including separation of Bornean and Sumatran orangutans, Eastern and Western gorillas, and bonobos and chimpanzees with high confidence. In contrast, however, to trees built from mitochondrial haplotypes or autosomal single nucleotide polymorphism (SNP) data from the same population (Prado-Martinez et al. 2013), Central chimpanzees emerge as an out-group to the other chimpanzee subspecies (96% support). Interestingly, we observed a slight distortion towards increased branch length for the chimpanzee-bonobo ancestral lineage, which becomes more pronounced for larger deletions (see below section *Rates and CNV Load*; Appendix C

section 9). Principal component analysis (PCA) of segregating structural variants also captures the subspecies relationships in addition to inter-population diversity (**Figure 4.2b**). Our analysis shows that estimates of SNP diversity and segregating copy number variants (as measured by Watterson's θ) are correlated (r^2=0.5 Pearson, P=0.02).

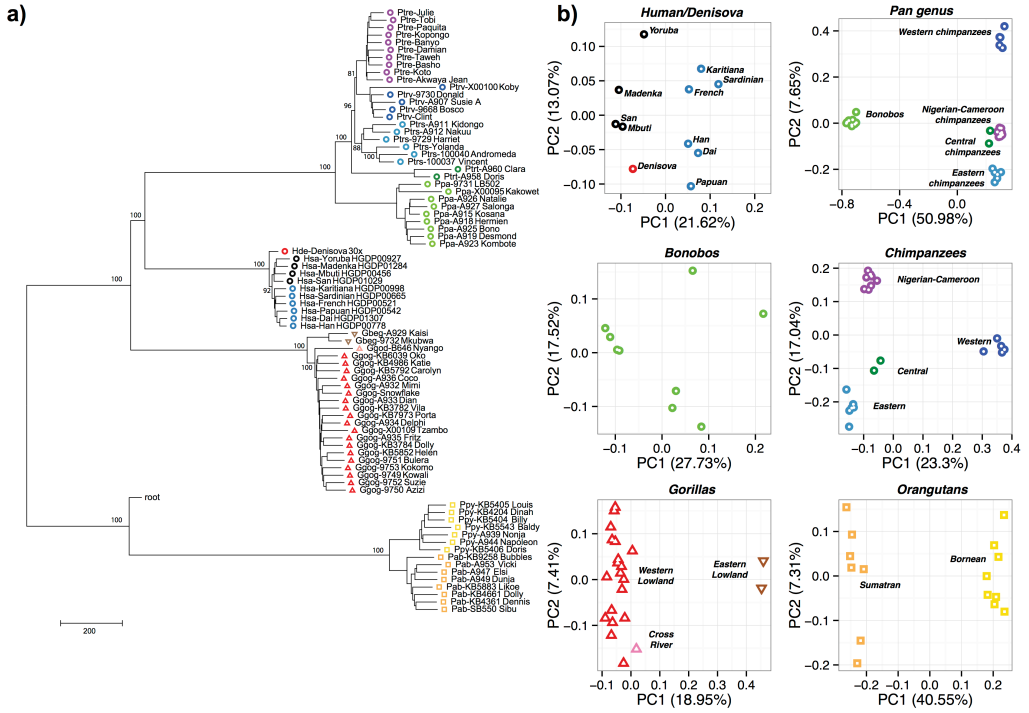


Figure 4.2. Hominid deletion phylogeny.

a. Neighbor-joining tree constructed from pairwise edit distance of genotypes for fixed and segregating deletions >5 kbp. Branch length confidence estimates were generated by repeatedly subsampling 50% of the variants and regenerating the topology. All species and subspecies relationships are reconstructed with high confidence and are concordant with the topology identified from SNPs with the exception of Central chimpanzees, which form an out-group to the other chimpanzee subspecies as a result of their increased diversity. SNP-based trees cluster Central and Eastern chimpanzees on a single clade. Among chimpanzees, the three individuals Yolanda, Andromeda and Vincent, the Eastern-most individuals assessed in this study from Gombe National Reserve in Tanzania, cluster together with strong support. Additionally, the individuals Tobi and Julie, a distinct subpopulation of Nigerian chimpanzees by SNP analysis, cluster together. Eastern lowland gorillas form an out-group to the gorilla clade and the Cross River gorilla clusters as an out-group to Western lowland gorillas. The archaic Denisova individual clusters as an out-group to all humans with 97% support. b. PCA of segregating deletion genotypes recapitulates intra-population relationships and additionally the relative diversity within the populations assessed.

4.3.2 Genes

The availability of multiple sequenced genomes allows us to generate a comprehensive list of fixed deletions and duplications that disrupt genes along each branch of the ape lineage (see Appendix C section 5). We identified 407 lineage-specific gene duplications and 340 deletions with complete or partial exon loss (**Figure 4.3a,b,c**) with an excess of gene duplication events in the African great ape and chimpanzee-human ancestor. Lineage-specific duplications include a chimpanzee expansion of *PRDM7*, a high-identity paralog of *PRDM9*, in common chimpanzees (10-20 copies) and bonobos (35-40 copies) that is stratified among chimpanzee populations; a 75 kbp gorilla-specific expansion of *C1QTNF* and *AMACR*—genes important in brain and skeletal development; and 33 genes duplicated specifically in human since divergence from chimpanzee. This includes two genes that appear to have been duplicated, or to have increased in frequency, in the human lineage after the divergence from Denisova, ~700 kya (Meyer et al. 2012), with the caveat that only a single Denisovan individual was assessed. These potential Homo sapiens-specific genes include *BOLA2*, which resides just inside the critical region of the 16p11.2 locus, the deletion of which results in developmental delay, intellectual disability, and features of autism (Kumar et al. 2008; Weiss et al. 2008).

Among the 340 exonic gene-loss events, orangutans show the highest number (90), commensurate with their divergence from African great apes ~16 mya. Strikingly, the second highest number of gene-loss events occurs in the chimpanzee-bonobo ancestral lineage where 57 genes exhibit exonic deletions. As expected, we find a massive enrichment for olfaction genes (96/340) in addition to fixed deletions of immunity (*IL36*, *IL37* in chimp, *CCL26* in gorilla), drug detoxification (*CYP3A43* in Denisova, *CYP2C18* in humans and chimps), and sperm surface membrane genes (*ADAM2* in gorilla; *ADAM3A* in gorilla and *Pan* genus). Some genes appear to have undergone both lineage-specific duplication and loss. Of note is the carboxyl-esterase gene family (**Figure 4.3c**; *CES1*, *2*, *3*), which appears to have expanded independently in all great ape lineages with the exception of human where it remains diploid or alternatively has been subjected to deletion.

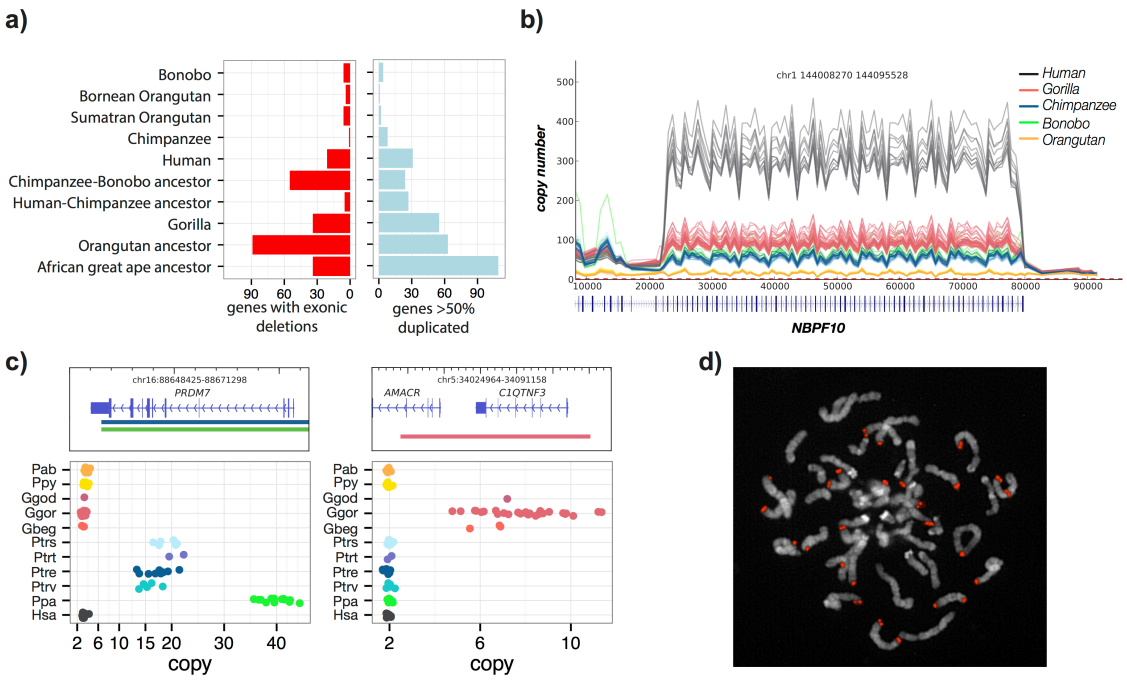


Figure 4.3. Genic variation.

a. Summary of the number of genes with exonic deletions and genes duplicated in each of the lineages assessed in this study. We identify 340 genes lost throughout the great ape lineage. While orangutans show the highest number of gene-exon-loss events (89), strikingly, the second highest number of gene-exon-loss events was in the chimpanzee-bonobo ancestral lineage where 55 were lost. b. A line plot of the copy number over the DUF1220 domain of NBPF10. This domain has expanded specifically in the African great ape lineage with human exhibiting ~300 copies compared to 50-100 in chimpanzee, bonobo, and gorilla. c. Lineage-specific great ape duplication events encompassing genes. Gene models are drawn with the duplication breakpoints shown below colored by lineage and dot-plots of the copy number in all individuals assessed in this study. PRDM7 is a close paralog of PRDM9, binding of which associates with recombination hotspots in humans. We find PRDM7 is specifically duplicated in the Pan genus and highly stratified; Nigeria-Cameroon chimpanzees have 10-15 copies while Eastern and Central chimpanzees have 16-20. Bonobos exhibit 30-40 copies of the gene. FISH assays demonstrate the extra copies to be the result of subtelomeric duplicative transposition (d). AMACR and C1QTNF3 are also specifically duplicated in gorillas. Mutations in AMACR have been shown to result in adult onset neurological disorders (Ferdinandusse et al. 2000) and C1QTNF plays a key role in skeletal development, inducing increased growth of murine mesenchymal cells with overexpression (Maeda et al. 2001).

We were also interested in genes that were lost in the human lineage, and therefore absent from the human genome, since these have been hypothesized to contribute disproportionately to the evolution of human adaptive traits (Olson 1999). We, thus, analyzed the 13.54 Mbp of human fixed deletions (see above) for the presence of open reading frames (ORFs) where there was also support for a multi-exon spliced transcript from RNAseq data from multiple nonhuman primate tissues (Brawand et al. 2011) (Appendix C section 6). By this definition, we identified 86 putative gene losses along the branches leading to the human lineage—40 since divergence from chimpanzee. A search of these ORFs against the RefSeq protein database yielded not only previously annotated gene-loss events, such as the human-specific *SIGLEC13* (Wang et al. 2012) and *CLECM4* (Ortiz et al. 2008) deletions, but 42 previously

unannotated or only predicted protein-coding genes with homology to other genes, 28 of which intersect highly conserved elements (HCEs, (Siepel et al. 2005)). In total we identified 180 kbp of highly conserved sequence within these fixed deletions, a marked depletion compared to the 3-8% of the human reference genome encompassed by HCEs. However, 18% and 12% of regions were located within introns or within 10 kbp upstream or downstream of annotated genes, respectively, suggesting some of these loci may have a potential regulatory impact as has been previously suggested (McLean et al. 2011).

4.3.3 Rates and CNV load

Comparing deletions and duplications among different great ape lineages (>2 kbp), we find that the number of base pairs added by duplication significantly exceeds that of deletions by a factor of 2.8, although this ratio varies considerably depending on the specific lineage (**Table 4.1**). In this analysis, we considered only those base pairs added by new duplication excluding the ancestral locus. Overall, we find that the contribution of fixed base pairs by deletion and duplication is ~1.4-fold greater than that of single base-pair substitutions. We estimated rates of duplication and deletion throughout great ape evolution by normalizing the number of fixed base pairs that were lost or gained as a function of genetic branch length as well as divergence time (**Figure 4.4a,b**). All analyses were additionally computed in units of the number of events per millions of years (Appendix C Table 9.1) and exhibited the same observed trends. Although there was an acceleration of duplicated base pairs along the ancestral African great ape lineage ($P=9.786 \times 10^{-12}$), we predict that the rate of fixation subsequently declined in the ancestral lineage of human and chimpanzee and at a slower rate in the gorilla lineage. Our analysis shows that the rate of duplication in base pairs exceeds by threefold the rate of substitution in the African great ape lineage and is ~7-fold higher than the rate of duplication in the human lineage. This results in a significant excess of fixed gene duplication events occurring at this time point (**Figure 4.4c**; $P=1.66 \times 10^{-20}$).

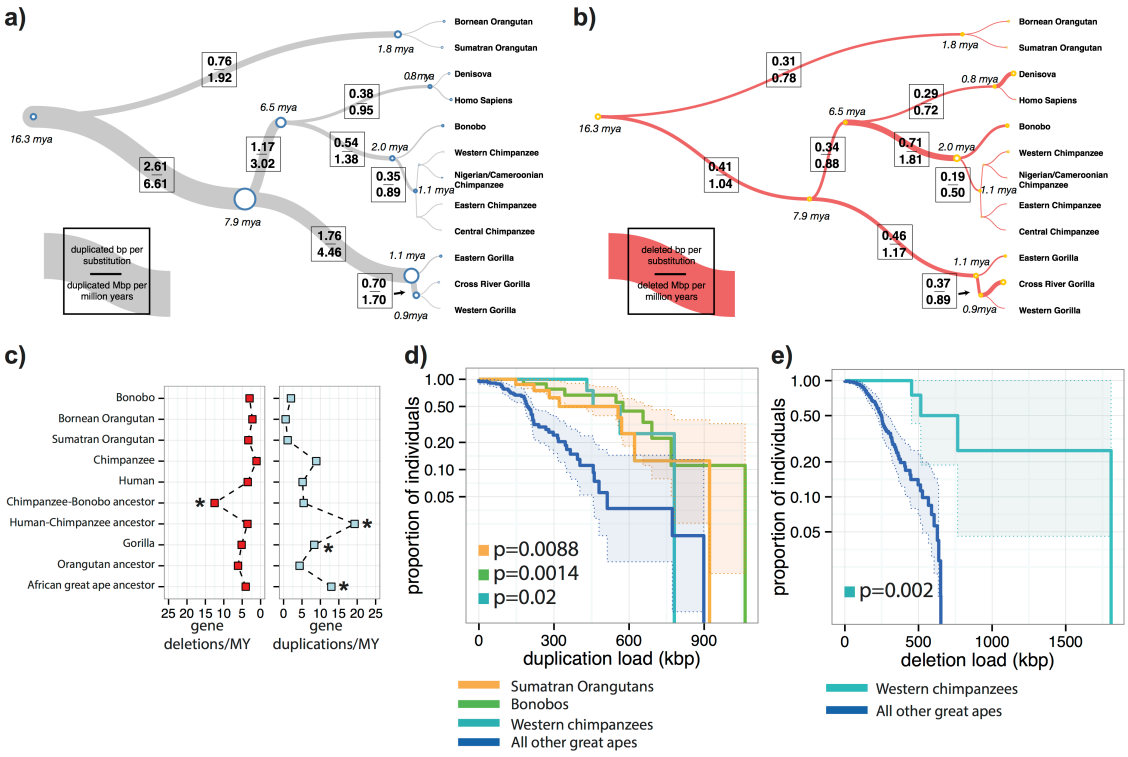


Figure 4.4. CNV rates and polymorphism.

Rates of a) duplication and b) deletion accumulation as a function of the number of substitutions along each branch of the great ape phylogeny. Tree branch lengths are scaled proportionally to the number of substitutions while tree widths are scaled proportionally to the number of duplicated base pairs per substituted base pair. Duplicated base pairs were added ~2.6-fold the rate of substitution along the African great ape ancestral branch, which rapidly declined in the chimpanzee-human ancestral lineage and more slowly in the gorilla lineage. In contrast, the rate of deletion in the great ape lineage is fairly consistent along all branches (mean of 0.32 deleted bp per substitution) with the exception of the chimpanzee-bonobo ancestral lineage where an approximate twofold increase in the rate of deletion is observed (0.71 deleted bp per substitution). c. The rate of genic deletion events and gene duplication events per million years plotted for each of the lineages assessed in this study. The rate of gene deletion events is significantly higher in the chimpanzee-bonobo ancestor (p=5.262e-9). An acceleration in the number of gene duplications is observed in the African great ape ancestor, the human-chimpanzee ancestor, and the ancestral gorilla lineage (p=1.663e-20). d. A survival curve of the total load of segregating duplications >30 kbp in Western chimpanzees, Sumatran orangutans, and bonobos compared to all other great apes shows that these three populations harbor an increased total number of duplicated base pairs (results significant for each individual population and combined). e. A survival curve for the total load of deletions >30 kbp in Western chimpanzees compared to all other great apes shows a significant excess of deletions in this population. Western chimpanzee populations show the lowest diversity of any of the populations assessed in this study and the most fixed deletions of all chimpanzee species assessed.

The corresponding analysis for deletions shows a markedly different pattern with the rate occurring in a more clocklike manner throughout most of the tree with the notable exception of the ancestral lineage of chimpanzees and bonobos. We observe an approximate twofold increase in the rate of deleted base pairs leading to a distortion specifically along this branch ($P=4.79 \times 10^{-9}$). This increase results from an excess of large (>5 kbp) chimpanzee-bonobo ancestral deletions, which affect significantly more genes when compared to all other great ape lineages (**Figure 4.4c**; $P=4.397 \times 10^{-8}$). Notably, this excess of deletions corresponds to a predicted

collapse in the ancestral chimpanzee-bonobo effective population size (Ne) ~3 mya ((Prado-Martinez et al. 2013); Appendix C section 5).

As demography may have played a significant role in the excess rate of deletion in the chimpanzee-bonobo ancestor, we sought to estimate the relative burden of segregating duplications (**Figure 4.4d**) and deletions (**Figure 4.4e**) in each of the great ape populations by comparing CNV and SNP diversity (Methods). Specific populations showed an increased burden of CNV load, both in the total number of base pairs affected and in the number of events (Appendix C section 10), although humans were not remarkable in this regard as has been hypothesized (Varki et al. 2008). Western chimpanzees, bonobos, and Sumatran orangutans all showed an excess of segregating duplications >30 kbp consistent with an increased duplication burden in these populations (P=0.02, 0.0014 and 0.0088, respectively; Appendix C section 9). Western chimpanzees were the only population to show an additional excess of segregating deletions >30 kbp (P=0.002). All of these populations are predicted to have experienced striking collapses in their effective population sizes during recent evolution ((Prado-Martinez et al. 2013), in press; Appendix C section 5). Western chimpanzees in particular exhibit the lowest overall nucleotide diversity and effective population size (8×10^{-4} Het/bp, Ne=9800) among all populations assessed. This subspecies also harbors the largest number of fixed deletions (34 events encompassing 276 kbp), consistent with a population that experienced a severe bottleneck.

4.3.4 A putative chimpanzee genomic disorder

Among the Western chimpanzees assessed, we identified one particularly striking private structural variant—a ~1.7 Mbp microdeletion on 17p11.2 in the individual Susie-A (BPRC, Holland) (**Figure 4.5a**). This deletion encompasses 29 genes, including *RAI1* (retinoic acid-induced 1). In humans, deletions of this locus cause Smith-Magenis syndrome (SMS). SMS is a rare syndrome with an incidence of 1 in 15,000-25,000 (Elsea and Girirajan 2008) resulting in severe behavioral abnormalities, mental retardation, and developmental delay. The clinical features of this chimpanzee bear striking similarity to many of the phenotypes observed in SMS

patients (**Table 4.2**), including common SMS maladaptive behaviors such as aggression and disobedience, obesity, a humped back indicative of kyphoscoliosis, renal abnormalities, and velopharyngeal insufficiency (Appendix C section 10). The chimpanzee deletion is flanked by multiple loci that have undergone expansion in the *Pan* genus (**Figure 4.5b**). The typical human SMS deletion spans an additional 2 Mbp and has breakpoints mapping to different locations and different segmental duplication blocks (**Figure 4.5c**). In order to resolve the chimpanzee duplication organization, we sequenced to high quality a total of 20 large-insert BAC clones (2.9 Mbp, ~1.73 Mbp nonredundant sequence) identifying ~765 kbp of sequence absent from panTro3. We find these blocks have increased in size and complexity in the chimpanzee lineage with at least an additional 600 kbp of duplicated sequence compared to humans (**Figure 4.5d**). These results predict that the chimpanzee genome harbors a novel 17p11.2 architecture whose more complex organization predisposes to a deletion resulting in an SMS-like phenotype. This identifies the first chimpanzee-specific genomic disorder mediated by lineage-specific expansion and restructuring of segmental duplications creating a putative chimpanzee-specific hotspot for deletion.

Table 4.2: Common clinical features of Smith-Magenis Syndrome and related features of Western chimpanzee Susie-A with a corresponding 17p11.2 deletion.

Clinical features of Smith-Magenis Syndrome	Related clinical features of Western Chimpanzee SusieA
Maladaptive behavioral issues including: • Frequent outbursts and tantrums • Aggression • Disobedience • Emotional volatility • Tendency toward attention seeking behaviors • Lack of respect for personal space during conversation	Susie-A is described as exhibiting • 'marked impairment in her behavioral skills,' • 'mean and more aggressive than usual' behavior • When in the close proximity to people SusieA would palpate her genitals, a challenging "culturally abnormal" sexualized behavior in chimpanzees
Abnormal curvature of the spine and scoliosis are present in 50-75% of individuals with SMS (Greenberg *et al*, 1996)	Susie-A had a hump on her back indicative of kyphoscoliosis (abnormal spine curvature)
Edelman *et al* 2007 found that individuals with RAI1 mutations and deletions were likely to be obese. Mouse models of RAI1 deletions additionally demonstrated an obese phenotype for deletions but not duplications of the critical locus (Walz *et al*. 2003, Yan *et al*. 2004); duplications conferred an underweight phenotype. A null RAI1 allele in mice generated by Bi *et al.* 2005 also exhibited obesity.	Susie-A was an obese chimp with a body weight of ~90kg. The normal body weight of a mature chimpanzee is between 50kg and 65kg and often less for female Western chimpanzees.
>75% of SMS patients exhibit otolaryngologic abnormalities including a hoarse deep voice (Greenberg *et al,* 1996). 50-75% exhibit tracheobronchial problems and velopharyngeal insufficiency.	Susie-A exhibited tracheitis and had grossly overlapping tracheal cartilage ends which are suspected to contribute to documented hoarse breathing noises she would make.
Renal abnormalities have been shown to occur in 20-35% of SMS patients (Greenberge *et al,* 1996) Additionally, 25%-50% of SMS patients have been found to have cardiac abnormalities.	The most significant postmortem histolopathological finding of SusieA was chronic interstitial nephritis which was clinically observed with increased creatinine and leukocytosis consistent with renal failure. Obesity and tracheitis presumably played a role in her final cardiorespitory failure as well.

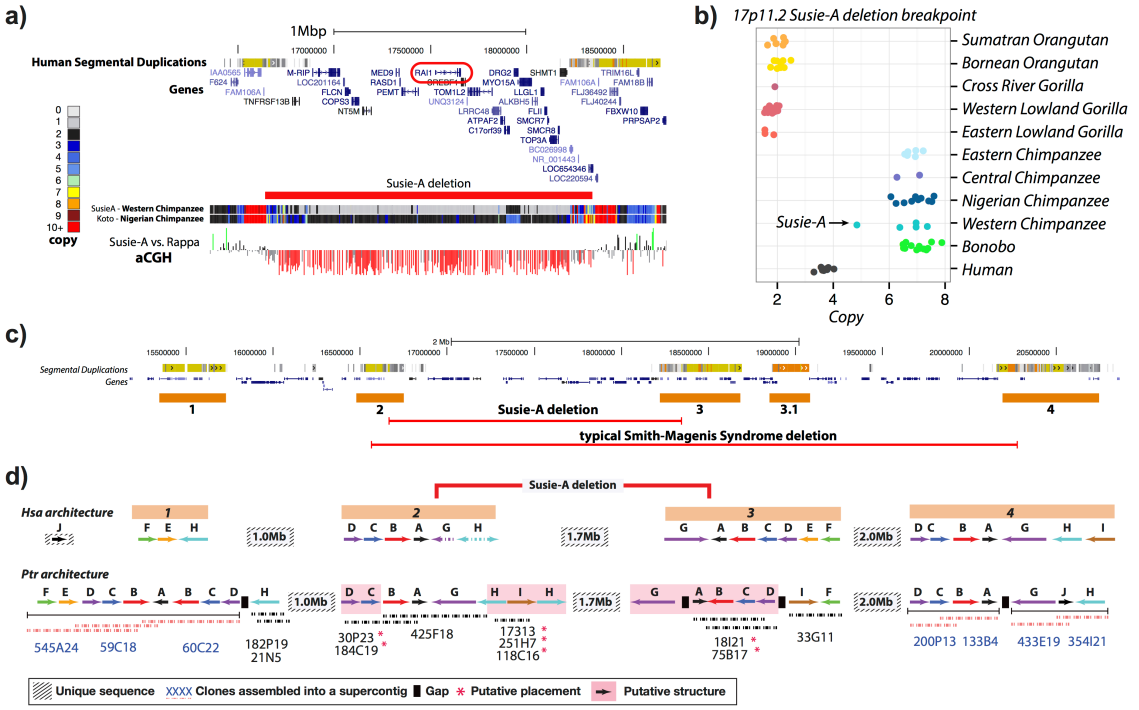


Figure 4.5. A chimpanzee genomic disorder.

a. A genome browser snapshot of 17p11.2 Smith-Magenis syndrome (SMS) critical region with a copy number heat map of the Western chimpanzee Susie-A and the Nigeria-Cameroon chimpanzee Koto. Susie-A has a 1.7 Mbp deletion of this locus, which encompasses RAI1, the critical gene associated the SMS phenotype. We confirm this deletion by array CGH. b. Copy number of great apes assessed in this study over the Susie-A deletion breakpoint 2 H-duplicon. c. Organization of the 17p11.2 SMS locus and 17p12 in humans with four blocks of segmental duplication. The typical human SMS deletion spans ~3.7 Mbp with different breakpoints than the Susie-A deletion (Elsea and Girirajan 2008). d. Segmental duplication architecture of the 17p11.2 locus as represented in the human reference genome and constructed in chimpanzees from high-quality sequencing of 22 BAC clones. We were able to assemble and anchor 11 of these clones into 7 contigs. The remaining 11 contigs were placed at their most likely locations and orientations based on their underlying duplication architecture and read-depth analysis of Susie-A compared to normal chimpanzees. We hypothesize a non-allelic homologous recombination event between the directly oriented chimpanzee G duplicons resulted in Susie-A's deletion.

4.4 Discussion

We present the first genome-wide assessment of duplication and deletion diversity where single nucleotide substitutions have been used to calibrate CNV accumulation over the course of great ape evolution. There are three novel findings in this study. First, chimpanzees show an excess of large deletions early in their history. This is in stark contrast to almost every other population of great ape where deletions have accumulated in a more clocklike fashion. The ancestral human lineage does not show an excess in the number of duplicated or deleted base pairs despite previous predictions (Varki et al. 2008; Olson 1999). Second, specific populations of great apes show an excess of copy number polymorphic duplications, notably Western chimpanzees, bonobos, and Sumatran orangutans. Only the Western chimpanzee shows evidence of increased deletion polymorphism. These three populations stand

out in that they are predicted to have experienced sudden rises and crashes in effective population size. The Western chimpanzees are the most extreme in this regard, showing the strongest signal of genetic drift and the largest excess of ancestry-informative markers—consistent with the strongest bottleneck.

One possibility may be that CNPs (both duplications and deletions), in general, increase with small effective population sizes but that a severe bottleneck is necessary in order to result in an increase in deletion burden as a result of strong selection against deletions. The neutral nature of the vast majority of SNPs suggests that reductions in diversity may, in some cases, have little effect on overall fitness, in contrast to large structural variants. Human investigations as well as Drosophila studies have additionally shown that deletions affecting genes are significantly more deleterious than duplications (Cooper et al. 2011; Emerson et al. 2008). Indeed, analyses of the theoretical relationship between N_e and rates of deletion and duplication have suggested fluctuations in effective population size may play a significant role in overall variations in genome size among organisms (Lynch 2007). These findings would explain the excess of deletions specifically in the ancestral chimpanzee branch as this species shows the most drastic decline in effective population size when compared to orangutan, human, and gorilla. Humans once again are similar to other great apes with respect to CNP burden and do not particularly stand out, although the number of genomes compared are few.

Finally, we report the first evidence of a genomic disorder in the chimpanzee lineage. The phenotype is remarkably similar to SMS but the breakpoints are not shared with the common recurrent deletion seen in humans. Our sequencing analysis shows that the chimpanzee 17p11.2 breakpoints have radically changed in structure and content facilitating non-allelic homologous recombination. Owing to the evolution of this chimpanzee-specific architecture, we predict that this locus represents a chimpanzee genomic hotspot of mutation and that additional recurrent microdeletions may be encountered among the chimpanzee population. It is somewhat surprising that Susie-A was captured from the wild, albeit as a young chimp. In light of her behavioral anomalies, it is unlikely that she would have survived to adulthood outside of captivity. This raises the intriguing possibility that

additional cases, and perhaps novel recurrent genomic disorders, may be encountered as apes continue to be bred in captivity. Most comparative sequencing studies of human genomic disorder breakpoint regions have reported increasing complexity in the human lineage as a predisposing factor to rearrangement associated with disease (Antonacci et al. 2010; Boettger et al. 2012; Rochette et al. 2001). Our results show that loci of increasing complexity are present in other great ape lineages creating species-specific hotspots prone to deletion and disease.

4.5 Methods

Read-depth profiles were initially constructed from whole-genome sequence from 120 great ape individuals. We assessed the quality of each of these genomes by assessing the sequence read-depth in regions of the genome (1.1 Gbp) regarded as copy number invariant (Appendix C section 1). We excluded 23 individual genomes that showed considerable heterogeneity in their read-depth presumably due to non-uniformity (Appendix C Figure 1.1). We report analysis on the remaining 97 genomes: 75 were sequenced as part of the Great Ape Genome Diversity Project (Prado-Martinez et al. 2013) to a mean coverage of ~25X on an Illumina HiSeq 2000 while an additional nine orangutans, ten humans, and the Denisova individual were sequenced as part of the Orangutan Genome Project and the Denisova Genome Project (Locke et al. 2011; Meyer et al. 2012). Individuals sequenced as part of the Great Ape Genome Project were originally selected to best represent wild natural diversity by focusing on captive individuals of known wild-born origin in addition to individuals from protected areas in Africa (Table S1). Individual genome subspecies designations were assigned as reported by sample sources and confirmed by SNP genotyping and PCA analysis. All reads were first divided into their 36 bp constituents and mapped to the human reference genome (NCBI36) using the mrsFASTc read aligner (Hach et al. 2010). Read-depth estimates across the genome were corrected for the underlying GC content, and a calibration curve from regions of known copy number was used to assign copy number estimates to windows of the genome. These regions were then segmented using a scale-space filtering algorithm (Appendix C section 3).

Briefly, the scale-space filtering algorithm transforms the windowed copy number waveform, f(x), into a set of waveforms, f(x,σ), where values of σ represents the standard deviation of a Gaussian smoothing kernel applied to the original waveform. Contours of this transform are then traversed from large values of σ as σ→0 and the resulting segments are hierarchically clustered. We also masked regions of high GC content (>57%, corresponding to 2.23% of the genome). Array CGH validation experiments were performed in duplicate for every sample tested with Cy3 and Cy5 labeling dyes swapped. Probes giving opposite signals in the dye swap experiment were discarded. Only loci with at least three probes were considered for validation. CNV load comparisons were performed using Kaplan-Meier survival curves and statistical tests were corrected for sample size. BAC clones were selected from the chimpanzee BAC library CHORI-251 corresponding to the male chimpanzee Clint. Clones were sequenced using a PacBio RS system using standard protocols. The library was prepared with a 10 kbp insert size and sequence generated with C2 chemistry in 90-minute movies.

4.6 Notes

Data resources

Copy number maps for the 97 individuals assessed in this study are available online (http://eichlerlab.gs.washington.edu/greatape-cnv). All lineage-specific and segregating copy number variants are additionally reported in Appendix C Tables S2-S11. All structural variants have been deposited into the database of genomic structural variation (dbVar, http://www.ncbi.nlm.nih.gov/dbvar/) under accession nstd82. Underlying raw sequence reads have been deposited in the short-read archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) under accession SRP018689.

Additional methods

A complete description of the material and methods is provided in the Supplementary Note, which is provided as Appendix C to this document.

5 - Summary and Future Directions

5.1 Assessing hominid genome diversity: insights and challenges

This thesis describes the development and application of methods used to explore the diversity of hominid genomes using short-read second-generation sequencing technologies. In particular, I set out to catalog variation in the duplicated portion of the genome and compare and contrast this variation among and between hominid populations and species. While the advent of second-generation sequencing technologies has enabled significant advances forward in understanding hominid genome diversity, many challenges persist. Here, I highlight a summary of the conclusions and insights offered by this thesis in addition to some of the remaining challenges.

The second chapter of this thesis describes methods to accurately genotype the copy number of segmental duplications from high-throughput sequence data and the application of these methods to assess variation among hominids. These techniques are shown to be highly accurate with 93% of copy number estimates being within +/- 1 of the true copy number. These methods were then applied to assess gene and copy number diversity in 159 diverse human genomes and describe for the first time the extent of variation of highly duplicated portions of the genome among human populations. Among the three continental populations assessed, 64 genes were highly stratified between populations and comparisons with nonhuman great apes identified 52 gene families expanded specifically in humans, 23 of which were diploid in great apes. Importantly, a number of duplicated genes were identified that are present in most or all humans and yet are absent from the human genome. This suggests that efforts to sequence and assemble the most difficult duplicated regions are still necessary. Additionally, the functions of many of the duplicated genes identified are as yet unknown. Determining the functions of these genes is clearly of importance to better understand what effect these genes might have on human biology.

Next, a map of 4.1 million SUN identifiers was developed and used to discriminate between different paralogs demonstrating a high level of specificity (93.2%) to

identify paralog-specific CNVs. This method was then exploited to genotype duplicated genes revealing two different classes of paralogs: 49.2% that are largely copy number invariant and 50.8% that are highly variable.

To expand upon these observations and explore patterns of duplication diversity throughout hominid evolution, the sequencing and characterization of 79 wild and captive-born great apes was undertaken, describing 86.5 million SNPs. A threefold range in heterozygosity was determined among great ape species and an Eastern/Central, Nigeria-Cameroon/Western tree topology created for the chimpanzee subspecies. Historical and current effective population sizes were estimated demonstrating the demographic history of these populations. All of the analyses performed, however, were based on short reads mapped to the human reference genome and to the individual great ape species references. Yet nonhuman great ape reference genomes, lack a complete set of gene annotations and are largely collapsed over complex duplication architectures. A complete understanding of the sequence and structure of great ape genomes requires complete sequencing, finishing and annotation of the genomes of all extant species.

Among those same individuals characterized for SNP diversity, 23,770 CNVs were identified between and among species and used to estimate relative rates of gene duplication and loss along the individual great ape lineages. Exact rates of duplication and deletion were then calculated along each of the lineages, identifying an excess of deletions in the chimpanzee-bonobo ancestral lineage, potentially the effect of an extreme drop in the effective population size. A similar pattern was observed in recently bottlenecked ape populations, which exhibited an increased load of large segregating copy number polymorphisms. These data suggest demography may play an important role in the structure of hominid genomes contributing to patterns of gene gain and loss throughout recent evolution. One possibility is the burst of segmental duplication activity in the ancestral lineage of African great apes was the result of a severe reduction in population size. Yet, these analyses were only able to calculate average rates of duplication and deletion along each of the branches surveyed—the exact timing of each of the duplication and deletion events was not determined.

5.2 Expression trajectories of duplicated genes

While this thesis describes gene duplication and loss events occurring both between species and segregating among individuals of the same species, the functional import of these copy number changes is unclear. To follow up on the impact of copy number polymorphisms currently segregating in the human lineage, a number of RNAseq datasets assessing expression in lymphoblastoid cell lines (LCLs) can be leveraged (Pickrell et al. 2010; Montgomery et al. 2010; Cheung et al. 2010). Exploring individuals sequenced in the Pilot Phase of the 1000 Genomes Project (1000 Genomes Project Consortium 2010) and using LCL-derived RNAseq data from European and Yoruban HapMap individuals (Montgomery et al. 2010; Pickrell et al. 2010), 55 segmentally duplicated genes and 9 genes in the unique portion of the genome were found to show significant correlations between copy number and expression (FDR of 10%, Sudmant, unpublished). For example, the *GSTM1* gene encoding glutathione S-transferase and the *UGT2B17* gene encoding uridine di-phospho-glucuronosyl-transferase each show highly significant correlations with copy number **(Figure 5.2.1)**. Interestingly, in addition to significant positively correlated associations between copy number and expression level, 7 of the 55 segmentally duplicated genes identified demonstrated negatively correlated associations with copy number, suggesting perhaps repeat-induced silencing might be acting to reduce the expression levels of some highly duplicated transcripts.

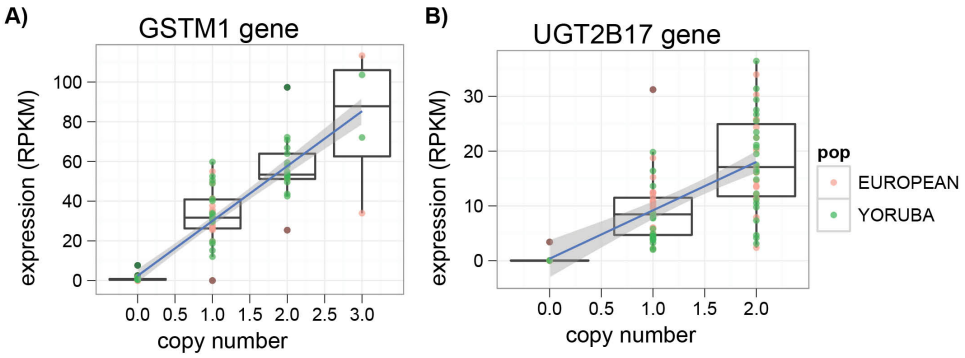


Figure 5.2.1. Transcript abundance versus copy number.

Copy number of *GSTM1* and *UGTB17* and associated expression levels in LCLs for European and Yoruban individuals. *GSTM1* encodes for the glutathione S-transferase protein, a drug detoxification enzyme. *UGT2B17* encodes for uridine di-phospho-glucuronosyl-transferase, which metabolizes steroids and is associated with risk for osteoporosis.

Such behavior has been demonstrated in the D4Z4 macro-satellite array on chromosome 4, each unit of which encodes a copy of the DUX4 protein. Normally 11-100 copies of the D4Z4 repeat are present, however, contraction to 10 or fewer repeats results in hypomethylation and decreased repressive chromatin (Lemmers et al. 2010; 2012), allowing expression of the *DUX4* gene. On specific genetic backgrounds this manifests as facioscapulohumeral dystrophy (FSHD).

At least seven additional macro-satellite repeats exist and are actively transcribed, four encoding for protein products (Schaap et al. 2013). Each of these repeat elements varies in copy among humans and it would be interesting to test if they too exhibited a repeat-induced silencing signature of negative correlation between copy number and transcript level. One barrier to this experiment and the above analysis is the current restriction of datasets to LCLs. Ideally, a comprehensive resource, including high-coverage DNA sequence in addition to RNAseq data from a host of tissues, would be required to fully explore the effects of copy number variation on gene expression. One potential resource to exploit would be the data generated by the ongoing GTEx project (GTEx Consortium 2013). The project will ultimately consist of ~900 donor individuals collected to ensure a low postmortem interval and will include samples of 47 different tissue collection sites throughout the body. The project does not include genome sequencing, thus, performing deep genome sequencing would be a valuable contribution to this resource.

While leveraging the data from the GTEx resource in combination with whole-genome sequencing would facilitate an in-depth study of the relationship between copy number variation and gene-expression dosage, the GTEx resource could also be used to more generally explore the evolutionary trajectories of recently duplicated genes (**Figure 5.2.2)**.

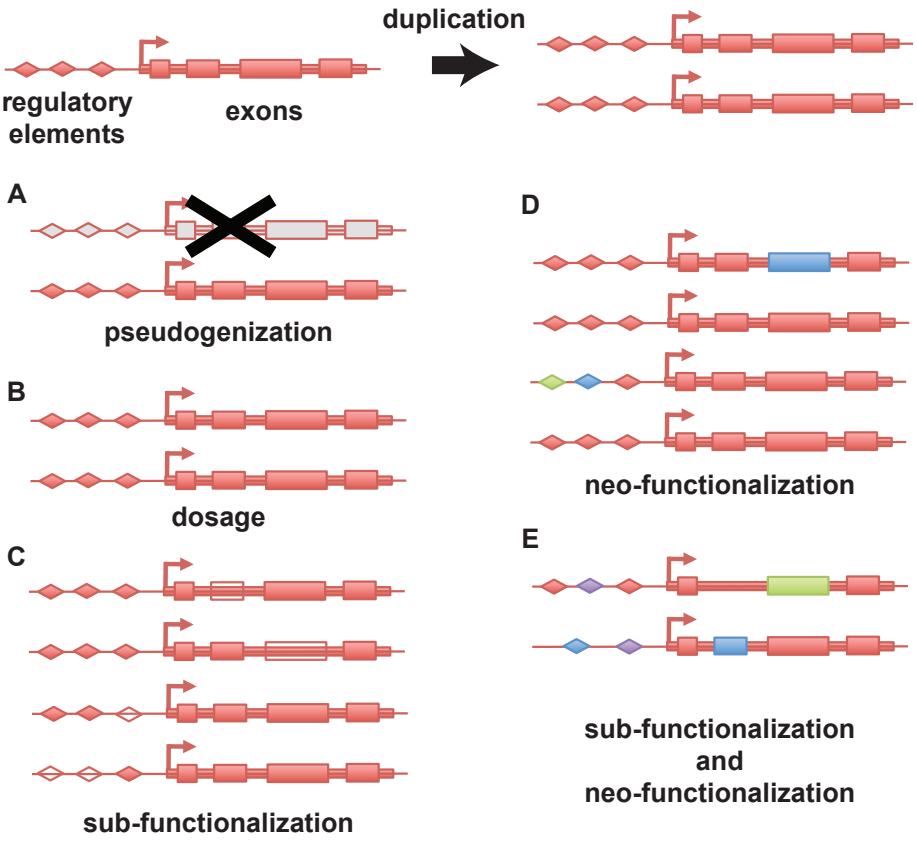


Figure 5.2.2. The fates of duplicated genes.

Following duplication a gene can undergo a number of fates—most commonly death (A). However, the duplication of functional substrates also allows for various other fates. A gene may increase its overall dosage as a result of its additional copies (B), portions of the coding or regulatory region of the ancestor or progenitor may be lost, yet the ancestral function remains with the redundancy of the other copy (C), the ancestor or progenitor may develop new regulatory programs or functions (D), or some combination of the above (E).

Exploring the functional trajectories of recently duplicated genes in the hominid lineage is important to study in light of the burst of segmental duplications in the African great ape ancestor (Marques-Bonet et al. 2009), but also due the unique interspersed nature of hominid duplications. Interspersed duplications are less likely to vary in copy or undergo homogenizing gene conversion events allowing them to evolve more independently. Specifically, to test for neo- or sub-functionalization of the regulatory programs of recently duplicated humans genes, SUNs can be utilized to track the expression patterns of paralogous transcripts. Using data from Wang and Sandberg (Wang et al. 2008a) and the Illumina Body Map Atlas 2.0, numerous genes with vastly different, paralog-specific expression patterns were observed, for instance, the *GOLGA* gene, for which the reference genome includes >25 paralogous gene models (**Figure 5.2.3**). The goal of such a project

would be to first construct a set of gene models for duplicated genes, identifying each of the different expressed isoforms of the paralogs. Next, the tissue-specific patterns of expression would be catalogued identifying those paralogs that appear to show distinct tissue-specific patterning. Finally, generating a diverse panel of tissue-specific RNAseq data from chimpanzees and macaques would provide information as to the complete evolutionary trajectory of these genes.

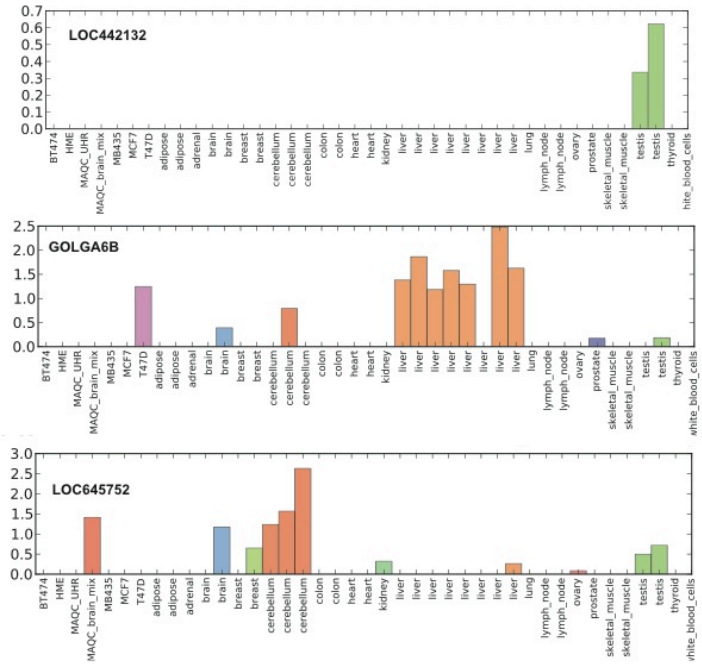


Figure 5.2.3. Expression of *GOLGA* paralogs.

The expression patterns of three different paralogs of *GOLGA* exhibit vastly different tissue specificity.

A shortcoming of such an analysis, however, is that it relies on SUNs in coding regions of genes. This is problematic as exons demonstrate a paucity of SUNs compared to introns. A potential solution to this problem would be to analyze hetero-nuclear (pre-spliced) transcripts so that SUNs in intronic sequence could be assayed. Though an alternative approach would be to further develop the map of SUNs to include higher-order nucleotide identifiers. For example, DUNs, or doubly unique nucleotide identifiers, are present in exactly two segmental duplications, TRUNs, three, and so on. A genome-wide map of these higher order identifiers, or, more specifically, the K-mers tagging them, would be a powerful way to assign reads to duplicated transcripts and construct a graph of all of the potential read-to-transcript relationships. From such a graph, the most parsimonious assignment of

reads to transcripts could be inferred, and thus the most likely paralog-specific transcript abundances.

5.3 Further insights from ancient genomes

The conclusions made in this thesis are largely derived from analyzing the genomes of extant hominid species: humans, chimpanzees, gorillas and orangutans. However, the ability to sequence the genomes of extinct hominids provides a wealth of additional information on the history of our lineage. For instance, while Chapter 2 of this thesis described the identification of 23 human-specific gene duplications with the advent of higher sequencing coverage and the complete genome of the archaic Denisova individual, Chapter 4 describes *BOLA2* as being a recent human-specific duplication not present in either nonhuman primates or the Denisovan.

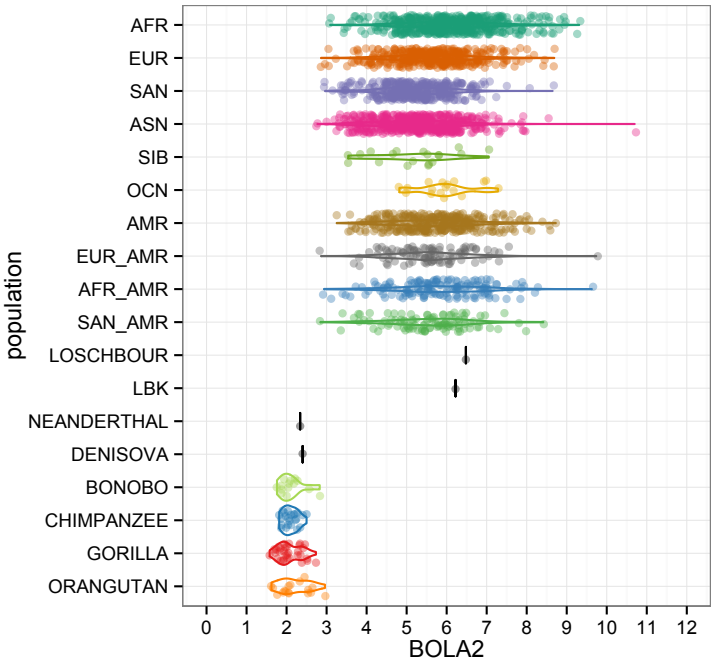


Figure 5.3.1. *BOLA2* copy number in 2641 hominids.

Copy number shown for *BOLA2* in 2,641 individuals from various hominid species and populations. The Neanderthal and Denisova each represent species extinct ~30 kya while the Loschbour and LBK genomes are ~8,000 and ~7,500 years old, respectively.

More recently, the availability of a complete Neanderthal genome confirms that *BOLA2* is indeed a modern human-specific duplication (**Figure 5.3.1**). Two additional ancient genomes, the Loschbour, an ~8,000-year-old pre-Neolithic hunter-gatherer, and the LBK, a ~7,500-year-old Neolithic farmer exhibit the same *BOLA2* copy number as contemporary modern humans demonstrating that *BOLA2*

has likely been a fixed human duplication for some time. Indeed, analysis of 2,641 hominid genomes demonstrates *BOLA2* to be duplicated in every one of 2,553 human genomes while diploid in all 88 archaic and nonhuman primate genomes assessed (**Figure 5.3.1**). As mentioned in Chapter 4, *BOLA2* is of particular interest as it lies near the breakpoint of the 16p11.2 microdeletion, which is associated with developmental delay, intellectual disability, and autism (Kumar et al. 2008; Weiss et al. 2008). This phenotype, combined with the unique human-specific nature of this gene, makes it an intriguing candidate to follow up for both functional analysis and potential disease association.

Archaic genomes such as the Denisova and Neanderthal in concert with more recent ancient genomes also provide further specificity to date ancient duplication events (**Table 5.3.1**). For example, comparing the Denisova and Neanderthal genomes to 20 high-coverage diverse humans highlights that since the divergence of humans from Neanderthals and Denisovans, only 55.3 kbp of duplication have occurred, intersecting three genes (**Table 5.3.2**).

Table 5.3.1: Summary of the number of duplicated base pairs along and shared between lineages. Copy-corrected duplicated bp refers to the number of duplicated bp corrected for over-counting segmental duplications represented in the human reference. H,N,D,C,G and O stand for Human, Neandertal, Denisova, Chimpanzee, Gorilla and Orangutan, respectively.

Lineage	Duplicated bp	Copy-corrected duplicated bp	Events
H-D	67520	66194	7
H	75207	53388	3
D	76585	139355	8
N	181101	279646	6
H-N	271110	219981	19
C-H-D-N	4791330	4341010	297
H-D-N	6450328	5679030	306
C	8039549	12086082	1751
G	14849699	39623204	1562
O	19458418	33373255	3378
G-C-H-D-N	25275862	30800081	1153
O-G-C-H-D-N	69582673	273666449	5955

Table 5.3.2: Lineage-specific segmental duplications along each of the terminal branches and genes encompassed.

Locus	length	lineage	genes	Genotypes			
				Modern Humans (median)	Denisova	Altai	Mezmaiskaya
chr12:122079832-122087495	**7663**	Altai-Neandertal	ORAI1	2	2	4	3
chr12:132295389-132391442	**96053**	Altai-Neandertal	MMP17,ULK1	2	2	4	2
chr19:9284044-9291195	**7151**	Altai-Neandertal		2	2	4	4
chr20:281880-290717	**8837**	Altai-Neandertal		2	2	10	9
chr3:12639069-12641393	**2324**	Altai-Neandertal	RAF1	2	2	7	3
chr6:95473793-95532866	**59073**	Altai-Neandertal		2	2	3	2
chr11:39901956-39909545	**7589**	Denisova		2	4	2	2
chr1:161272681-161274838	**2157**	Denisova	MPZ	2	4	2	2
chr12:49894191-49897733	**3542**	Denisova	SPATS2	2	4	2	2
chr19:55302094-55315197	**13103**	Denisova	KIR3DP1,KIR2DL4	2	4	2	2
chr2:48781187-48787915	**6728**	Denisova		2	3	2	2
chr4:68542692-68577288	**34596**	Denisova	UBA6,LOC550112	2	3	2	2
chr4:68579206-68581585	**2379**	Denisova	LOC550112	2	3	2	2
chr7:140872574-140879065	**6491**	Denisova	LOC100131199	2	6	2	2
chr1:108924526-108990191	**65665**	Modern Human		4	2	2	2
chr16:30200098-30206185	**6087**	Modern Human	CORO1A, LOC606724,BOLA2	6	2	2	2
chr2:87417089-87420544	**3455**	Modern Human		4	2	2	2

The paucity of recent duplications highlights the possibility that the bulk of duplications along the human lineage may have occurred long before our divergence from Neanderthals. As suggested in Chapter 4 of this thesis, the accretion of fixed duplications in hominid genomes may be associated with population bottlenecks. If this is the case, perhaps many of the human-specific duplications in our genome, including *BOLA2,* occurred ~500 kya to 1 mya during a pre-African diaspora bottleneck (**Figure 5.3.2**).

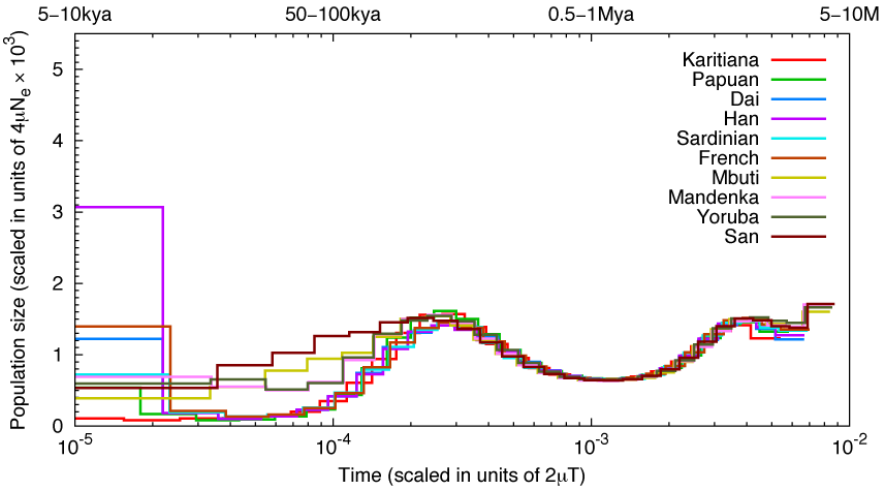


Figure 5.3.2. PSMC of the human lineage.

Previous to the out-of-Africa bottleneck, the ancestral human population appears to have undergone a reduction in effective population size ~500 kya to 1 mya.

Ideally, more archaic genomes of increasingly ancient origin will be sequenced in the near future allowing precise delineation of the timing of duplication events. However, the upper limit on the age of an ancient DNA sequence appears to be ~1 million years old and is highly dependent on the temperature at which the specimen has been maintained (Millar and Lambert 2013). Another way to date recent human-specific duplications would be to attempt to calculate their coalescent times from the accumulated sequence diversity of duplication blocks. This approach would be challenging for a number of reasons, including the complex architecture and polymorphism of these duplication blocks in addition to the presence of non-allelic gene conversion. If these processes were properly accounted for however, it would be possible to test if there is any relationship between this human population bottleneck and an accumulation of duplications.

An upper bound of ~1 million years on the age of a potential hominid genome sequence conceivably means the genome of a *Homo erectus* specimen may someday be available; however, perhaps a more reasonable expectation is that additional sequences of Neanderthals and ancient modern humans will be made available to provide insight into the population diversity of these individuals. Utilizing this resource, it will be important to construct a full catalogue of the genomic changes specific to humans and to Neanderthals, particularly the coding changes and potential regulatory changes. To determine the potential functional importance of truly human-specific genetic changes, they could first be screened in a large number of human individuals using publicly available resources such as the 1000 Genomes sequences or, more comprehensively, in a panel using targeted sequencing techniques such as molecular inversion probes (MIPs) (Turner et al. 2009). A subset of the mutations outside of the range of normal human variation and selected to most likely be functional could then be engineered into iPS cell lines using transcription activator-like effector nucleases (TALENs) (Zhang et al. 2011). RNAseq experiments on these cell lines as they are induced into various different cell types would provide insight into the potential functional consequences of these mutations and, ultimately, into the specific nucleotide changes that shaped our species.

5.4 Exploring human diversity

Throughout this thesis, data from the 1000 Genomes Project has been leveraged to explore human copy number diversity. Though phases I (1000 Genomes Project Consortium et al. 2012), II and III of the project are sufficiently improved from the initial Pilot Project (1000 Genomes Project Consortium 2010) in both the sequence coverage of individual genomes and the diversity of samples collected (2,577 individuals from 26 populations collectively), the data are far from ideal. Specifically, increased sequence coverage and a more uniform worldwide sampling of individuals would greatly improve our ability to explore variation. Recently, a resource of ~300 diverse human individuals has been generated with median sequence coverage of ~40X (**Figure 5.4.1** – Reich and colleagues https://www.simonsfoundation.org/life-sciences/project-awards/).

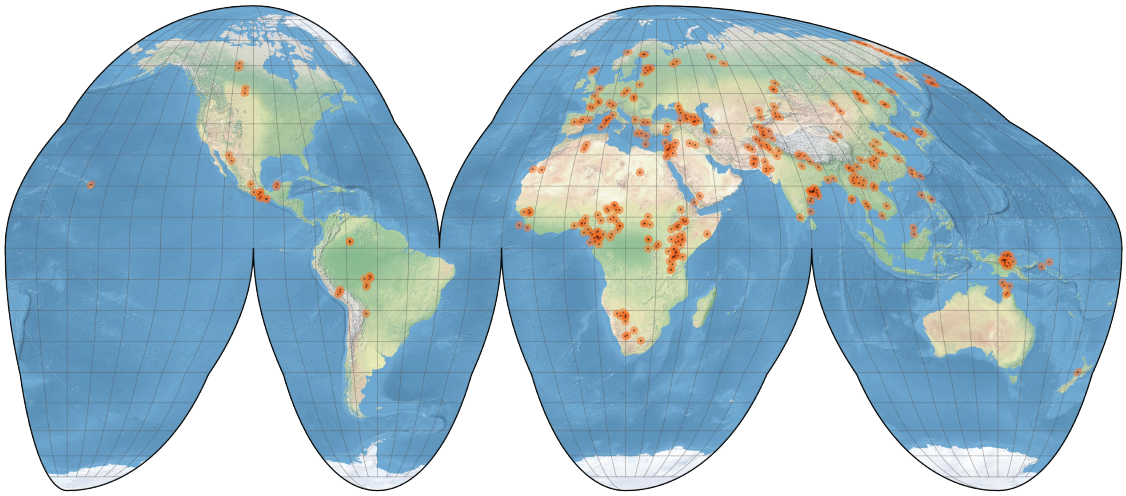


Figure 5.4.1. The C-team sequencing project.

Orange circles represent sampling locations of individuals across the world.

This resource provides a unique opportunity to assess human genome variation on a per-individual basis, unbiased by the underlying frequency of variants. In addition to cataloguing a more complete set of rare structural variants and duplications across these individuals, their high coverage makes it possible to identify new paralogous variants and single nucleotide variants within duplications. By anchoring reads on known SUNs, the reads can be assigned to specific duplications and novel variants in their paired ends identified. This approach potentially allows new duplication haplotypes to be discovered that may appear identical between individuals in their copy number state but have arisen independently. Known

alternate haplotypes, incomplete portions of the human reference genome that have since been assembled (for example, portions of 1q21 (Dennis et al. 2012)), and sequence that appears to be lost along the human lineage, can also be explored by remapping the raw shotgun reads to these data.

Despite the power conferred by sequencing to high coverage a set of diverse human genomes, some regions will unfortunately remain out of the reach of short-read Illumina sequencing. For example, the aforementioned D4Z4 repeat array of *DUX4* is not captured well by Illumina sequencing due to its extreme GC content. Additionally, many repetitive regions, such as centromeric α-satellite and β-satellite sequence, mobile elements and complex duplication architectures, are intractable by short-read sequencing. Inversions and SUNs lying in new duplication haplotypes are also challenging to assess. To ultimately tackle these regions, technologies beyond 'second generation' sequencing must be employed. One such technology, PacBio (Eid et al. 2009), boasts read-lengths in excess of 30 kbp (mean read-length ~8 kbp), though, due to the low quality of reads, high coverage is required to construct assemblies. Generating a resource of multiple human genomes sequenced to high coverage by a long-read technology would enable an almost complete characterization of human genetic diversity over much of the genome, though likely still excluding some of the more complex duplications. In addition to developing a better understanding of genetic diversity among humans, high-coverage, long-read, full-genome sequencing of chimpanzee and gorilla genomes would additionally provide incredible insight into primate evolution.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nature Publishing Group* **12**: 363–375.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**: 745–750.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* **18**: 2555–2566.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**: 552–564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res* **17**: 1505–1519.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M. 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Mol Biol Evol* **28**: 615–623.

Boettger LM, Handsaker RE, Zody MC, McCarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**: 881–885.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, et al. 2009. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *Science* **325**: 318–321.

Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* **40**: 1466–1471.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16237444&retmode=ref&cmd=prlinks.

Caldecott JO, Miles L. 2005. *World Atlas of Great Apes And Their Conservation*. Univ of California Press.

Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L, et al. 2011. Population-genetic properties of differentiated human copy-number polymorphisms. *Am J Hum Genet* **88**: 317–332.

Campbell G, Kuehl H, N'Goran KP, Boesch C. 2008. Alarming decline of West African chimpanzees in Côte d'Ivoire. *Current Biology* **18**: R903–R904.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923–935.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68**: 444–456.

Chen LT. 1999. A Candidate Target for G Protein Action in Brain. *Journal of Biological Chemistry* **274**: 26931–26938.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93.

Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS. 2010. Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol* **8**: e1000480.

Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, et al. 2001. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**: 953–958.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Consortium TIH3, investigators P, leaders PC, group MW, QC GA, discovery E3SAS, analysis CNVTA, analysis P, analysis LFV, analysis LDAHS, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.

Consortium TWTCC. 2010. Genome-wide association study of CNVs in16,000 cases of eight common diseasesand 3,000 shared controls. *Nature* **464**: 713–720.

Coop G, Przeworski M. 2006. An evolutionary view of human recombination. *Nat Rev Genet* **8**: 23–34.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* **40**: 1199–1203.

Dai L, Bellugi U, Chen XN, Pulst-Korenberg AM, Järvinen-Pasley A, Tirosh-Wagner T, Eis PS, Graham J, Mills D, Searcy Y, et al. 2009. Is it Williams syndrome? GTF2IRD1implicated in visual-spatial construction and GTF2Iin sociability revealed by high resolution arrays. *Am J Med Genet* **149A**: 302–314.

Darwin C. 1909. *The Origin of Species*.

Darwin C, Wallace AR. 1958. *On the Tendency of Species to Form Varieties*.

Dennehey BK, Gutches DG, McConkey EH, Krauter KS. 2004. Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83**: 493–501.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M,

Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* **17**: 1266–1277.

Edelmann L, Prosnitz A, Pardo S, Bhatt J, Cohen N, Lauriat T, Ouchanov L, Gonzalez PJ, Manghi ER, Bondy P, et al. 2006. An atypical deletion of the Williams-Beuren syndrome interval implicates genes associated with defective visuospatial processing and autism. *Journal of Medical Genetics* **44**: 136–143.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

Elsea SH, Girirajan S. 2008. Smith–Magenis syndrome. *Eur J Hum Genet* **16**: 412–421.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**: 1629–1631.

Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui L-C. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet* **11**: 1987–1995.

Fan Y, Linardopoulou E, Friedman C, Williams E, Trask BJ. 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res* **12**: 1651–1662.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**: E207.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* **21**: 1626–1639.

Goidts V, Szamalek JM, de Jong PJ, Cooper DN, Chuzhanova N, Hameister H, Kehrer-Sawatzki H. 2005. Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res* **15**: 1232–1242.

Gonder MK, Locatelli S, Ghobrial L, Mitchell MW, Kujawski JT, Lankester FJ, Stewart C-B, Tishkoff SA. 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proc Natl Acad Sci USA* **108**: 4766–4771.

Gonder MK, Oates JF, Disotell TR, Forstner MR, Morales JC, Melnick DJ. 1997. A new west African chimpanzee subspecies? *Nature* **388**: 337.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y. 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–722.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330–336.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577.

Hageman GS, Hancox LS, Taiber AJ, Gehrs KM, Anderson DH, Johnson LV, Radeke MJ, Kavanagh D, Richards A, Atkinson J, et al. 2006. Extended haplotypes in the complement factor H (CFH) and CFH-related (CFHR) family of genes protect against age-related macular degeneration: Characterization, ethnic distribution and evolutionary implications. *Ann Med* **38**: 592–604.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. 2000. AIDS as a zoonosis: scientific and public health implications. *Science* **287**: 607–614.

Hey J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol Biol Evol*

27: 921–933.

Hirayasu K, Ohashi J, Tanaka H, Kashiwase K, Ogawa A, Takanashi M, Satake M, Jia GJ, Chimge N-O, Sideltseva EW, et al. 2008. Evidence for Natural Selection on Leukocyte Immunoglobulin-like Receptors for HLA Class I in Northeast Asians. *The American Journal of Human Genetics* **82**: 1075–1083.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* **3**: e7.

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**: 1270–1278.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**: i350–7.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–951.

Innan H. 2003. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc Natl Acad Sci USA* **100**: 8793–8798.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**: 1361–1368.

Jobling MA, Hollox E, Hurles M, Tyler-Smith C, Kivisild T. 2013. *Human Evolutionary Genetics*. Garland Science.

Junker J, Blake S, Boesch C, Campbell G, Toit LD, Duvall C, Ekobo A, Etoga G, Galat-Luong A, Gamys J, et al. 2012. Recent decline in suitable environmental conditions for African great apes ed. M. Bode. *Diversity and Distributions* **18**: 1077–1091.

Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE. 2012. Detection of structural variants and indels within exome data. *Nat Methods* **9**: 176–178.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, et al. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**: 523–526.

Kehrer-Sawatzki H, Sandig C, Chuzhanova N, Goidts V, Szamalek JM, Tänzer S,

Müller S, Platzer M, Cooper DN, Hameister H. 2005a. Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). *Hum Mutat* **25**: 45–55.

Kehrer-Sawatzki H, Sandig CA, Goidts V, Hameister H. 2005b. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and the homologous chromosome 12 in humans. *Cytogenet Genome Res* **108**: 91–97.

Kehrer-Sawatzki H, Szamalek JM, Tänzer S, Platzer M, Hameister H. 2005c. Molecular characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to human chromosome 9. *Genomics* **85**: 542–550.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**: 894–897.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin J-J, Hänni C, Fortea J, la Rasilla de M, et al. 2007. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. *Current Biology* **17**: 1908–1912.

Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH, Dobyns WB, et al. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17**: 628–638.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**: 4181–4193.

Lalueza-Fox C, Gilbert MTP. 2011. Paleogenomics of Archaic Hominins. *Current Biology* **21**: R1002–R1009.

Lalueza-Fox C, Rompler H, Caramelli D, Staubert C, Catalano G, Hughes D, Rohland N, Pilli E, Longo L, Condemi S, et al. 2007. A Melanocortin 1 Receptor Allele Suggests Varying Pigmentation Among Neanderthals. *Science* **318**: 1453–1455.

Lemmers RJLF, Tawil R, Petek LM, Balog J, Block GJ, Santen GWE, Amell AM, van der Vliet PJ, Almomani R, Straasheijm KR, et al. 2012. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* **44**: 1370–1374.

Lemmers RJLF, van der Vliet PJ, Klooster R, Sacconi S, Camaño P, Dauwerse JG, Snider L, Straasheijm KR, van Ommen GJ, Padberg GW, et al. 2010. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**: 1650–1653.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, Roe B, Rocchi M, Eichler EE. 2003a. Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol* **4**: R50.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529–533.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE. 2003b. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* **13**: 347–357.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates Inc.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38**: 86–92.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PIW, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**: 1166–1174.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**: 216–219.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**: 222–226. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22936568&retmode=ref&cmd=prlinks.

Millar CD, Lambert DM. 2013. Ancient DNA: Towards a million-year-old genome. *Nature* **499**: 34–35.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65.

Mirnics K, Middleton FA, Marquez A, Lewis DA, Levitt P. 2000. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* **28**: 53–67.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773–777.

Morin PA, Moore JJ, Chakraborty R, Jin L, Goodall J, Woodruff DS. 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science* **265**: 1193–1201.

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**: 1344–1356.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**: 1113–1118.

Nsubuga AM, Holzman J, Chemnick LG, Ryder OA. 2009. The cryptic genetic structure of the North American captive gorilla population. *Conserv Genet* **11**: 161–172.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**: 18–23.

Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* **4**: 20–28.

Ortiz M, Kaessmann H, Zhang K, Bashirova A, Carrington M, Quintana-Murci L, Telenti A. 2008. The evolutionary history of the CD209 (DC-SIGN) family in humans and non-human primates. *Genes Immun* **9**: 483–492.

Ovchinnikov IV, Götherström A, Romanova GP, Kharitonov VM, Lidén K, Goodwin W. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**: 490–493.

Pallanck L, Ordway RW, Ganetzky B. 1995. A Drosophila NSF mutant. *Nature* **376**: 25.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* **91**: 275–292.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Res* **18**: 1698–1710.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–772.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**: 207–211.

Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution* **61**: 3001–3006.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature*.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 1–5.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencingplatforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 1–1.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–1060.

Rochette CF, Gilbert N, Simard LR. 2001. SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum Genet* **108**: 255–266.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.

Schaap M, Lemmers RJ, Maassen R, van der Vliet PJ, Hoogerheide LF, van Dijk HK, rk NB, de Knijff P, van der Maarel SRM. 2013. Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics* **14**: 1–1.

Sharp AJ, Cheng Z, Eichler EE. 2006. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7**: 407–442.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, et al. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* **40**: 322–328.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145.

Shimada MK, Kim C-G, Kitano T, Ferrell RE, Kohara Y, Saitou N. 2005. Nucleotide sequence comparison of a chromosome rearrangement on human chromosome 12 and the corresponding ape chromosomes. *Cytogenet Genome Res* **108**: 83–90.

Shinawi M, Schaaf CP, Bhatt SS, Xia Z, Patel A, Cheung SW, Lanpher B, Nagl S, Herding HS, Nevinny-Stickel C, et al. 2009. brief communications. *Nat Genet* **41**: 1269–1271.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, et al. 2001. Assembly of microarrays for

genome-wide measurement of DNA copy number. *Nat Genet* **29**: 263–264.

Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res* **14**: 2209–2220.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37**: 129–137.

Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* **455**: 232–236.

Stoneking M, Krause J. 2011. Learning about human population history from ancient and modern genomes. *Nature Publishing Group* **12**: 603–614.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* **28**: 289–301.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. 2009. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**: 315–316.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732.

Urban TJ, Weintrob AC, Fellay J, Colombo S, Shianna KV, Gumbs C, Rotger M, Pelak K, Dang KK, Detels R, et al. 2009. correspondence. *Nature Medicine* **15**: 1110–1112.

Varki A, Geschwind DH, Eichler EE. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat Rev Genet* **9**: 749–763.

Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, et al. 2011. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* **21**: 1640–1649.

Wallace AR. 1869. *The Malay Archipelago*.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008a. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al. 2008b. The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS Biol* **4**: e52.

Wang X, Mitra N, Secundino I, Banda K, Cruz P, Padler-Karavani V, Verhagen A, Reid C, Lari M, Rizzi E, et al. 2012. Specific inactivation of two immunomodulatory SIGLEC genes during human evolution. *Proc Natl Acad Sci USA* **109**: 9935–9940.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MAR, Green T, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**: 667–675.

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, et al. 2008. Adaptive Evolution of UGT2B17 Copy-Number Variation. *The American Journal of Human Genetics* **83**: 337–346.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865–2871.

Yue Y, Grossmann B, Tsend-Ayush E, Grützner F, Ferguson-Smith MA, Yang F, Haaf T. 2005. Genomic structure and paralogous regions of the inversion breakpoint occurring between human chromosome 3p12.3 and orangutan chromosome 2. *Cytogenet Genome Res* **108**: 98–105.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science* **215**: 1525–1530.

Zhang F, Cong L, Lodato S, Kosuri S, Church GM. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature*.

Zody MC, Jiang Z, Fung H-C, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, et al. 2008. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* **40**: 1076–1083.

Appendix A - Supplementary Material for Chapter 2

Supporting online material for

Diversity of Human Copy Number Variation and Multicopy Genes

Peter H. Sudmant*, Jacob O. Kitzman*, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, Laurakay Bruhn, Jay Shendure, 1000 Genomes Project, Evan E. Eichler

TABLE OF CONTENTS

MATERIALS & METHODS 92
SUPPORTING TEXT 93
1. COPY NUMBER ESTIMATION 93
1.1. Data and methods 93
1.2. False discovery estimation 94
1.3. Digital comparative genomic hybridization and the human CNV landscape 96
2. EXPERIMENTAL VALIDATION 97
2.1. SNP microarray validation 97
2.2. FISH validation 98
2.3. Quantitative PCR validation 99
2.4. Array CGH validation 99
2.5 Copy number aware array comparative genomic hybridization 100
3. GENE ANALYSIS 101
3.1. Missing human genes 101
3.2. Most variable genes 101
3.3. Highly population stratified genes 102
3.4. Human-specific gene duplications 102
4. SINGLY UNIQUE NUCLEOTIDE (SUN) ANALYSES 103
4.1. SUN discovery and genotyping 103
4.2. Validation of paralog-specific copy number estimation 106
4.3. Paralogous regions of genomic variation 108
4.4. Gene family analyses 109
4.5. Validation of novel structural variants within segmental duplications 111
4.6. Gene conversion analysis 113
SUPPORTING FIGURES 116
SUPPORTING TABLES 177
SUPPORTING REFERENCES 195

MATERIALS & METHODS

Short read sequencing data were obtained from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra and Table S1) and mapped to a masked reference genome (Build36) masked with RepeatMasker and Tandem Repeat Finder. Mapping was performed with mrsFAST (*S1*) allowing for up to 2 mismatches. Reads exceeding 36 bp in length were truncated to 36 bp or divided into non-overlapping 36 bp constituents. Read depth profiles were then constructed independently for each sequencing library, corrected for G+C bias introduced during library construction (*S2*) using a smoothed correction factor, and combined. Copy number prediction was performed by regression against a standard curve fit to regions of known copy. Individual loci were genotyped by computing the median copy number across nonoverlapping windows of 1000 unmasked bases each. We constructed a map of 4.06×10^6 paralog-specific tags, or SUNs (Singly Unique Nucleotides) as follows: briefly, the human reference genome (Build36) was divided into its constituent k-mers (k=30) for both forward and reverse complement DNA. Kmers were screened for uniqueness by mapping to the reference with the mrFAST aligner (*S3*). We rejected k-mers with more than one perfect match in the genome as well as those with highly repetitive content. Within segmental duplications, we defined SUNs as the set of paralogous sequence variants within the underlying pairwise alignment between segmental duplications (Table S2).

We used fluorescence in situ hybridization (FISH), quantitative PCR (qPCR), and array CGH to validate our read depth-based copy number estimates.

FISH Analysis Methods: Metaphase spreads were obtained from 32 lymphoblast cell lines from Coriell Cell Repository, Camden, NJ, and one YH (Han Chinese). FISH experiments were performed using fosmid clones (Table S3) directly labeled by nick-translation with Cy3-dUTP (PerkinElmer) as previously described (*S4*), with minor modifications. Briefly, 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, and 3µg sonicated salmon sperm DNA, in a volume of 10µL. Posthybridization washing was at 60°C in 0.1xSSC (three times, high stringency). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI and Cy3 fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images was performed using Adobe Photoshop software. A minimum of 50 interphase cells were scored for each probe.

Quantitative PCR assays: Quantitative PCR assays are listed in Table S4 using primer sequences listed in Table S5. Primer oligos were obtained from Operon (Huntsville, AL). Copy number estimates were obtained by the $\Delta\Delta C_T$ method and normalized using a primer set directed against a diploid control, the human albumin gene. Reactions were carried out in quadruplet using SYBR Green I Master Mix on a Roche LightCycler 480 thermocycler.

Array CGH: Copy number-aware CGH was performed by hybridizing two differentially Cy-labeled samples to a custom, high-density oligonucleotide 4x180K CGH microarray (Agilent, Santa Clara, CA) targeted at 17q21.31 with a density of 1 probe per 100bp. Labeling, hybridization, scanning,

and data processing were performed as directed by the manufacturer. For the reference sample, an individual having intermediate copy in the target region was chosen (NA19240).

Additional validation was performed by comparison to SNP-chip based events from the same individuals called on the Affymetrix 6.0 and Illumina 1M DUO SNP platforms and CNV events and genotypes from the NimbleGen 41M probe and Agilent single-channel CGH platforms (*S5-8*). Paralog-specific copy number estimation was validated by comparison to fully sequenced clones corresponding to deletions within segmental duplications (*S9*), by paralog-specific quantitative PCR, and by Illumina short-read fosmid clone inset sequencing.

SUPPORTING TEXT

1. COPY NUMBER ESTIMATION

1.1. Data and methods

Copy number prediction was performed with read depth-based methods adapted from those we previously described (*S3*). A detailed description of the data and modifications to this method are described below.

Data: Sequence data were collected primarily from *Pilot 1* and *Pilot 2* of the 1000 Genomes Project, which includes two trios sequenced to high coverage and 179 individuals sequenced from 1-4X. A high coverage trio, including the published NA18507 genome and the genome of Jay Flatley, was also obtained from Illumina (*http://www.ncbi.nlm.nih.gov/sra/SRP000978, http://www.ncbi.nlm.nih.gov/sra/SRP001050*). In addition, 10 other published genomes of varying coverages were collected: NA10851 (*S10*), AK1 (*S11*), SJK (*S12*), KB1 ABT (*S13*), YH (*S14*), and five individuals from a Human Genome Diversity Panel (*S15*), as well as a collection of great apes including a gorilla, a chimpanzee (capillary sequence segmented into 36-bp fragments to mimic Illumina datasets as previously described, ref. (*S3*)) and a Bornean orangutan (Fig. S1). Additionally, we simulated the reference genome at 6X segmenting the assembly (Build36) into 36-bp fragments.

Read mapping: Short-read sequence data were mapped to the masked human reference genome (Build36) using the software *mrsFAST* (*S1*) with a maximum of two substitution mismatches allowing for no indels. *mrsFAST* returns all locations in the genome at which a specific read can map, given the specified parameters. All reads exceeding 36 base pairs (bp) in length were truncated to 36 bp, or divided into their constituent nonoverlapping 36-bp sequences to eliminate potential mapping biases between genomes sequenced at different read lengths. Using *mrsFAST*, we mapped 35.4 billion reads from this dataset to the human reference genome in one month using a 160 CPU Linux cluster farm. Masking was defined as all base pairs identified by RepeatMasker 3.2.7 and TandemRepeatFinder (*S16*) as defined in the UCSC genome browser (*http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/*). Masking was extended 36 base pairs upstream and downstream of the original mask during construction of read depth maps to account for potential edge mapping effects.

G+C correction: Sequence from each genome library was mapped individually to the human reference genome (Build36) and each library was tested to account for potential biases that may have occurred during library preparation (Fig. S2). We applied a multiplicative G+C correction to every library to account for biases in construction and/or sequencing. This correction was determined by calculating the average read depth across the control diploid regions (Build36 - [Database of Genomic Variants + gaps + genomic super dups]) binned by G+C content. The correction factor k_{gc} was defined as $k_{gc} = \frac{\mu_{total}}{\mu_{gc}}$, where μ_{total} is the total average read depth across all copy number 2 regions of the genome and μ_{gc} is the average read depth across regions of a specific G+C content. μ_{gc} was smoothed by LOESS regression, with piecewise linear regression at high and low G+C outliers. The corrected read depth at any base ***x*** in the genome $d'(x)_{gc}$ is then calculated as $d'(x)_{gc} = d(x)_{gc}k_{gc}$ where $d(x)_{gc}$ is the original depth at base ***x.*** A maximal G+C correction factor of two was enforced.

Quality control: To assess the quality of individual libraries we analyzed the correlation of the (G+C)- corrected read depth with the copy number of regions of experimentally validated fixed copy (Fig. S3).

As the effective coverage of a genome increases, the correlation between read depth and copy increases asymptotically as demonstrated by analyzing this metric on genomes generated from subsampling a single, high coverage genome (Fig. S4). We exploited this relationship as a quality control (QC) metric to judge the quality of each individual library (Fig. S5). Libraries that appeared to deviate from this asymptotic trend of increased correlation were discarded (i.e. all libraries that failed a correlation threshold of 0.85 were omitted). All libraries from a single genome passing QC were then combined by summing their (G+C)-corrected read depths at each base. 177/198 libraries from *Pilot 1* low coverage genomes passed our QC metric.

Copy Number Prediction: Copy number was predicted using a linear regression model based on our read depths versus copy number in regions of known fixed copy number (Fig. S3). We also developed a method to estimate copy number using a simple linear model maximum likelihood approach, as read depths were approximately normally distributed for each integer copy number state (Fig. S6). Using the predicted mean and variance of the Gaussian distributions underlying different copy numbers, a series of models (representing copy 0, 1, 2, 3 and so on) can then be generated to represent the likely distribution of read depths underlying a region of specific copy number. Given a region of interest, each of these models can then be tested to determine by maximum likelihood which model, and subsequently which copy number, best describes the underlying read depth distribution of that particular locus.

1.2. False discovery estimation

To determine an approximate upper bound on our false discovery rate (FDR) and how this changes as a function of sequence coverage and event size, we analyzed the copy number of diploid regions that are copy-number invariant in most humans. We defined these invariant diploid regions (cumulatively ~$1.2x10^9$ non-RepeatMasked bp) by excluding from the whole genome known CNV regions (*S5, 17*) and all segmental duplications (WGAC+WSSD) (*S18, 19*). We limited our analysis to all contiguous autosomal sequence greater than 100 kb in length (i.e., regions of uninterrupted diploid copy). As a stringent measure of FDR, we analyzed all 1-kb windows of unmasked sequence in high coverage genomes (>8X) and found that 97-99% of individual windows in these regions were predicted at copy 2, as expected. However, a dramatic decrease in the fraction of predicted copy 2 regions was observed as coverage decreased for these small regions (Fig. S7), with the copy 2 fraction varying from 72-98%.

At 1-kbp resolution, we predict an FDR of ~10% for genomes of 4X sequence coverage or greater, however, the FDR rapidly decreases as larger regions are considered (Fig. S8). At 3-kb resolution the FDR is <5% in genomes with as little as 1.5X coverage. For regions >10 kb the FDR reduces to <2%. Similar results were observed for data simulated from subsampling a single, high coverage genome to lower coverage (Fig. S9), however, simulated genomes showed slightly better concordance to genomes at similar coverage.

Using these subsampled genomes, we explored factors contributing to false CNV discovery, again considering 1-kbp windows of diploid invariant sequence. False positives favored losses (predicted copy <2) as opposed to gains (predicted copy >2) (Fig. S10). These corresponded with reduced overall coverage in these regions arising from biases in library preparation and sequencing. Interestingly, at low G+C content, calls <2 contributed almost exclusively to the error, whereas at high G+C, there were errors in both directions: this explains the slight excess of deletion calls. These effects are constrained to a small fraction of the genome as a whole: ~73% of the genome lies between 35%-55% G+C content, where we observe high accuracy and little bias in the direction of error.

We estimated the copy number of all genes (discarding those <3kb in length) in individual NA18507 at full coverage (~43X) and in 15 simulated reduced-coverage genomes created by subsampling reads from NA18507 to ~1-25X coverage. Using the full-coverage genome as a gold standard, we computed deviation in copy number predictions for genes in different copy number classes (Fig. S11). As expected, deviation from the full-coverage estimates widened (i.e., accuracy was lower) as coverage decreased. Additionally, because the variance in read depth scales with copy number state, higher copy number states become increasingly difficult to predict within +/- 0.5 copies. Despite this trend, the magnitude of the errors remains small for genes of moderately elevated copy number, even with greatly reduced sequence data. For a genome at 3X coverage, ~96% (1702/ 1764) of genes >3kb in length with copies ranging from 3-10 remain concordant, changing by < +/- 0.5 copies of the full-coverage estimate (Fig. S12). For an 8X coverage genome, >98% of genes >1kb in length stay within this range.

We next evaluated the effect of coverage upon accuracy using two loci validated across a large panel of low and high coverage genomes. We compared read-depth and single channel array-based copy number predictions for the greatly copy-number expanded and highly variable gene *TBC1D3* (Fig. S13A,B). Consistent with the subsampling results, both low (<3X) and high coverage

(>=3X) genomes showed strong correlation between sequencing- and array-based genotypes (r=0.93 and r=0.97, respectively). Another duplicated gene which we extensively validated was *CCL3L1*, a duplicated gene 1.9kb in length. We designed a quantitative PCR assay for this gene and genotyped 150 samples. This demonstrated a very strong correlation of r=0.948 across the entire range of variation (0-14 copies, Fig. S13C). Stratifying the individual genomes by sequencing coverage, we again observed strong correlation with only a modest reduction for low coverage genomes (<=3X depth, r=0.947 with qPCR) versus those at higher coverage (>3X, r=0.958).

1.3. Digital comparative genomic hybridization and the human CNV landscape

In order to detect copy number polymorphism among our sequenced genomes we developed a comparative method similar to array comparative genomic hybridization (CGH), which we term digital CGH (dCGH). Windowed copy number estimations from one genome are subtracted from the windowed copy number estimations of a second genome, or from the mean value of a population of genomes. Similarly to array CGH, this provides a differential estimate to detect regions of gain or loss in a particular genome compared to a reference genome, or a population of genomes (comparable to using a pooled reference experimentally). We utilized this method to detect both common and rare polymorphisms among our genomes.

To detect large copy number variants we performed dCGH on each of our samples versus a high coverage (28X) European individual, NA10852. This experimental procedure was analogous to that implemented with CGH in a recent publication (*S5*). The use of the same reference and test genomes facilitated direct comparison between these two datasets. Events were selected for which at least 15/20 contiguous 1-kb unmasked windows showed a differential of at least 0.5 (positive or negative, corresponding to at least a single copy gain or a loss, respectively). We spanned across gaps in the reference genome, and regions with a total length <20 kb or with fewer than 10 kb of unmasked base pairs were discarded. An average of 338 large CNVs were discovered per individual, however, the mean number of CNVs discovered in individuals with >10X coverage was 298 (n = 14) indicating an increase in the number of false positives in low coverage genomes because of their higher variability (Fig. S14). As this analysis is not optimized for boundary detection and may capture segmental duplication *mirror* effects (see below), it may inflate the number of copy number variable bases.

We merged these calls into a set of 1101 (~202,793 kb) copy number variable regions (CNVRs) and compared them to 1273 (~91,815 kb) calls matching the same size threshold criteria (>20 kb total, >10-kb unmasked sequence) reported by Conrad *et al.* (*S5*), discovered in 41 individuals. 81% of large Conrad CNVR base pairs were detected in our discovery set; after excluding private events from both call sets, this fraction increased to 85%. We focused on 952 large events >50 kb and noted that the majority (55%, 522/952) of these regions overlapped segmental duplications (>20% overlap). As expected, events of increasing size were at progressively lower frequencies (Fig. S15).

A substantial fraction of the observed polymorphism exists at <5% frequency within the population, with 53% (500/952) of events detected in fewer than eight individuals (Fig. S16). Interestingly, we noticed a significant change in the frequency spectrum after stratifying events by

segmental duplication content (Fig. S17). Events containing no segmental duplication were overwhelmingly rare in frequency. For example, 71% (390/546) of such events were observed in less than three individuals, while 55% (302/546) without duplication were private events. In contrast, events containing segmental duplication were found at both high and low frequencies. For example, 16% (66/406) were observed in <=8 individuals, while 42% (170/406) were highly polymorphic across most individuals.

Combining dCGH results with absolute copy number estimations from a population sample allows us to both discover and genotype variation. Additionally, combining paralog-specific copy number maps (see below, section 4) allows us to demarcate the boundaries of structural variants falling within segmental duplications. For example, we identify two distinct, large CNPs (210-kbp and 205-kbp) on chromosome 17q21.31 based on signals of gain from dCGH extending into the more highly duplicated region (Fig. S18). Paralog-specific copy number maps assist in determining the approximate breakpoint locations. We genotyped the resulting regions and observed absolute copy number differences between 2-6 as clearly distinguishable, discrete integer values, which we subsequently validated (Fig. S19 and see section 2.2).

We have formatted these genome-wide copy number maps as heatmaps (e.g., see main text Fig. 1, 2) for visualization across large loci and many individuals, as well as in a UCSC Genome Browser instance (*S20*) for fine-scale visualization. Both resources may be accessed at (*http://hgsv.washington.edu/*).

2. EXPERIMENTAL VALIDATION

We used four independent methods to validate the accuracy of our read depth copy number estimates: fluorescence *in situ* hybridization (FISH), SNP microarrays, quantitative PCR (qPCR) and array CGH. In addition, we demonstrate how our genome-wide assessment of copy number may be used to improve discovery and genotyping using array comparative genomic hybridization.

2.1. SNP microarray validation

CNV detection: To assess the power and accuracy of read depth CNV estimation against currently accepted experimental approaches for detecting copy number variation, we compared our absolute copy number estimates to CNV calls across 253 HapMap samples made on the Illumina 1M Duo SNP genotyping platform. Loci were discovered on the SNP microarray using a Hidden Markov Model or HMM as previously described (*S7*) and classified as a 'gain,' 'loss,' or 'homozygous loss' with respect to the population average (median size 39.6 kb; smallest event 1.6 kb). For each of these loci we took the median regression-based copy number call. We analyzed 152 samples (*Pilot 1*, *Pilot 2* and NA18507 trio) and determined the state of each individual as 'loss' or 'gain' with respect to the average copy number call across all individuals. 109 individuals were common to both sets of samples containing 2270 events with at least 1000 bp of unmasked sequence (503 gains, 1241 losses, and 526 homozygous deletions, respectively). Depending on the size and type of the event, we correctly predicted 94-100% of events using our read depth metric

(Fig. S20). As expected, larger events are easier to detect, but the accuracy also increases as a function of effective read depth (Fig. S21).

CNP Prediction: SNP microarray genotypes and subsampling analysis

To assess our ability to accurately genotype copy number (as opposed to simple gains or losses), we compared our calls to those recently made by another group using Affymetrix 6.0 SNP microarray across 270 HapMap samples (*S21*), of which 114 overlapped with our samples. We selected 1015 of the 1320 regions assayed by McCarroll *et al.* containing at least 1000 bp of unmasked sequence for genotyping (*S21*). Our initial results indicated a 70% genotyping concordance with the SNP microarray copy number estimates (82,373/117,847). We identified 243 regions (n = 243) that showed little correspondence with our genotype calls (<20%). Of these, 94% (228/243) overlapped or were contained entirely within segmentally duplicated portions of the human genome, and encompassed 300 gene models.

After adjustment, we observed 86% concordance (101,631/117,847) between the SNP genotyping calls and our copy number estimates. Considering only genomes with high sequence coverage (seven genomes ranging from 13-29X effective coverage), this increased to 90% (8101/8980). Unadjusted concordance only increased to 73%, further supporting the misassignment of the population average copy number call among the SNP chip-based genotype calls. In contrast, analysis of unique regions of the genome (i.e. not containing segmental duplications) yielded 94% concordance (79,202/84,145) between call sets for unadjusted SNP-chip CN genotypes and 95% concordance (80,086/84,145) with adjusted SNP-chip CN genotypes (Fig. S22). The decreased concordance of genotype calls for segmental duplications stems in large part to the complex nature of these regions. In many cases though, the average copy number across the population for a region fluctuates, as detected in the genotypes from McCarroll *et al* (*S21*). The underlying structure of the region is, thus, much more complex, contributing to the observed discordance.

To further analyze the effect of variable genome sequence read depth on our ability to genotype copy number, we randomly subsampled reads from NA18507 and simulated genome coverage ranging from ~1-30X. We then analyzed the concordance between the McCarroll *et al.* genotype calls (n = 989) and calls adjusted according to the consensus difference observed in high coverage genomes (>13X, n = 6, NA18507 excluded) (Fig. S23a). We observe an asymptotic, rapid increase in the concordance of genotypes as coverage increases, leveling out at ~8X. To explicitly determine why higher coverage enables more accurate genotyping, we plotted the mean and standard deviation of coverage in 1-kb windows tiling a copy number invariant portion of the human genome across individuals of varied coverage (Fig. S23b). Though the variance of read depth increases linearly with increased coverage of a genome, the mean of the read depth increases at a more rapid rate facilitating more robust estimates of the underlying copy of a region.

2.2. FISH validation

We experimentally validated by fluorescence *in situ* hybridization (FISH) 25 predicted copy number differences of the genomes of 33 individuals using cell lines from the same individuals from which the computational predictions were generated (Fig. S24). There were three general

categories of experiments: 1) highly copy number variable regions where individuals with the most extreme copy number difference were selected; 2) large, rare events seen in a single individual; and 3) duplications predicted to be duplicated in all individuals but represented as single copy in the human reference. FISH experiments were perfectly concordant with the computational predictions in 80% (47/59) of the cases, and in 93% (55/59) of the cases the FISH copy numbers differed from the prediction by one copy. In the rest, 7% (4/59) of the predicted copy numbers differed by two or three copies. Of these, three out of four predictions are in the *NPEPPS* region, for which we observed a very strong correlation between copy number predictions and those from the qPCR experiment (r = 0.96) and the single-channel array CGH validation (r = 0.85). This indicates a likely problem in the copy number estimation by FISH. In three highly copy number variable regions, FISH failed to accurately estimate copy number differences between individuals because the signals visualized on interphase nuclei were too numerous to provide an exact number of copies. In one case within the protocadherin gene cluster, FISH could not validate the computational predictions since the duplicated region was too small to be detected by FISH. The observed duplication structure has previously been confirmed in the literature (*S22*) confirming the failure of the FISH assay in this case (Table S3).

2.3. Quantitative PCR validation

We designed two primer pair sets across eight genic regions predicted to be both multi-allelic and highly polymorphic to validate our sequence-based copy number predictions. We additionally designed one paralog-specific qPCR assay. For several loci, we selected the most robust primer set among multiple ones tested. To determine the qPCR dynamic range we attempted to select individuals spanning the full copy number range. Copy numbers were estimated using the $\Delta\Delta C_T$ method normalized using a control directed against the diploid albumin gene. We observed very strong correlations between our copy number predictions and those from the qPCR experiment with R^2 values greater than 0.84 for seven of nine regions and, and only two regions failing all primer sets (Table S4, Figure S25).

As an example, we estimated the copy number of the gene *CCL3L1* in 150 individuals using qPCR to independently verify our sequence-based copy numbers. The *CCL3L1* gene is 1.9 kb long and known to be highly stratified between populations (*S23-27*). We observed a strong concordance between the sequencing-based copy number estimates and those made by qPCR, r = 0.95 (Fig. S26). Furthermore, we demonstrate distinct population stratification of the *CCL3L1* region captured by our sequence-based copy number estimates and confirmed by qPCR.

2.4. Array CGH validation

One of the perceived limitations of array comparative genomic hybridization is its inability to distinguish differences for regions of high copy number. This is due to the individual variability of probe binding characteristics among loci as well as signal saturation within duplicated regions resulting in reduced sensitivity to detect smaller changes (e.g., 9 copies vs. 8 copies being harder to distinguish than 1 copy vs. 2 copies). Using our absolute read depth estimates for loci, we reassessed the accuracy of array-based methods by taking advantage of single-channel intensity

data and leveraging the repeated hybridization of the reference genome over a set of experiments performed on multiple individuals.

We developed a CGH-based approach that leverages our absolute read depth copy number to perform more accurate genotyping of multi-copy CNVs. As a test dataset we used CGH data from (*S5*) in regions determined to be variable. These regions were run through a segmentation algorithm and merged across samples to obtain 7,498 distinct regions. Copy number was predicted from array data by first determining the median signal intensity for a region of diploid copy in the reference, S_{MR}. The reference sample copy number of a test region was then predicted by taking the median reference signal across all probes and arrays assayed, divided by S_{MR}. The copy number state of a sample was then predicted from the $\log_2$ ratio of the test versus the reference in a region multiplied by the inferred reference copy. This allows us to take advantage of the fact that the reference DNA sample used for array CGH is typically hybridized several hundred times, thus, copy number estimates of the median single-channel intensity of a region are likely to be robust.

Sequencing-based copy number predictions were compared to array-based copy number predictions by analyzing the correlation of copy number calls between samples. We noted in many regions that the underlying distribution of signal intensity on the array from the reference sample did not vary greatly with respect to the test samples, precluding this kind of analysis due to the lack of underlying variance in the copy number predictions (Fig. S27a), or inability of the array platform to detect variation at this locus. In order to filter by variance to select regions in which copy number variation was captured by the array, we developed a metric 'rCV'. rCV = CV(sample channel)/CV(reference channel) where CV = std(sample signals)/mean(sample signals). Higher numbers indicate a greater spread of signal intensities in the test samples than in the reference (Fig. S27b). For regions with >1 kb of unmasked sequence, we found that for rCV values greater than 2, 77% (501/648) regions have a correlation >=0.7. For regions with rCV greater than 3, 91% (248/272) have a correlation of >=0.7 (Figs. S28 and S29).

Among regions with sufficient copy number response with respect to the reference as determined by the rCV metric (Fig. S30), we can calibrate array CGH copy number calls with read depth-based calls from a reference genome such as NA10851, which was used as a reference for CNV discovery in a large array CGH study (*S5*) and recently sequenced (*S12*) .

2.5 Copy number aware array comparative genomic hybridization

Our analysis of 159 genomes provides a reference set of DNA and cell lines from which a specific reference DNA sample of known absolute copy may be used to maximize copy number sensitivity. We refer to this pre-selection of a particular reference of known absolute copy number as *copy number aware array CGH*. We demonstrate its utility and the accuracy of our predictions by reanalyzing the 17q21.31 region discussed above. We selected a DNA sample, NA19240, whose copy number across the mosaic region ranged from 1-5 copies (Fig. S31). This particular individual was an ideal reference sample because its copy number range approximated the median copy number for the populations suited. Comparative genomic hybridization using this reference sample was performed on five HapMap individuals with custom, high-density

oligonucleotide 4x180K Agilent chips targeted to the 17q21.31 region with a density of 1 probe per 100 bp. Array CGH showed strong concordance with the computational predictions. We identified three distinct copy number polymorphisms of size 155, 205 and 135 kbp as well as numerous smaller common CNPs within this region. Additionally, using the copy numbers estimated in our reference genome, we were able to infer the true exact copy of the variable loci targeted in the test samples.

3. GENE ANALYSIS

We analyzed the genic portion of the human genome to assay for putative functional copy number variants. We performed copy number prediction on 25,832 nonredundant RefSeq gene models from the UCSC Build36 refFlat table using the median regressed copy number estimates for contiguous 1000-bp windows tiled across the genes. Gene models shorter than 1 kbp were excluded from this analysis (see Section 1.2). Genes were grouped into paralogous gene families using the relationships defined by NCBI Homologene and then further manually curated. As expected, the vast majority of genes were predicted as diploid, with 97% of non-duplicated genes having a median copy number of ~2 across all 159 individuals (median copy >=1.5 and <2.5, fig. S32); 91% of these genes with >10-kb unmasked sequence were determined to be fixed in all 159 genomes analyzed. As expected, simulating copy number estimates for the reference assembly by splitting it into short read-sized fragments (36bp) and calculating copy number as before yielded almost exclusively estimates of diploid copy (fig. S32).

3.1. Missing human genes

We tested whether specific gene families were underrepresented in the human genome by searching for genes with median estimated copy among individuals greater than that in the reference genome Build36 (difference >= 5 copies between median copy in 159 individuals and reference genome) (Table S6; Fig. S33a). 167 gene models were identified corresponding to 44 gene families.

3.2. Most variable genes

We identified the most variable gene families among the 159 individuals analyzed (Table S7) selecting those with a high variance (> 3 copies[2]) across individuals. 260 genes were identified corresponding to 56 individual gene families (Fig. S33b). Enrichment for segmental duplication was calculated using these 56 gene families and a list of 17,601 nonredundant human genes, manually curated to include only one representative gene from each duplicated gene family (*S3*). Genes were considered to be in segmental duplications when >50% of their length was covered by one or more segmental duplications. We found striking enrichments for segmental duplications in variable genes (OR 311.3, $P<2.2\times10^{-16}$, Fisher's Exact). Notably, many of the genes found to be missing copies from the reference were also the most variable among humans.

We also observed 28 large regions showing extreme copy number variability with highly variable estimates within each region, suggesting complex variation in the underlying duplication

architecture (available from *http://hgsv.washington.edu*). One such example spanning nearly one megabase and containing several brain and testis-expressed genes is shown in Fig. S34.

3.3. Highly population stratified genes

We identified the most population stratified gene families across individuals using the statistic V_{st} developed by Redon *et al.* (*S28*) (Table S8). Briefly, V_{st} was calculated as $(V_T-V_s)/V_T$, where V_T is the variance in copy number among all unrelated individuals and V_s is the population-specific variance. V_{st} values were calculated separately for each pair of populations. 176 genes were identified having V_{st} values greater than 0.2 clustering to 64 highly stratified gene families many falling within 22 large (>100 kbp) regions of increased population stratification (Table S9). Segmental duplication enrichment was calculated as above indicating a significant enrichment for segmental duplications in highly stratified genes (OR=311.3, $P<2.2x10^{-16}$). Moreover, genes exhibiting population stratification ($V_{st}>0.05$) tended to have higher V_{st} values when they overlapped with segmental duplications (P<2.2x10^-16, two-sided KS test, fig. S35), indicating that segmental duplications harbor some of the most highly population-stratified sites in the genome. No correlation was observed between V_{st} and copy number indicating that increase of the V_{st} statistic is not an artifact of increased copy (P=0.2; Fig. S36).

Among the most stratified gene families identified by our analysis, many have been previously described, including 3 among our top 12 candidates (*UGT2B*, *CCL3L1*, and *LILRA3*) (*S23-27, 29-31*) (See main text Fig. 3a; Fig. S37). Additionally, we identify highly duplicated gene families previously unassayable for copy number showing striking signatures of population stratification.

3.4. Human-specific gene duplications

We compared copy number from 159 human genomes to those of three outgroup ape species: the genome of a gorilla, a Bornean orangutan, and the "illuminized" chimpanzee genome. We identified 74 genes that were diploid among all nonhuman primates but showed an increase in the median copy in the 159 individuals analyzed. These genes cluster to 23 gene families that have been specifically duplicated in the human lineage (Table S10; see also Main Text Fig. 3b) and localized to 16 regions. We determined that 8 of these 23 gene families are largely fixed in the 159 samples we analyzed, having a variance <0.15 in our population of 159 individuals.

To further develop our analysis of human lineage-specific gene duplications, we identified an additional 100 genes clustering to 30 gene families that have undergone human-specific expansions compared to all of the analyzed primate lineages (Table S11, Fig. S38) requiring the median human copy to show a differential of at least three from the nonhuman primates. Of interest, a number of genes involved specifically in human neuronal development and disease were indentified among human-specific gene duplications and human expanded gene families including *HYDIN* for which we confirmed two rare deletion/duplication polymorphisms by FISH (Fig. S39), GRPIN2 and SRGAP2 (Fig. S40).

We compared the relative divergence of duplications underlying gene families specifically duplicated in the human lineage to those of all nonredundant segmentally duplicated genes. Both

human-specific gene duplications (diploid in great apes) and genes increased in copy along the human lineage were found to have significantly higher sequence identity ($P<7.0\text{x}10^{-5}$ after multiple testing correction, Welch's 1-tailed *t*) providing independent confirmation of the recent, human-specific origin of these gene duplications. The median percent identity of genes with >80% overlap in segmental duplications was 98.7, 97.0 and 95.1, respectively, for human-specific, human increased and all duplicated genes.

4. SINGLY UNIQUE NUCLEOTIDE (SUN) ANALYSES

Massively parallel short-read sequencing is becoming widely adopted for the discovery and typing of genetic variation. Most efforts to date have been confined to unique portions of the genome owing to an inability to uniquely map short reads originating from repeats or duplication tracts. We address this shortcoming by focusing on reads matching the small subset of divergent positions (<10%) within otherwise identical duplications. Using the underlying segmental duplication pairwise alignments from the human reference genome, we searched for paralogous sequence variants that could be used to uniquely distinguish and tag each paralog (Fig. S41). We termed these singly unique nucleotide (SUN) identifiers. We show even with short reads and modest coverage (>=30 bp, <5X), it is possible to genotype both the individual content and copy number of paralogs within most of the highly identical segmental duplications in the human genome.

4.1. SUN discovery and genotyping

We created a database of singly unique nucleotides (SUNs) that uniquely tag a paralog within the length of short read (~30 bp). To construct this database, we first segmented the human genome (UCSC hg18/NCBI build 36.1) into its constituent 30mers (SUN kmers, or "SUNKs") and their reverse complements ($n = 6.19\text{x}10^9$). Next, we excluded all 30mers occurring more than once in the genome as these positions cannot uniquely tag a single genomic locus without additional information (e.g. longer reads or mate-pairing information). For each of the remaining $2.40\text{x}10^9$ Watson-strand 30mers, we found all matches to the genome within edit distance 2 using the *mrFAST* aligner (*S3*) and discarded $1.40\text{x}10^8$ 30mers (5.83%) that, although unique, had a large number of close matches (>500 within edit distance 1 or 2) or were comprised of highly repetitive 15mers (not shown). Such kmers are therefore at greater risk of spurious mappings from distant loci. After filtering, $2.256\text{x}10^9$ SUNK positions remained throughout the genome (including $4.488\text{x}10^7$ in segmental duplications).

Each cluster of SUNK positions within a duplicated tract is made unique by one or more underlying SUNs resulting from paralogous sequence variation. To catalog these, we analyzed all pairwise global sequence alignments between segmental duplications. We first consolidated the 44.8 million SUNKs within segmental duplications into 535,992 clusters spanning 68.3 Mb of the genome covered by one or more SUNKs. We then extracted intervals corresponding to each SUNK cluster from every pairwise alignment of segmental duplications in which that cluster overlapped. Within the resulting extracted fragments of pairwise alignments, we identified the underlying nucleotide differences and categorized them by class: substitutions (transitions and transversions), insertions, and deletions.

Because we sought to define per-position markers useful within the context of short reads, we counted SUNs by single-base physical coordinates within the reference genome. For example, a unique, 1-kbp insertion in one paralog relative to another would be counted as 1000 SUN positions in the former and only one in the latter, even though the underlying indel in this example may represent one discrete variant event. Nevertheless, for the purposes of short-read-based genotyping, the insertion-bearing paralog can be uniquely tagged by these 1000 inserted positions, while the cognate paralog is only tagged by the corresponding deletion junction position. This asymmetry was corrected by counting SUNs at each pairwise-aligned duplication (except in cases when one was on an unanchored chromosome, in which case only SUNs at the anchored copy were counted).

In total, we identified ~12.6 million SUN identifiers within segmental duplications. The transition/transversion (Ti/Tv) bias for these was 1.54, comparable to the overall ratio within segmental duplications as a whole. We excluded SUNs found in or near (+/- 36 bp) repetitive elements as identified by RepeatMasker and TandemRepeatFinder. This step did not grossly alter the density nor makeup of the resulting SUN set. In total, there remained 4.08 million unique identifying base pairs, or 1-bp difference per 13.4 bp, which could be used to infer the specific identity of a paralogous sequence (Table S2).

We sought to quantify what fraction of the duplicated portion of the genome that could be ascertained for copy number using SUNs. As expected, the density of SUN positions diminished with increasing identity (Fig. S42). Indeed, the density of SUN positions might be expected to follow a linear relationship to the percent divergence or equivalently (100 - %ID): the more highly identical a duplicate is, the fewer divergent bases and thus the fewer SUNs it contains. This relationship breaks down for some duplicon blocks, and the density of SUNs exceeds the minimum level of pairwise divergence. The following pathological example illustrates how this can arise: consider a 100-bp block that underwent nine duplications in close succession. Following these duplications, each new copy diverged at one different position. Relative to any one of these ten blocks, the nine duplicates are identical at 99/100 positions, but within all the blocks, there are by definition 10 SUNs differentiating that block from its nine duplicates. This leads to a SUN density of 10/100, which exceeds by the mean divergence by tenfold. Thus, pairwise divergence of duplicates is not always indicative of the density of markers suitable for paralog-specific genotyping.

To be useful for paralog-specific copy number genotyping, SUN markers must be in fixed association with a duplicate locus as defined by the reference assembly. Markers appearing to be SUNs may, in principle, instead represent SNPs within duplicated sequences, rare variants, or errors in the reference. In these alternative cases, a SUN's presence or absence among short reads does not necessarily reflect the copy number of its associated locus. Because we required a perfect read placement (edit distance 0) to consider a SUN present, several effects could cause a SUN to appear absent, including single-base variants affecting SUN itself or its flanks (+/- the read size, 36 bp), deletions or rearrangements eliminating the SUN or altering flanking sequence, and low sequencing coverage.

To determine the extent of these effects, we checked each SUN marker for presence or absence among reads from 12 unrelated human genomes sequenced to high coverage (>10X). We noted an

overwhelming majority (~91%) of SUNs found in >=11/12 individuals (Table S12, Fig. S43), suggesting that these positions largely represent differences that can be used to tag specific paralogs. Turning to the full set of 159 individuals, we noted that >81% of SUNs were present in two-thirds or more of the genomes, with SUNs present at similar rates among each of the three populations sampled. At a more stringent threshold (presence among >80% of individual genomes), we noted a significant reduction in the SUNs' presence (~58%).

We noticed that SUNs dropped out of low-coverage genomes much more rapidly than in high coverage genomes (94.2% SUNs present among at least 10/12=83% of high coverage genomes compared with only 57.7% present among 128/159=81% of all genomes). Therefore, we reasoned stochastic "drop-out" due to low sequencing coverage was primarily responsible for the higher rate of SUN absence among the full set of genomes. Indeed, the fraction of SUNs present in each individual genome closely correlated with coverage — those genomes sequenced at lower coverage were overwhelmingly the ones contributing to the 'missing' SUNs (Fig. S43d). The rate of SUN drop-off appeared similar in all three populations rather than being skewed towards the African population which has relatively more rare variation (*S32*), suggesting that there are relatively few allelic variants disguised as SUNs, or that their alternate alleles (i.e., not the ones represented in the reference) are quite rare.

We next modeled the effects responsible for missing SUNs by simulation. SUN markers were assumed to be polymorphic with allele frequencies drawn from a mixture with three components: (1) fraction p_{fix} of SUNs being present in all individuals (allele frequency=1), (2) fraction p_{most} of SUNs being present in most individuals (allele frequency=f) and (3) fraction (1-p_{fix}-p_{most}) drawn from a uniform distribution over allele frequencies U[0,1]. To determine the simulated genotype at each SUN, we divided each chromosome's SUN markers among these three components and for each SUN, drew two alleles (allele 0, SUN absent; allele 1, sun present) from the respective component's allele frequency distribution, except for chrX and chrY in males, for which only one allele was drawn for each SUN. This process was repeated for each sample to produce a simulated genotype of all SUNs for that sample. The depth of reads covering each SUN was then modeled separately within each sample, and was drawn from a Poisson distribution with parameter $0.5G_iC_j$, where $G_i \in \{0,1,2\}$ was the genotype of SUN i (respectively, homozygous absent, hemizygous, and homozygous present) and C_j was the mean coverage by edit-distance 0 reads in sample j. We then counted for each SUN the number of simulated samples in which that SUN was present (i.e., had simulated depth of mapping > 0), and repeated this process for all combinations: p_{fix}=0,0.025,0.05,…,0.95,0.975,1.0, p_{most}=0,0.025,0.050,…,1-p_{fix} and f=0,0.05,0.1,…0.95,1.0. The parameter set with lowest root mean squared deviation (RMSD) between the observed and simulated rates of autosomal SUN dropout was (p_{fix},p_{most},f)=(0.425,0.275,0.800) meaning that 42.5% of SUNs were fixed, another 27.5% had allele frequency 0.8, and the remaining 30% had allele frequencies uniformly distributed between 0 and 1, for an overall mean SUN allele frequency of 0.795 (Fig. S44). Because of the multiple effects including flanking variation that can cause a SUN marker appear to be absent, this "allele frequency" actually refers to the allelic frequency at which (i). a SUN marker is present as annotated and (ii). there is no flanking variation. As such this is a lower bound on SUNs' actual allele frequencies – the alternate (non SUN) alleles may be much rarer and the extent of allelic polymorphism thus much lower. This further reinforces the

notion that the primary reason for not observing a SUN is stochastic drop-out rather than polymorphism, and confirms the utility of these markers for genotyping across populations.

Relative to unique portions of the genome, segmental duplications are enriched for annotated SNPs (*S33*), many of which may actually be paralogous sequence variants (PSVs) in fixed association with one duplicate allele that are misannotated as polymorphisms at highly identical loci. Indeed, among the 3.28 million autosomal SUN positions found in all 12 high coverage genomes, which likely represent PSVs present at high frequency in the human lineage, we found 275,245 SUNs that were annotated as SNPs (dbSNP v130, quality filtered from UCSC Genome Browser). This was a highly significant enrichment for overlap between these fixed SUN positions and annotated SNPs, even beyond the overall enrichment for SNPs within segmental duplications. Some of these SUNs may indeed be SNPs with low minor allele frequency (MAF) – high-coverage sequencing of additional individuals will be required to resolve which of these are fixed SUNs devoid of common allelic variation and which are low-frequency SNPs.

Having characterized these SUN markers for presence or absence across 159 genomes, we next applied them in a quantitative fashion for paralog-specific copy number prediction. As for total copy number prediction, we estimated paralog-specific copy number (psCN) by the depth of short-read mapping, taking into consideration only reads hitting paralog-specific markers (specifically, reads starting at SUN k-mer (SUNK) positions). Due to coverage effects and the scarcity of these markers within highly identical duplicates, we measured the average number of reads across all markers within each duplication interval. In order to obtain a psCN estimate, we performed linear regression using a model trained on regions of known copy (as described above for total copy number). For paralog-specific genotyping, we counted only reads mapping to SUNK positions with zero edits. This should exclude spurious mappings from paralogous copies at the expense of losing a minority of reads with one- or two-base sequencing errors or nearby SNPs. The paralog-specific copy number linear model was fit using only edit distance 0 placements over the same regions. One important difference between this and our total copy number estimation is that while for the latter we used depth-of-coverage (i.e., the number of reads overlapping each given position), for psCN estimation we considered instead the number of reads *starting* at each position. The reason for this difference is that reads starting at non-unique positions can "bleed in," adding coverage to unique positions and skewing psCN estimates. Using reads restricted to the start positions of SUN kmers alleviates this problem.

4.2. Validation of paralog-specific copy number estimation

Comparison with high-confidence, fosmid-mapped deletions

As one means of validating our paralog-specific copy number (psCN) genotyping, we computed psCN estimates for 383 high-confidence deletion intervals previously identified by cloning and resequencing. These deletion intervals were initially identified within eight individuals (NA18517, NA12156, NA19129, NA18956, NA18555, NA12878, NA19240, and NA18507) by discordant fosmid clone end sequence pair mapping and were resolved to the deletion intervals at the single-base level by full capillary-based resequencing of the respective clone inserts (*S9*). GenBank accessions for the deletion breakpoint sequences are listed in Table S14.

Treating these deletion calls as a gold standard, we used short-read depth at SUNK markers to obtain a psCN genotype of each deletion interval in the same individual where it was discovered. As a control, we also computed the number of concordantly mapping fosmid clone end sequences within each deletion—true homozygous deletions tend to have zero clone ends mapping within the deletion interval. For a hemizygous deletion, there remains one copy of the deleted sequence in the genome; thus, except for very short deletions, the deletion interval tends to contain at least one concordantly mapped clone end originating from the non-deleted allele. We first examined the distribution of psCN genotypes within each deletion interval. We classified deletions with psCN less than 0.5 as homozygous, between 0.5 and 1.5 as hemizygous, and above 1.5 as diploid or greater.

We treated deletions in unique portions of the genome (n = 310) separately from those overlapping segmental duplications (n = 73). The vast majority of deletions within unique regions were correctly predicted (psCN<1.5, 301/310, 97.1%, Fig. S45a). We called 81 (26.1%) as homozygous deleted. Of these, 65 (80.2%) and 78 (96.3%) were overlapped by 0 or <=1 concordant clone end, respectively, confirming that we effectively discriminated between homozygous and hemizygous state. We called 220 (71.0%) deletions as hemizygous. Of these, a smaller proportion had 0 or <=1 concordant clone ends as compared with the predicted homozygous deletions: 38 (17.2%) and 98 (44.5%), respectively. The remaining 9 deletion intervals (2.9%) were predicted as diploid or greater, reflecting errors in psCN genotyping (these 9 regions were on average shorter, with a median unmasked length of 605 bp compared to the 1567 bp for deletion intervals called with psCN<1.5).

Among deletions overlapping segmental duplications, the majority of intervals were predicted as significantly less than diploid in paralog-specific copy number (psCN<1.5 in 59/73, 80.8%, Fig. S45b). The fact that such a large fraction was correctly predicted despite being extensively duplicated (mean of 7.6X per deletion interval) is strong confirmation of the specificity of psCN genotyping. We genotyped 22 (30.1%) as homozygous deleted and 37 (50.7%) as heterozygous deleted. Again, a majority of predicted homozygous deletions (psCN<0.5) had either 0 or <=1 concordantly mapped fosmid end sequences (15/22 = 68.2% and 18/22 = 81.8%, respectively). There remained a larger fraction (18.2%) of called homozygous deletions containing >1 fosmid clone end, which may represent clone end sequences from distant paralogs spuriously mapping in the deletion interval. We predicted diploid or greater paralog-specific copy number for 14 (19.2%) of the deletion intervals, of which some are embedded within flanking amplification events.

To further evaluate cases where we predicted a diploid (or greater) state for the deletion interval, we computed psCN at 10 kb flanking both sides of each deletion interval and compared that with the psCN estimate within the deletion interval (Fig. S46). If the deletion is accurately genotyped, we expected the estimate within the interval to be less than that of its flanks. We observed this to be so in the majority of cases: 308/310 (99.4%) for deletions outside of segmental duplications, and 68/73 (93.2%) for those inside of segmental duplications. Of the 14 segmental duplication-associated deletions with psCN>1.5, almost two-thirds (9/14, 64.3%) had lower predicted copy inside the deletion interval relative to their flanks, affirming the accuracy of these genotypes even when nested within highly duplicated regions such as upstream of the *NOTCH2NL* gene on 1q21.

Such cases may arise when the reference genome is "missing" copies of a duplicated region (see above).

Experimental validation

Amylase

We compared predicted psCN genotypes at the amylase locus (Fig. S53) with single-channel array CGH copy number estimates (*S6*) and previously reported quantitative qPCR at this locus (*S34*). We summed our psCN measurements at *AMY1A*, *AMY1B*, and *AMY1C* and observed similarly strong correlations with estimates from the arrays (r = 0.714, Fig. S47) and from qPCR (r = 0.713, Fig. S48).

APOBEC

We compared our psCN genotyping with PCR-based genotypes at the *APOBEC3B* gene, which is commonly deleted among East Asian and Oceanic populations (*S35*). Treating these PCR-based calls as a gold standard, we achieved 86.8% genotyping accuracy this locus (Fig. S49).

CFHR

We validated our psCN genotyping at *CFHR3* using a quantitative PCR assay specific to this paralog. As with our other comparisons to qPCR data, we observed strong concordance (r>0.9, Fig. S50). qPCR confirmed the population stratification predicted by read depth psCN analysis. This deletion shows a higher allele frequency among YRI.

4.3. Paralogous regions of genomic variation

Our SUN-based genotyping allows us to delineate specifically which copy of a polymorphic paralogous region is variable. Analyses of these regions uninformed by SUNs will spuriously detect polymorphisms that we term *mirror* copy number variants—those where the location of the true polymorphism cannot be distinguished due to homology (Fig S51). We re-genotyped 406 of the 952 large (>50 kbp) CNV events discovered by dCGH that overlapped segmental duplications, estimating their paralog-specific copy number. We determined that 60% (245/406) of these regions showed no signature of variation using SUNs and hence represent mirror effects from cross-mapping between highly identical duplicated sequences.

To assess the extent of mirroring, we estimated total copy number and SUN-based psCN genotypes for 46,937 CNVs reported by a recent array CGH study (*S5*). We compared these genotypes among the 36 test individuals shared between that study and the current work and additionally took into account our estimates of total and paralog-specific copy number for the reference genome used in that study (NA10851). We focused on events within segmental duplications (>70% overlap) with a sufficient number of markers to allow paralog-specific genotyping (>400 SUNKs). Among these, we found that total copy number genotype concordance was 95% (4181/4412) between read depth estimates and array CGH. Strikingly, SUN-based paralog-specific genotyping concordance was only 62% (2734/4412). In contrast, concordance for regions outside of segmental duplications was essentially identical (read depth-based concordance 88%, 9235/10470; SUN concordance 91%, 9526/10470). Increasing the minimum number of required SUNKs to genotype a region to 2000 only increased the concordance to 61.5% (1820/2960), strongly suggesting that

the decrease in concordance from read depth-based genotypes is not due to a lack of power to genotype paralog-specific copy number. We find that 35.1% (1547/4412) of calls on an individual level are confirmed copy number variable by read depth-based CNV genotyping. However, these events show no paralog-specific variation in copy and, hence, have been assigned incorrectly as result of sequence homology. To specifically tag the location of *mirrored* CNVs we manually curated 191 of the largest (>100 kb) CNVRs discovered using our dCGH approach. In 42 of these CNVRs we identified likely CNV *mirrors* within segmental duplications. For 28 of these *mirrors* we were able to resolve the specific true location of the CNV to an alternate paralogous region of the genome (Fig S51 and see Fig. 4B of Main Text). Overlapping regions showing likely mirror effects were consolidated and are listed in Table S13.

In addition to correctly defining loci of copy number variation, the method we developed will facilitate refinement of breakpoints within high-identity duplications. For example, as part of our analysis we observed a large deletion mapping between BP1 and BP2 of the Prader-Willi region on chromosome 15q11.2 in 2/159 individuals. This deletion is mediated by tandem clusters of segmental duplications separated by ~700 kbp of unique sequence including the genes *CYFIP1, NIPA1*, and *NIPA2* (Fig. S70). Deletion of these genes is implicated in schizophrenia by GWAS (*S36*) and as a low-penetrance risk factor for developmental delay (*S37, 38*). These previous studies utilized SNP microarrays and array CGH thus only detecting the unique portion of the deletion. Based on SUN read depth, we determined that this deletion extends at least ~95 kbp beyond its previously assessed boundaries, encompassing at least three additional genes. We validated the unique portion of this deletion by FISH analysis and array CGH. We note that the SUN-based breakpoints showed Mendelian consistency between the two related individuals in which we observed the deletion. We also observed the reciprocal gain in this region in one individual (GM18555 CN = 3, validated by FISH). Within this background of amplification, we noticed a small diploid patch indicative of a nested deletion, which we validated by fosmid resequencing (see below).

4.4. Gene family analyses

In section 3, we used short-read-based absolute copy number estimation to discover gene families showing extreme variability, human-specific expansion, and population stratification in total copy. Exploring the dynamics of the individual paralogs making up these duplicated families has been largely infeasible to date. To address this, we applied our ~4 million SUN markers to genotype specific paralogs among these highly duplicated genes. We focused on those identified by our total copy number analyses as showing extreme stratification or variability, in some cases allowing us to resolve the variability to specific paralogs (e.g., *NBPF1*) to the exclusion of others that showed near total fixation in the population (e.g., *NBPF7*).

We started with a global survey of variability among duplicated genes to define the complement of genes, which although segmentally duplicated, showed little variability in the 159 genomes sampled. We classified these invariant copy number genes as fixed within the human lineage. Genes within segmental duplications are often redundantly annotated—i.e., with one physical locus being annotated as containing numerous paralogous gene models. To avoid overcounting duplicated genes, we initially clustered gene models. We obtained from the UCSC Genome Browser

32,571 RefSeq gene models aligned to the hg18 reference assembly. We consolidated gene models sharing more than 50% reciprocal overlap into 22,038 clusters to reduce redundancy from different splice forms and multiple mappings of paralogous genes. Gene model clusters shorter than 50 kb were padded to 50 kb in order to utilize flanking paralog-specific tags for genotyping. Of these, 1240 overlapped with segmental duplications by at least half their length and, of those, we discarded clusters having <100 SUNK markers as well as those on chromosomes X and Y. We estimated the psCN for each of the remaining 990 gene model clusters across all 159 individuals sampled, taking each cluster as a single window and taking the mean signal across each cluster. For each gene we found the median psCN across all 159 individuals and, as a measure of variability, computed the fraction of individuals showing a deviation of >1 copy in psCN from the per-gene median. Approximately half (51%) of the duplicated genes showed common variation in psCN. For the purpose of this study, we defined common copy number variation where >5% of individuals deviated from the median copy number.

We compared this variability with that of 7834 clusters of gene models that do not overlap segmental duplications or other regions of defined CNV. As expected, the duplicated genes were overwhelmingly more copy number variable than their unique counterparts, which showed only slight deviation from median psCN (<0.1% of these genes by the above metric). We reasoned that the greater density of genotyping markers in the unique genes might account for this difference—i.e. most sequence within a unique gene will be informative. To model this effect, we randomly rejected 80% of the markers in unique genes to mimic the distribution of marker densities among duplicated genes. Even after controlling for marker density, the estimated variability of these unique genes remained low (<3.6% of genes showing deviation among >5% of individuals).

Next, we focused on the analysis of specific duplicated gene families (Fig. S52-S60) . We were able to fully ascertain certain gene families (e.g., *LILR*, Fig. S52) as all paralogs are accounted for within the reference assembly and had sufficient density of SUN markers. For others, such as the highly expanded *NBPF* family (*S39*), several paralogs are not represented in the reference assembly and could not be accounted for by psCN analysis. We also excluded paralogs with spurious alignment to the reference (e.g., misaligned, crossing distinct duplication blocks and/or assembly gaps), along with those insufficient SUN markers (<200 SUN kmers) for reliable genotyping.

As one means of validation, we compared total copy number estimates for each of several gene families with the summed psCN estimates of the constituent paralogous. For example, as shown in the case of the duplicates *ESPN* and *ESPNP* (Fig. S59), we observed strong correlation between the sum of individual paralog psCN estimates and the total copy number estimate for the corresponding gene family (R>0.8 in all, R>0.9 in 4/6). Additionally, these showed little bias in copy number (i.e., best-fit line had *y*-intercept ~0, as shown for the duplicated genes *ESPN*) indicating that our psCN estimates at these loci are both accurate and precise.

CFHR

In addition to the two common deletion breakpoints at the *CFH* locus, we also noted two rarer, alternate copy number states in one individual each: single-copy gain of *CFHR1/4* and a deletion of *CFHR1/4* with novel breakpoints (Fig. S60). The former appears to be the reciprocal of the common *del2* deletion of *CFHR1/4.* The latter is shifted towards the centromere relative to the breakpoints of *del2* and, as with *del2*, appears to arise from homologous loci in two segmental

duplications, suggesting that non-allelic homologous recombination (NAHR) mediates this event as it is presumed to for the more common deletions at this locus (*S40*).

LILR

The family of duplicated *LILR* family is found by total copy number genotyping to be highly population-stratified, with decreased copy number in Asians and vastly increased diversity in Africans (Fig. S52). Individual paralogs are functionally distinct, with diverse ligand binding characteristics and alternatively stimulatory or inhibitory effects (*S29*). We find the variability within this gene family to be specific to 5 of the 12 duplicated genes. *LILRA6* showed the highest variability, contributing to the bulk of the increased African diversity. We confirmed the deletion of *LILRA3* at high frequency among Asians (*S30*) to the near exclusion of other groups sampled. The remaining 8 *LILR* copies are diploid in nearly all individuals.

NBPF

The *NBPF* genes have undergone dramatic expansion in the human lineage. These genes exhibit neuron-specific expression and signatures of positive selection (*S41*). Analyzing the individual NBPF paralogs with sufficient SUN marker coverage (Fig S54), we found that of several paralogs (*NBPF3*, *NBPF4*, *NBPF7*, *NBPF15*, and *NBPF22P*) were diploid across most individuals with little variation. In contrast, *NBPF1* was both highly amplified and highly variable (median psCN=8.04, +/- median absolute deviation [MAD]=1.68 copies), with increased copy number and diversity among Africans (African median: 9.55 copies, non-African: 7.62; P<2.48x10-12, Mann-Whitney U). The paralog *NBPF14* showed even stronger population stratification, with a common deletion strikingly enriched among Africans (Africans: 38/56=67.9% with psCN<2; non-Africans: 13/103=12.6%; OR=14.3, P<2.23x10-12).

Amylase

Our analysis also reveals previously uncharacterized differences among well-studied gene families. For example, at the amylase gene cluster, thought to be important for human adaptation to starch-rich diets (*S34*), we find that the gene *AMY1C* is largely copy-invariant while the salivary amylase genes *AMY1A* and *AMY1B* are commonly subject to deletion. 23% of humans are diploid for both copies, while ~20% are homozygously deleted for either of these two genes (Fig. S53). Pancreatic amylase genes (*AMY2A* and *AMY2B*) by contrast show less variability; 7% of individuals carry 3-5 copies of *AMY2A*. *AMY2B* shows more variability in copy number, with 1 individual predicted to be homozygously deleted for this gene, a potential candidate for congenital pancreatic deficiency (*S42*). By comparison, the *AMYP1* pseudogene shows the widest range in copy number with the highest copy number observed among Asians.

4.5. Validation of novel structural variants within segmental duplications

The specificity afforded by our SUN marker approach allowed us to discover novel structural variants within segmental duplications. We performed a genome-wide search for deletions within segmental duplications, selecting regions with psCN <=1.5 in at least 2 of 159 individuals for >=20 of 30 consecutive windows of 25-200 SUNKs (maximum physical window size of 1000 bp). We identified 6712 putative deletions, and selected 31 of them for further validation.

Large-scale validation of CNVs within duplicated regions is challenging because available platforms (FISH, microarrays, qPCR, etc.) are hampered by poor genotyping specificity within such regions. We instead developed a clone-based validation approach. We leveraged fosmid clone libraries from eight individuals, which we had previously applied to discover structural variants (*S9, 43*). Based on our psCN predictions, we selected 144 clones spanning 31 deletion intervals from fosmid libraries derived from the same individuals in which each event was called (Table S15). We re-sequenced the selected clones by pooled, shotgun short-read sequencing on the Illumina platform and mapped the resulting reads back to the same SUN positions as originally used to infer the deletion. We considered these clone-based data to support our deletion calls when we observed hits to SUN positions flanking a region (as defined by its capillary end sequence mapped location) with little or no coverage corresponding to the predicted deletion interval. In some cases, we observed SUN hits only from one end of the insert; these nevertheless supported the deletion call when the observed coverage from the fosmid clone was significantly shorter than its physical size (i.e., 32 kbp or less).

Short-read sequence data from the selected clone inserts yielded interpretable patterns of SUN hits at 18 loci (Figs. S61-S70). These hits validate the specificity of our approach within these highly complex regions and confirm that these events predicted using whole-genome sequencing data arose at the indicated loci rather than as signals spuriously mirroring other duplicates. In addition, this approach confirmed that most of the SUN markers indeed specifically tag one paralog, as predicted. For instance, we confirmed a ~32-kbp deletion on 4p16.1 encompassing the 3' end of the gene *DEFB131* (homozygous in 4/8 individuals, hemizygous in 1/8) despite the presence of highly identical duplicates at >50 other loci (Fig. S61). In another example, we validated a hemizygous deletion in the individual NA18956 at the heavily duplicated *PSG* gene cluster, which encodes glycoproteins with known immunomodulatory roles during pregnancy (*S44*) (Fig. S62). Some of selected loci (e.g., *SMN*) were too sparsely populated with SUN markers to unambiguously interpret. In a minority of cases, we observed SUNs mapping outside of the clone's predicted location; these may represent errors in resequencing, polymorphisms relative to the reference, errors in the reference, or in a number of cases discussed below, interlocus gene conversion. A summary of the results of the clone insert sequencing is provided in Table S15.

In several cases, clone sequencing revealed additional complexities or potential artifacts. Among some psCN deletion loci, for example, the corresponding clones were initially classified by end-sequence mapping as inversions. We hypothesized that such inversions might be 'masking' an underlying deletion or a more complex rearrangement. For most of these, the SUN pattern confirmed the original inversion call but was uninformative regarding the putative underlying deletion (not shown). These deletions may be masked by the surrounding inversion, or they may represent false positives among our calls in which polymorphisms prevented the corresponding reads from mapping to SUN positions with edit distance 0, creating patches without coverage and the false impression of a deletion. We noticed several other cases in which the majority of the SUNK hits mapped between the clone ends but were accompanied by patches of hits to paralogous sequences mapping outside the clone insert. We interpreted these either as errors in the reference (misassignment of sequence between highly identical paralogs) or as polymorphisms acting to convert unique markers from one paralog to those from another. A subset of these may represent interlocus gene conversion relative to the reference sequence, which we discuss further below.

Nextera Illumina sequencing of fosmid clones
We identified by their end sequence mapping locations (http://hgsv.washington.edu/) a total of 144 fosmid clones of shotgun genomic DNA from the following eight individuals: NA18517, NA18507, NA18956, NA19240, NA18555, NA12878, NA19129, and NA12156. Colonies were picked from freezer stocks, grown overnight at LB broth at 37°C, harvested, and clone DNA isolated by alkaline lysis as previously described (*S45*) with minor modifications for use in a 96-well format. After purification, clone DNA was arrayed on a single 96-well plate, in some cases combining two clones from unrelated loci. A sequencing library was created separately from each well using the Nextera Illumina-compatible DNA Sample Prep Kit (Epicentre, Madison, WI) as directed by the manufacturer, with ~100 ng of fosmid clone DNA per well as starting material. The subsequent PCR step was used to add a barcode tag specific to each well. Finally, the 96 resulting barcode-tagged libraries were pooled and sequenced with two lanes of paired-end, 76-bp reads on an Illumina Genome Analyzer IIx with an additional 9-bp index read to recover the barcode sequence. The resulting sequences were partitioned by well of origin corresponding to a single clone or two-clone pool on the basis of the barcode sequence, allowing for up to 2 edits to the nearest matching barcode. Reads from each clone or clone pair were then mapped to the genome using *mrsFAST* as described above. Lastly, we filtered the resulting read placements to perfect hits at SUNK positions to indicate the paralog-specific markers captured by each resequenced clone insert.

4.6. Gene conversion analysis

We observed patterns consistent with interlocus gene conversion among the paralog-specific read depths at several loci throughout the genome. Gene conversion arises from the non-reciprocal genetic exchange of information between two loci or two alleles. Conversion within the human genome is known to act preferentially upon segmental duplications as its activity is strongly dependent on the proximity and homology of acceptor and donor loci [reviewed in (*S46*)]. Previously characterized pathological gene conversion events in humans tend to be relatively short (<3 kbp per event), and we reasoned that we could detect evidence for these events by searching for small patches where the dropout of paralog-specific SUN signatures was accompanied by cognate patches of elevated copy at the corresponding paralogous position.

We noted several distinct patches following this pattern particularly among tandem segmental duplications (Fig. S71). One such 6-kbp patch appeared recurrently over the second exon of each of the duplicated genes *RHD* and *RHCE*—a known site of clinically relevant recurrent gene conversion (*S47-49*). Within our genome-wide psCN maps, this manifested as a localized loss in copy over exon 2 of *RHCE* accompanied by a gain over exon 2 of *RHD*. We interpret this as a signature of conversion acting to replace the paralog-specific markers at *RHCE* exon 2 with those from the paralogous exon and flanking intron of *RHD*.

To experimentally confirm this interpretation, we sequenced fosmid clone inserts spanning the conversion patch from one individual predicted to be homozygous for RHCE→RHD exon 2 conversion (NA18555) and one predicted to lack this conversion haplotype (NA19240). Because these clones are physically constrained to ~40 kbp and most (2/3) mapped unambiguously by capillary end-sequencing to *RHCE*, we considered reads mapping to SUN markers within the

reference *RHD* sequence as evidence for conversion where *RHD* serves as a donor and *RHCE* as an acceptor. Mapping sequence reads from these clone inserts back to SUN locations, we observed that the majority of SUNs corresponded to the *RHCE* locus, confirming that all three clones mapped to *RHCE*. This finding was in agreement with the clone end sequence locations in two of three clone inserts and corrected the map location of the third, (ABC11_48195000_H6) for which the forward end capillary sequence had mapped one end sequence into the putative acceptor patch, spuriously creating a end sequence signature of an inversion. In both clones derived from individuals predicted to have an *RHCE*→*RHD* conversion, we observed a lack of SUNs precisely at the conversion acceptor patch accompanied by an isolated 'island' of SUN hits at the conversion donor patch. This pattern was not observed among orthologous clones from an individual where no gene conversion was detected. These findings confirm the gene conversion event and demonstrate that it is a polymorphism segregating within the population along with other paralog-specific variants such as the European-enriched *RHD* deletion, which we detect and confirm (*S50*).

We noted several other duplicated genes showing similar signatures, although these were often difficult to discriminate from background noise due to the low coverage of the genomes and the length of the interval. To discover other possible instances of recent inter-locus gene conversion, as well as to determine the genome-wide extent of its effect among segmental duplications, we developed a statistic to quantify this signature across all 159 genomes sampled. We evaluated this statistic individually for each annotated segmental duplication within the genome. We first split segmental duplications into windows of length <1 kb, discarding windows with <50 SUN kmers (SUNKs) and limiting each window to <=400 SUNKs. We identified each window's cognate, paralogous window by projecting it through the Needleman-Wunsch pairwise global alignment for the corresponding segmental duplication record. Next, we estimated the psCN of each cognate pair of windows as before, except that for this analysis we discarded SUN markers that did not align within a cognate pair of windows (e.g., in the case of an insertion in one duplication relative to its paralog). We quantified the relative shift in psCN of each window relative to the surrounding duplication by calculating its difference from the duplication-wide average. This was quantified for each window *j* of each duplication *i* in each sample *s* as: $a'_{sij} = \frac{a_{sij} - \overline{a_{si\bullet}}}{\frac{1}{N_i}\sum_k \left(a_{sik} - \overline{a_{\bullet ij}}\right)^2}$ where $\overline{a_{\bullet ij}}$ was the mean psCN across all samples for that particular window, window *j* belonging to duplication *i*; $\overline{a_{si\bullet}}$ was the mean psCN of all windows making up duplication *i* in sample *s*; and N_i was the number of windows in duplication *i*. These were calculated analogously for the other aligned duplication as b'_{sij}. We discarded short duplications as well as very highly identical ones with insufficient SUN markers (specifically, those with <2 windows meeting the SUNK criteria above). Having computed relative shifts versus a per-sample baseline across each segmental duplication, we measured whether there was a negative correlation versus random expectation. In other words, we tested whether the increase in SUN depth for one duplicated locus was significantly associated with reduction in SUN read depth at the cognate positions in its paralog for the same sample. We quantified the degree of such correlation in the human population by counting the number of genomes with a positive shift at that window accompanied by a negative shift at the cognate

window, and vice versa. For each window, we computed c_{ij} = the fraction of samples s with $a_{sij}>0$ and $b_{sij}<0$ or $b_{sij}>0$ and $a_{sij}<0$. To obtain a background distribution for c_{ij}, we shuffled the psCN values among the windows within each duplicate locus and individual (i.e., not shuffling data across individuals). We computed the degree of excess observed in this statistic at each window versus 1000 shuffled replicates. We standardized our empirically determined c_{ij} using the mean and standard deviation from the 1000 shuffled replicates, clipping this score to zero to remove negative values. Finally, we ranked each segmental duplication by taking the mean across all windows making up that duplication of the square per-window standardized c_{ij} values.

The analysis yielded 7862 segmental duplications ranked by the degree of enrichment of this signature relative to controls where the psCN values at each locus had been scrambled within each sample. The segmental duplication encompassing the Rh blood group locus scored very highly by this measure (score = 7.14, rank = 78/7864). We examined other top scoring loci by this measure and identified several that appear to show a similar signal (Table S16). Many of these mapped to duplicated genes as well as other sites of interlocus gene conversion suggesting potentially widespread functional impact of this signature.

We sought to determine whether this signature was widespread among segmental duplications, or if it instead reflected only a small number of cases. To do this, we recomputed the per-duplication scores as before, after having randomly permuted the sample labels (but keeping the data within each segmental duplication intact). We observed a highly significant excess of conversion signature in the top half of duplications by this measure relative to the signature among the same regions with randomly permuted sample labels ($P < 2.2\text{x}10^{-16}$; when instead comparing signature of real vs. controls for top half of regions among permuted control, $P = 0.60$; one-sided Kolmogorov-Smirnov; Fig. S72).

Finally, we focused on different classes of segmental duplications to determine if the extent of this conversion signature was uniform throughout the duplicated portion of the genome. We observed a very striking enrichment for this signature among tandem (<1 Mbp, center-to-center) relative to distant or interchromosomal duplications, in accord with most known examples of gene conversion both in humans (*S51*) and mice (*S52*) (Fig. S73). Dividing segmental duplications by their level of sequence identity, we found this signature was nearly exclusive to duplications of >95% identity, in close agreement with known examples (*S51*) and that the gene conversion signals became more significant at higher levels of identity. Although beyond the scope of this study, the method we developed and the sites we identified should be valuable for future studies of gene conversion as more individual genomes (particularly from trios) become sequenced to high coverage.

SUPPORTING FIGURES

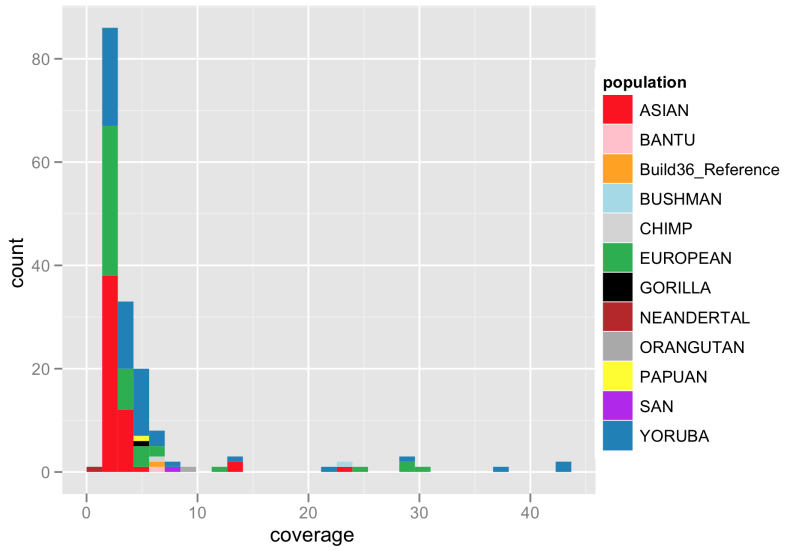


Figure S1. The number of genomes from each population as a function of sequence coverage.

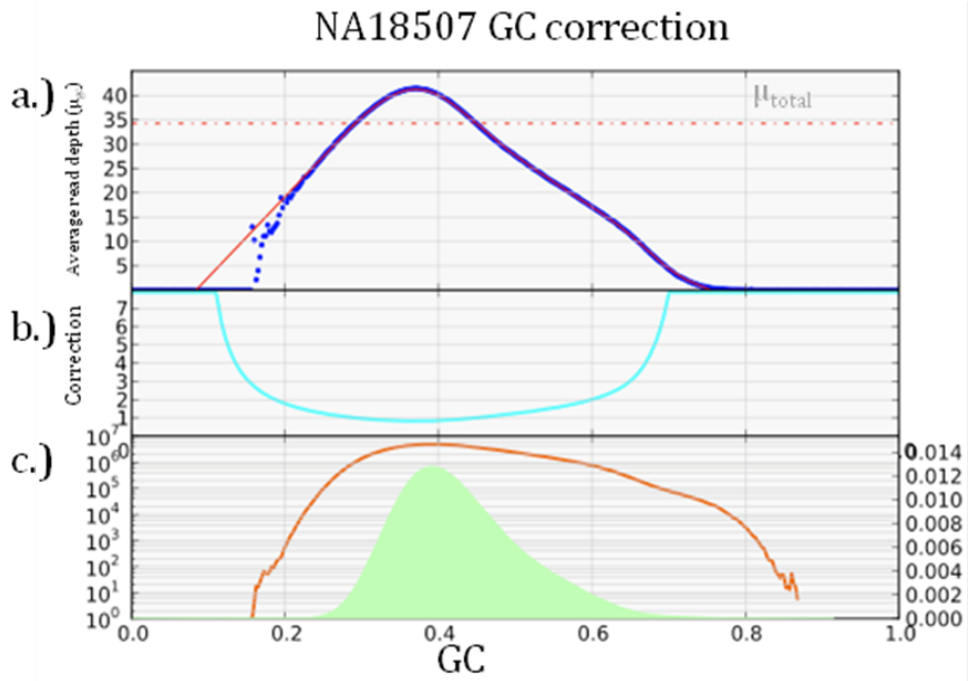


Figure S2. Multiplicative G+C correction for a high coverage genome NA18507. (a) Average read depth for specific G+C% is shown at blue points and fitted with orange line. Total average read depth (μ_{total}) is shown as a red dotted line. (b) The correction factor k_{gc} is shown (truncated at 8) for each individual %(G+C) bin. (c) The fraction of the genome at specific G+C% is shown as a histogram, in green, and overlaid in orange is the log transform of this histogram.

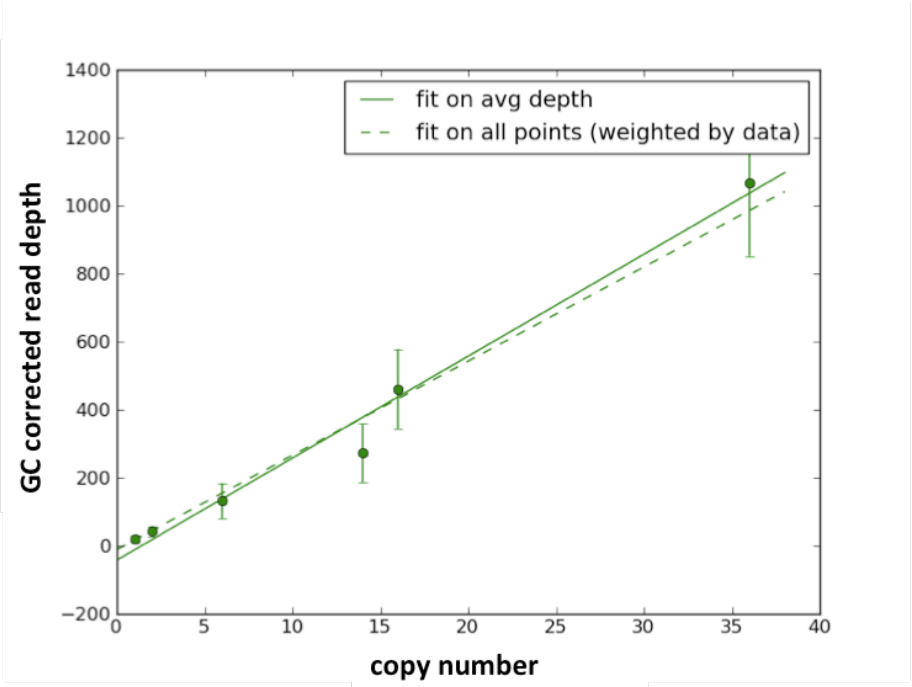


Figure S3. (G+C)-corrected read depth and copy number are linearly correlated.

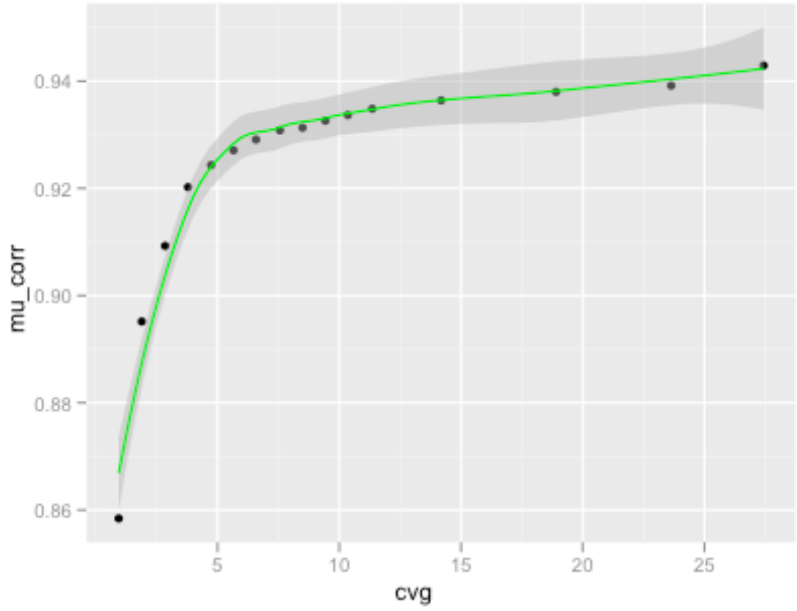


Figure S4. Correlation (*mu_corr*) of read depth to copy in control regions subsampled at different coverages (genome NA18507).

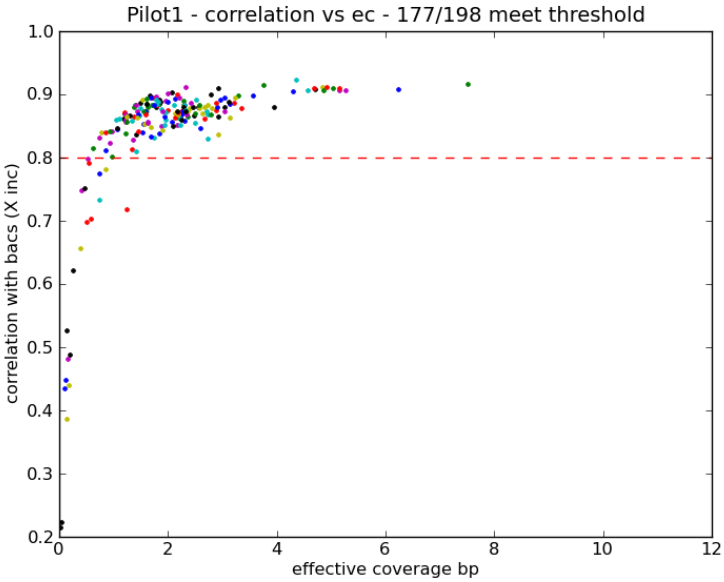


Figure S5. Correlation of read depth to copy in regions of known copy for 1000 Genomes *Pilot 1* libraries (n = 198) colored by genome, plotted as a function of effective coverage.

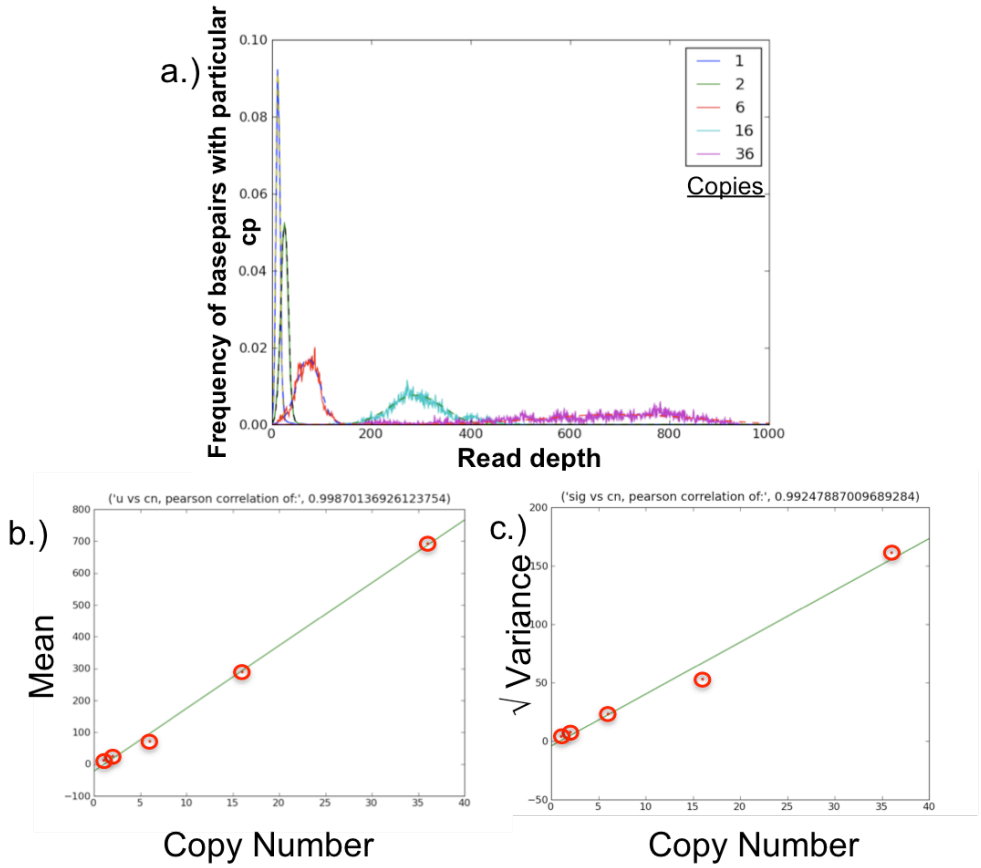


Figure S6. a) Histograms of read depth for regions of known copy with fit Gaussians overlaid. b) Mean of fit Gaussians versus copy number of known regions. c) Square-root of variance versus copy number of known regions.

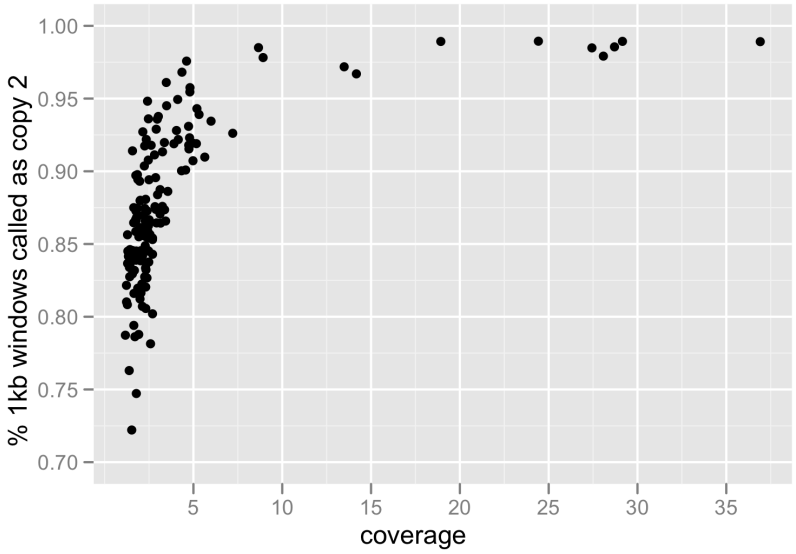


Figure S7. Percentage of 1-kb windows classified as diploid in invariant regions of the genome as a function of varying coverage.

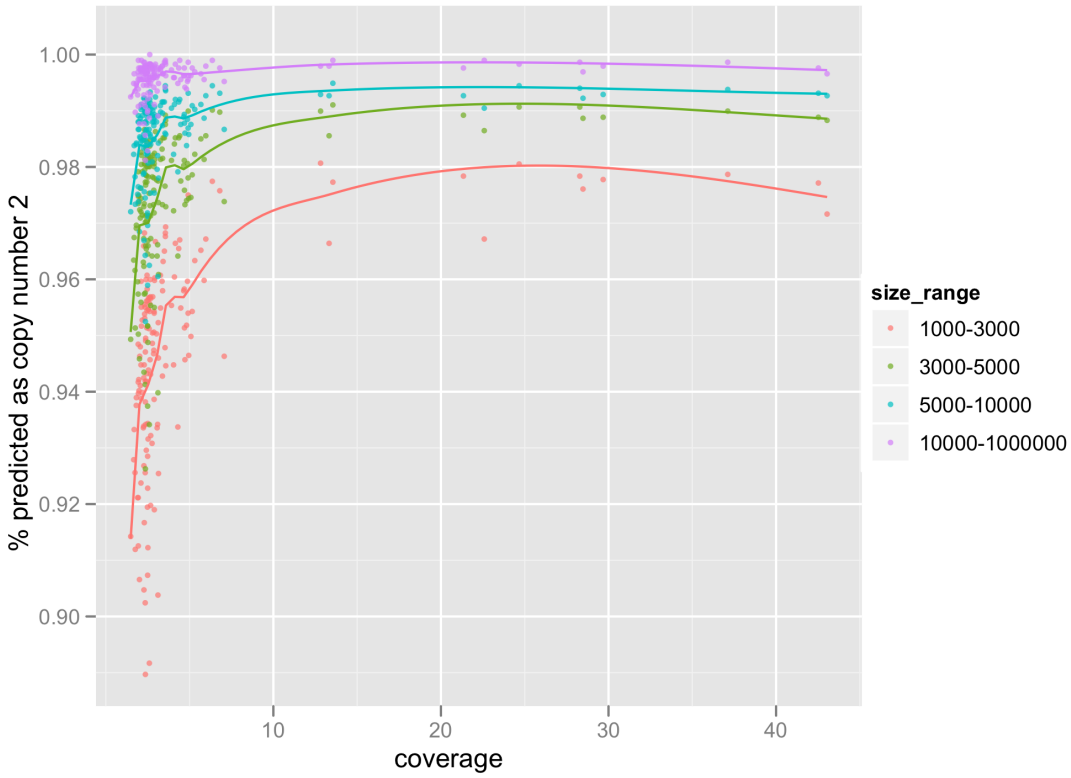


Figure S8. False discovery rate (FDR) as a function of coverage and size. At 3-kb resolution the FDR is <5% for genomes at >1.5X sequencing coverage.

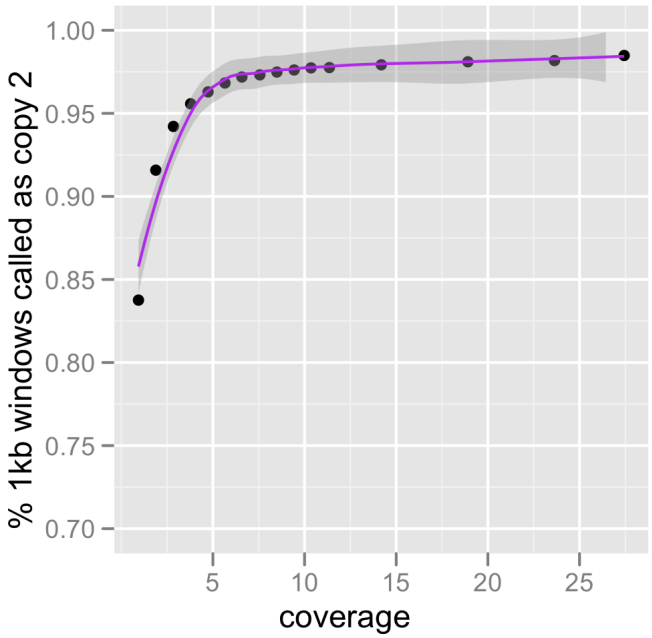


Figure S9. False discovery rate (FDR) estimated based on 1-kb diploid invariant regions from one genome (NA18507) subsampled at various coverages.

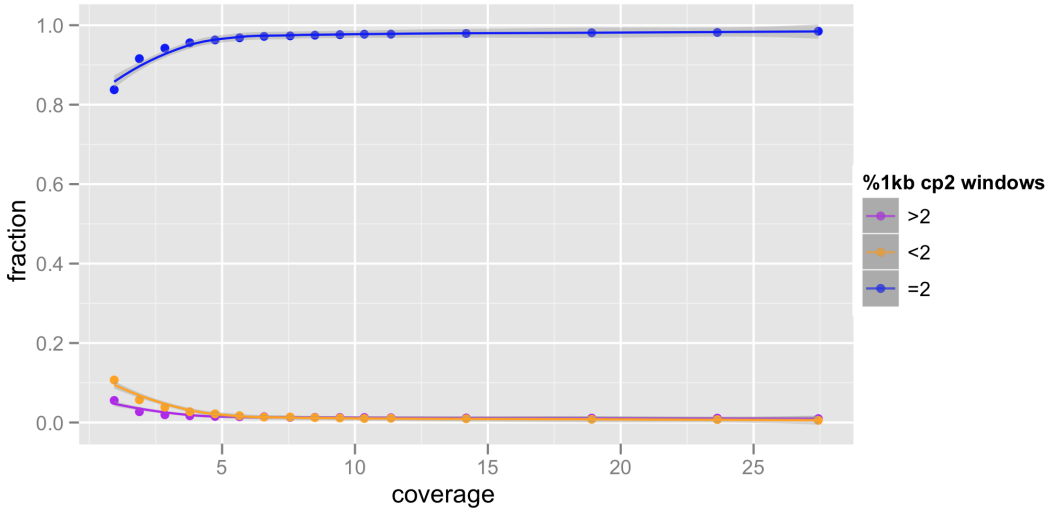


Figure S10. Fraction of 1-kb windows detected as copy number 2, less than 2, or greater than 2 as a function of coverage for one genome (NA18507) subsampled at various coverages.

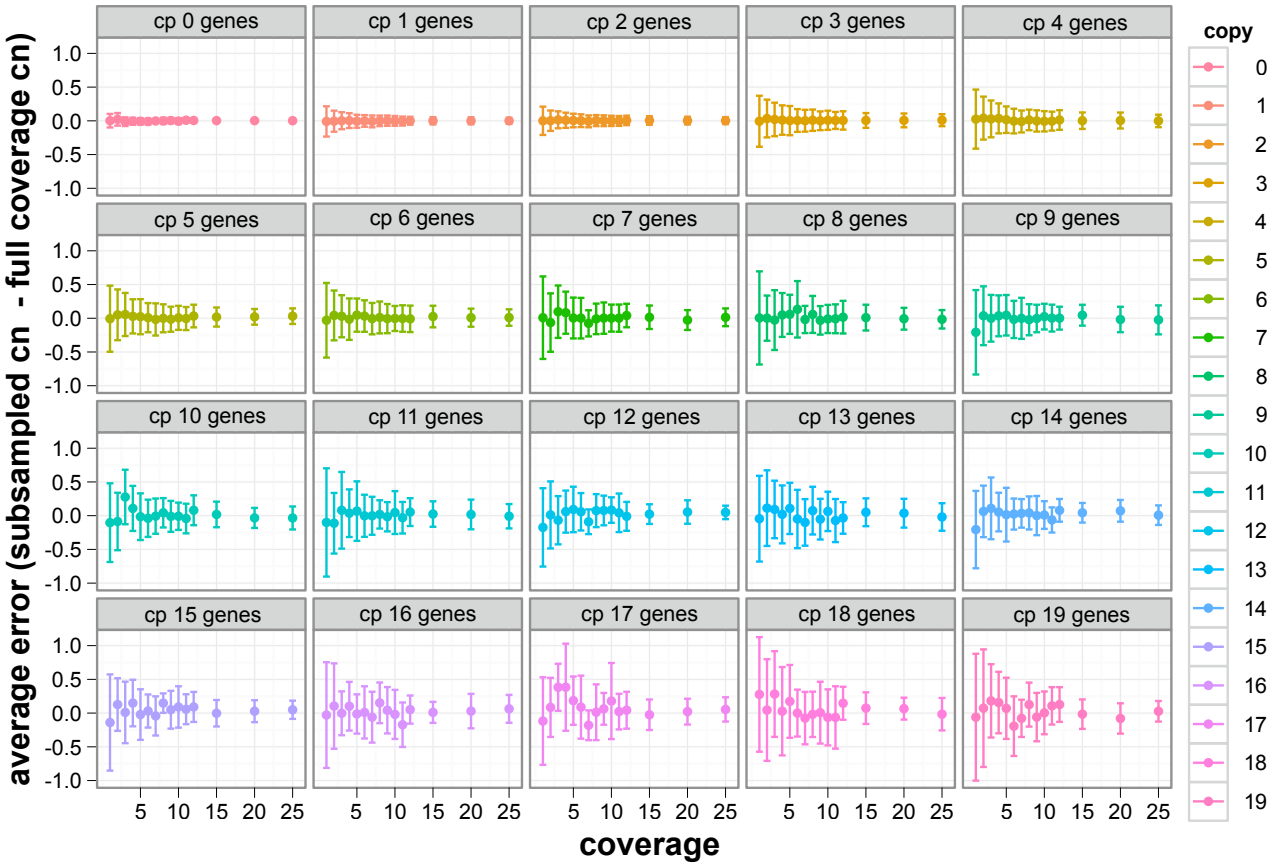


Figure S11. The deviation (mean error +/- 1 s.d.) in copy number prediction for genes >=3kb in length is shown for genomes subsampled from NA18507 to various coverages, using the full coverage (~43x) genome as a gold standard. Genes are stratified by their predicted copy number in the full-coverage genome. As expected, the accuracy rate is higher for genomes of increased coverage. Higher copy number states are increasingly difficult to predict accurately with low-coverage genomes.

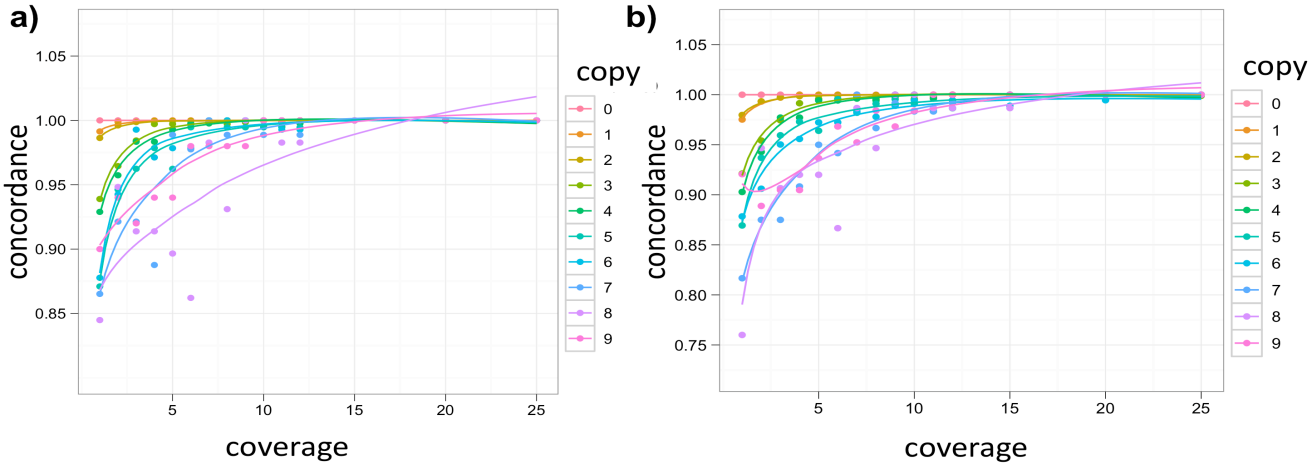


Figure S12. Concordance of copy number estimates versus sequence coverage for all genes a) >3kb and b) >1kb. A gene's copy number estimate was considered concordant following subsampling if it was within +/- 0.5 copies of the full-coverage estimate.

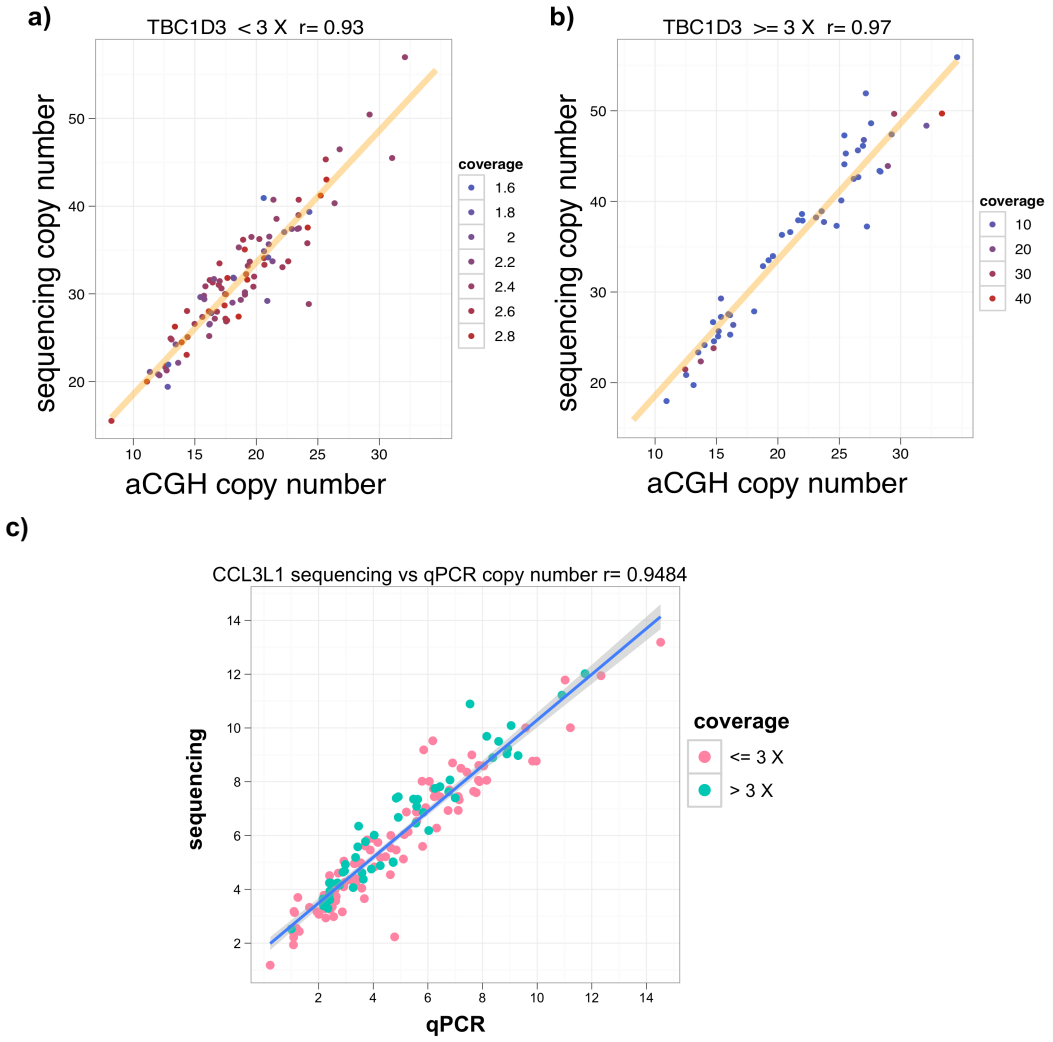


Figure S13. Independent validation of read depth-based copy number predictions for (A,B) TBC1D3 and (C) CCL3L1, stratified by depth of sequence coverage used for read depth-based genotyping.

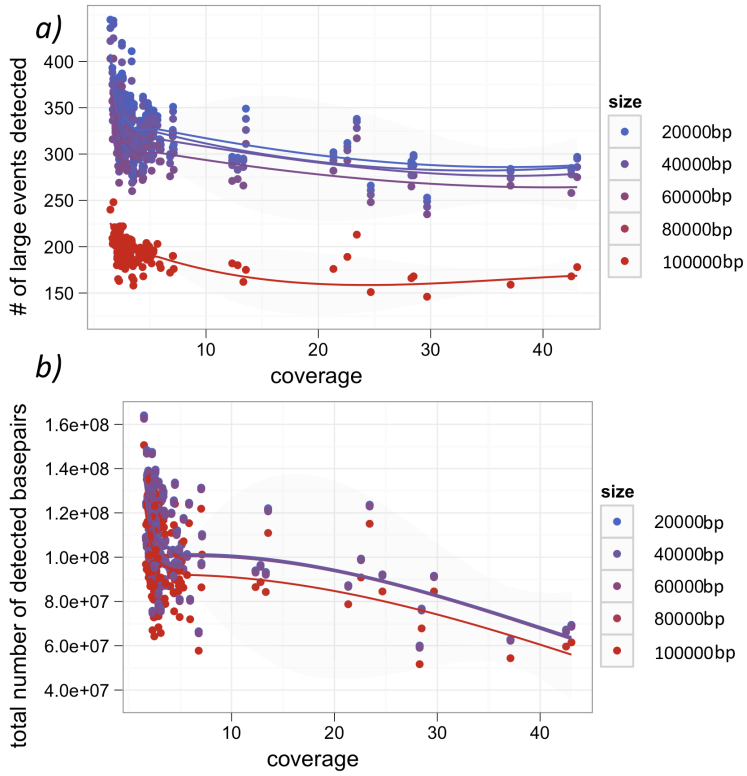


Figure S14. a) The number of events and b) total number of base pairs encompassed by events detected by dCGH in each genome as a function of overall coverage (x-axis). Colors indicate different minimum event-size thresholds. More events are detected in low coverage genomes due to their increased variability.

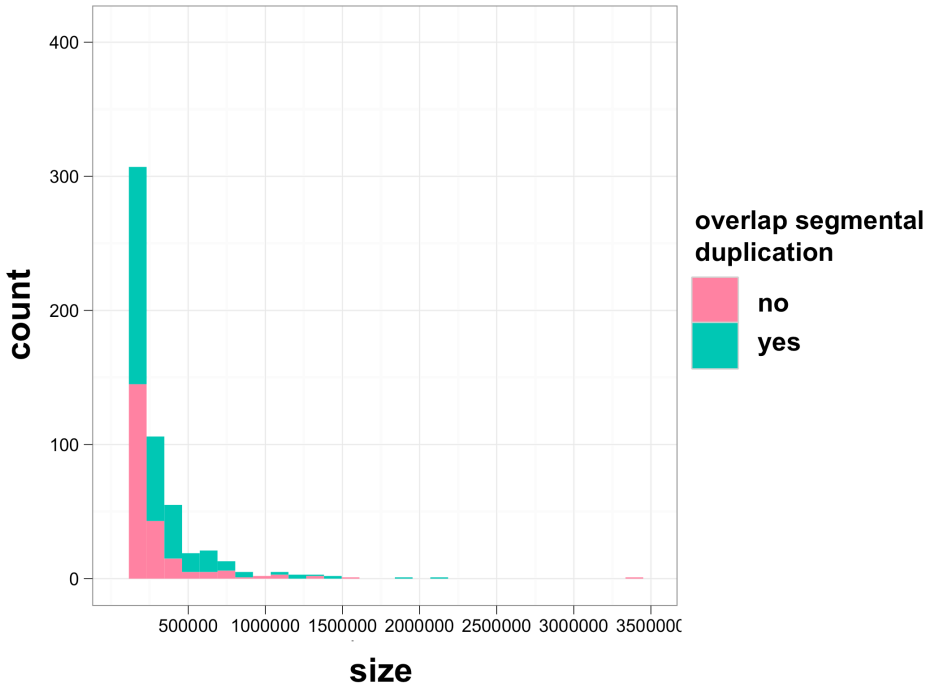


Figure S15. Size distribution of copy number polymorphisms (CNPs) >50 kb colored by segmental duplication overlap (>20% overlapping segmental duplications).

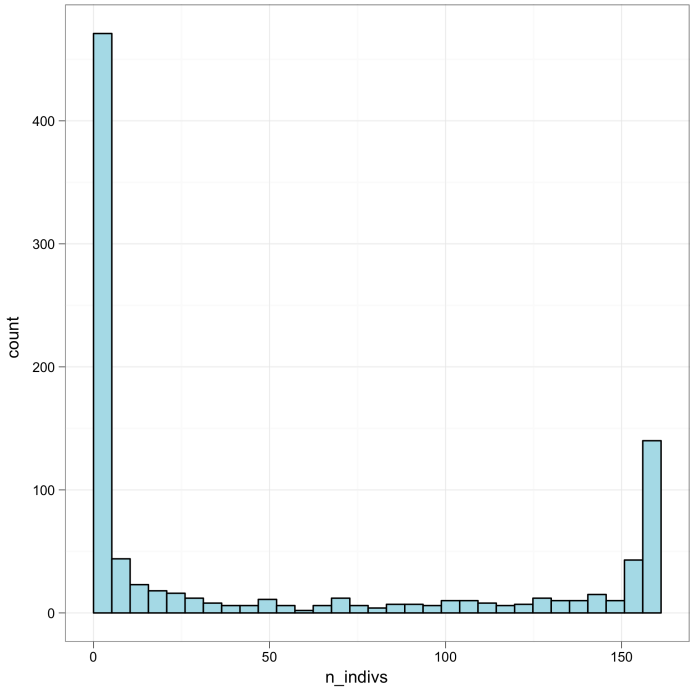


Figure S16. Frequency spectrum of copy number polymorphisms >50 kb shows the majority of events are rare.

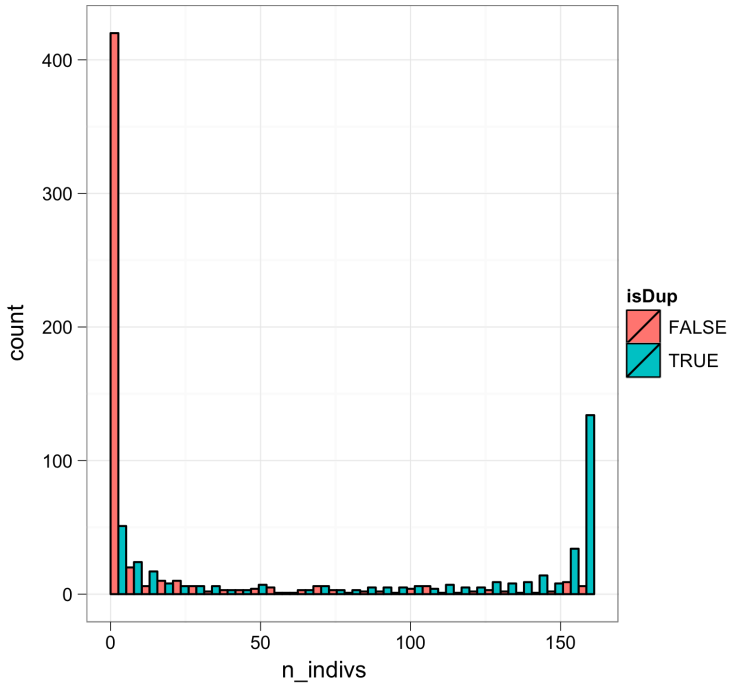


Figure S17. Stratifying the frequency spectrum of events by overlap with segmental duplications (50% overlap) shows CNPs not overlapping segmental duplications are overwhelmingly rare in frequency.

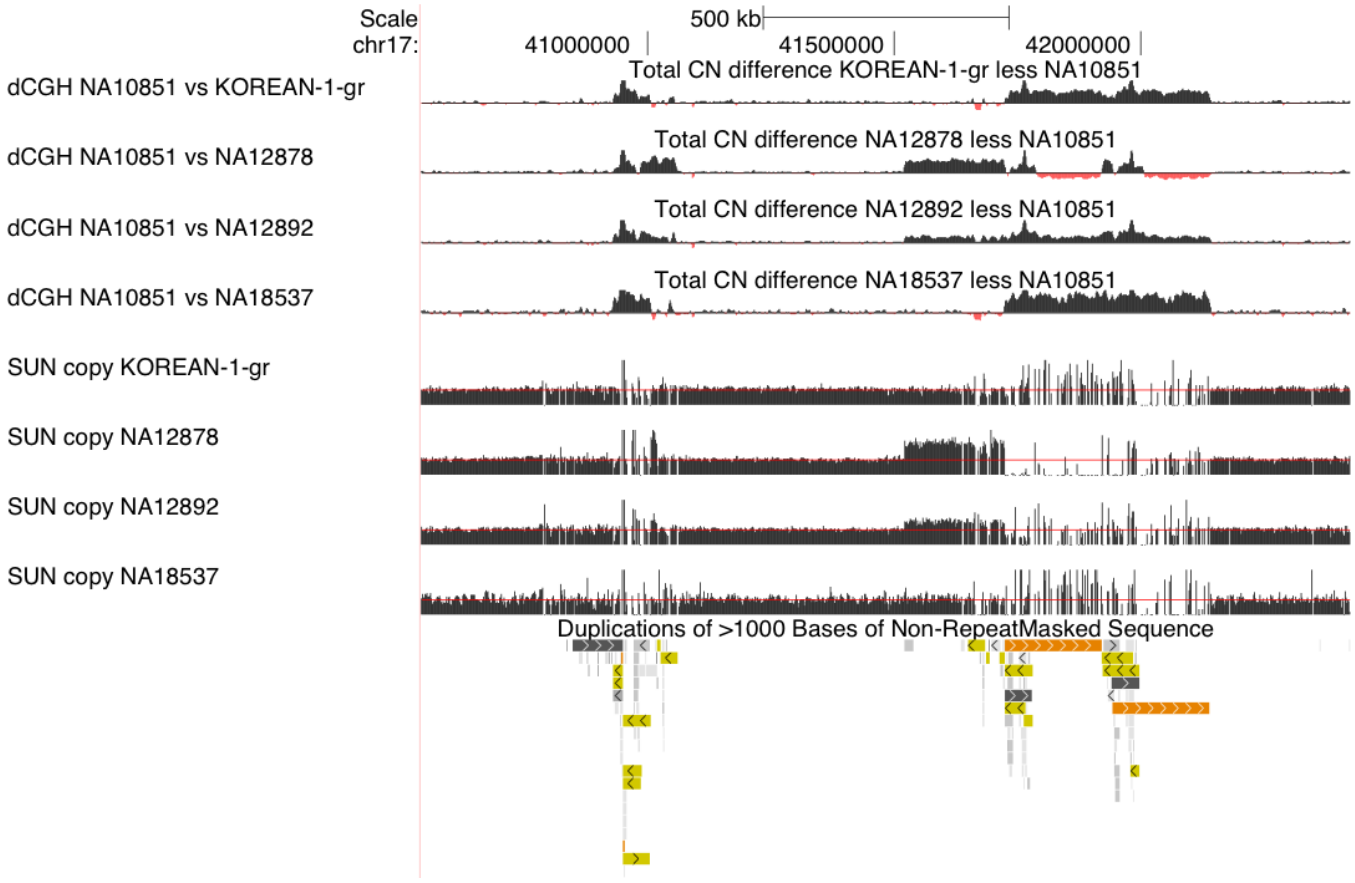


Figure S18. Digital CGH and SUN-based copy number maps demarcate two distinct duplication events on chromosome 17q21.31, for four representative individuals. Digital CGH signal is colored by call: black bars (above the baseline) indicate relative copy number gain; red bars indicate relative loss. Red line on SUN copy tracks indicate predicted paralog-specific copy number of 2.

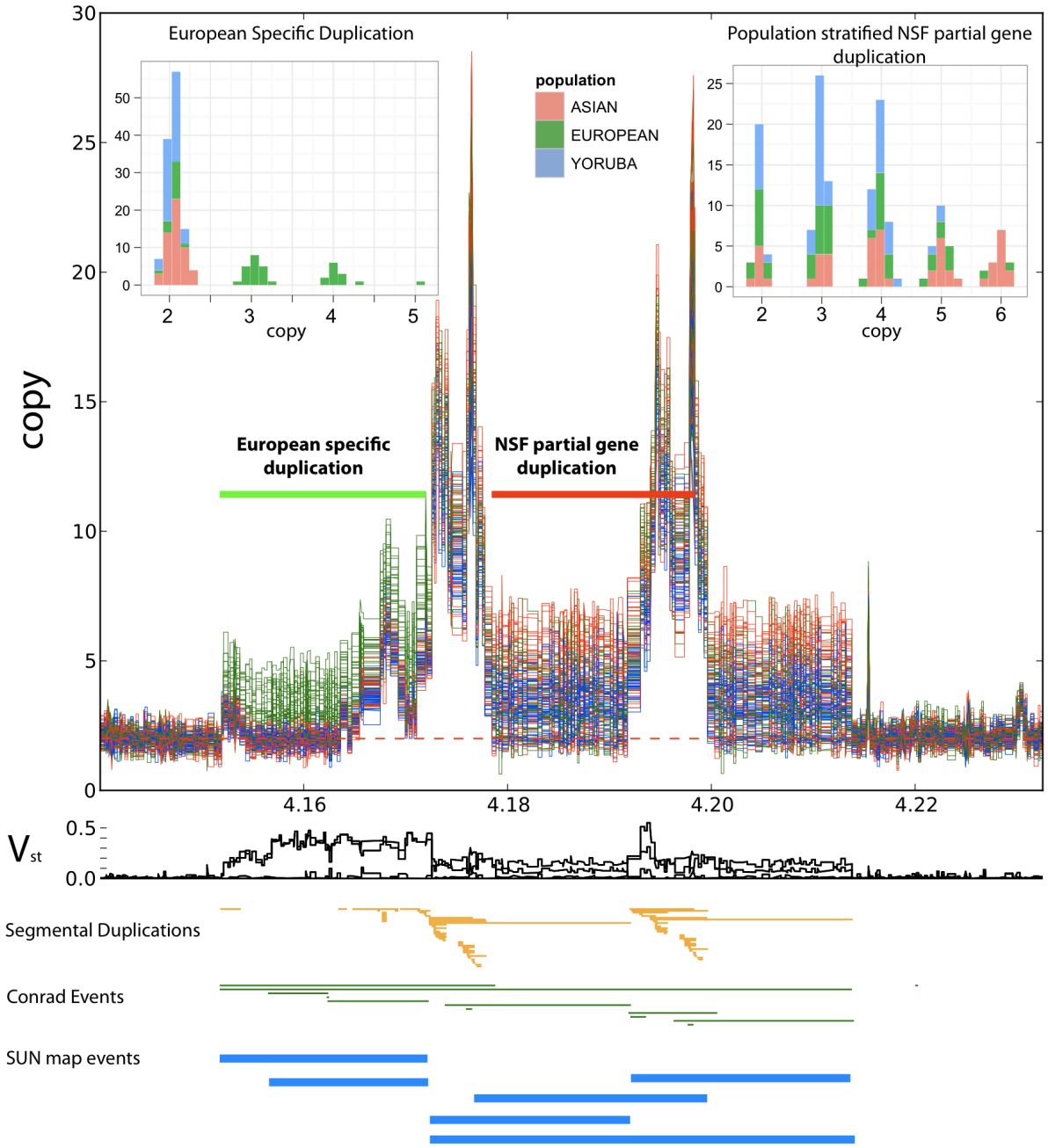


Figure S19. Population stratification in CNVs on chromosome 17q21.31. Total copy number estimates for 154 individuals are overlaid (related and duplicate individuals excluded) and shaded by population. Note the apparent increase in copy among Europeans (green) and Asians (red). Inset are the copy number estimates for two highly population stratified duplications of 210 and 205 kbp. Discrete copy number estimations were confirmed by FISH (see below). The V_{st} metric (as described below) is an indication of population stratification. Strong signals of population stratification are seen for both European-specific and *NSF* partial gene duplications.

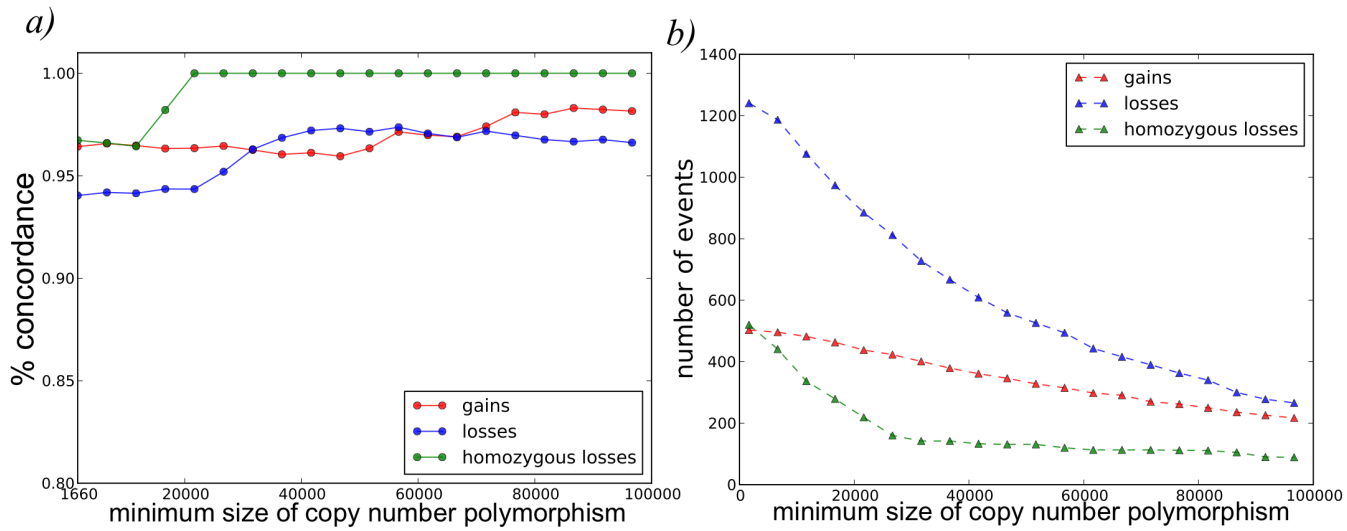


Figure S20. a) Percent concordance between read depth-based copy number genotypes and previous genotype calls made using the Illumina 1M Duo SNP genotyping platform on the same DNA samples (*S7*). Plotted cumulatively by event size (x-axis) b) The number of events assayed at each cumulative size range.

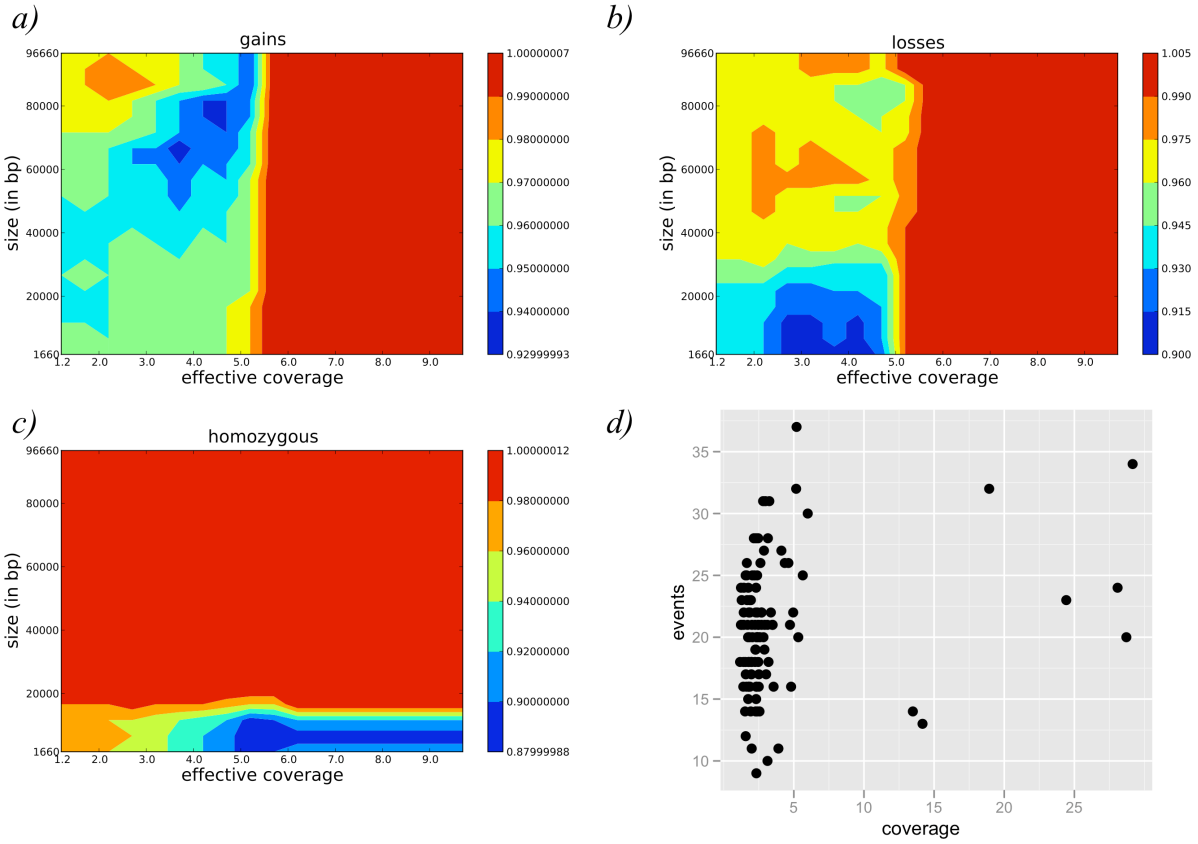


Figure S21. Contour plots showing the effect of sequencing coverage and event size on concordance (ranging from 0 to 1) for a) gains, b) losses and c) homozygous losses. Size and effective coverage both contribute to the overall concordance of events. d). The sudden jump in concordance at ~5X is an artifact of the paucity of genomes between ~5X to ~12X.

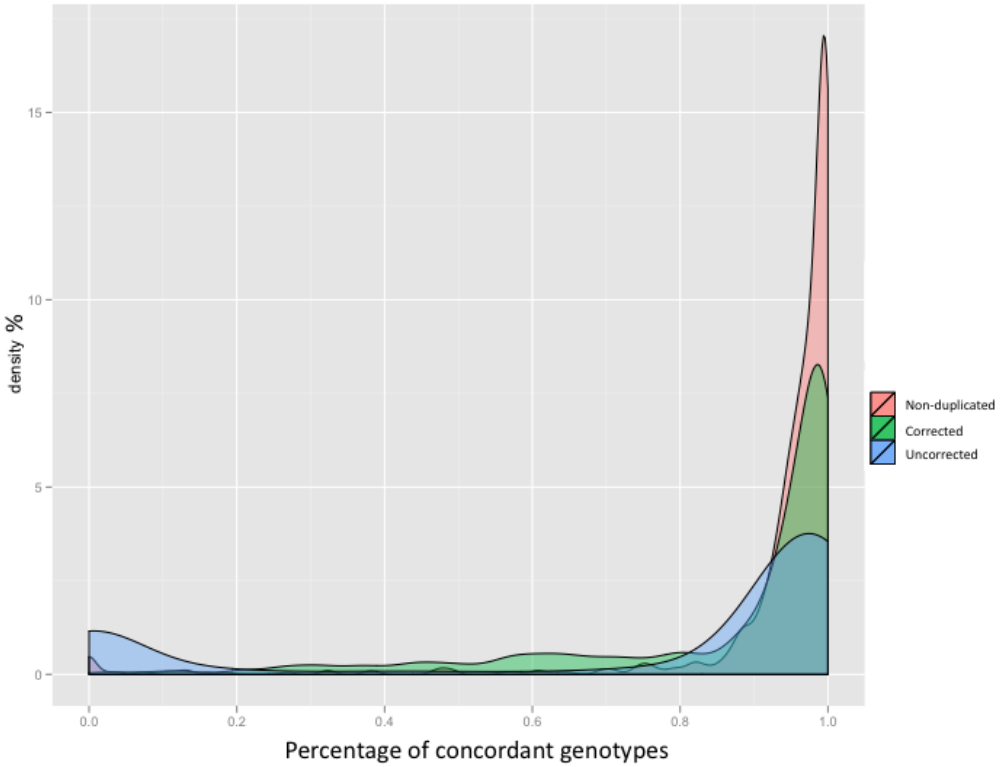


Figure S22. Genotype copy number concordance. Fraction of genotyping calls concordant between absolute read depth estimates and the Affymetrix SNP 6.0 Microarray (*S21*). 21% of calls show significant departure (uncorrected) but these map overwhelmingly to duplicated sequence. If adjusted (corrected) by a fixed integer amount reflective of the true population average copy number, the fraction of correct genotypes increases. Within unique regions, 94% of genotype calls are in accord.

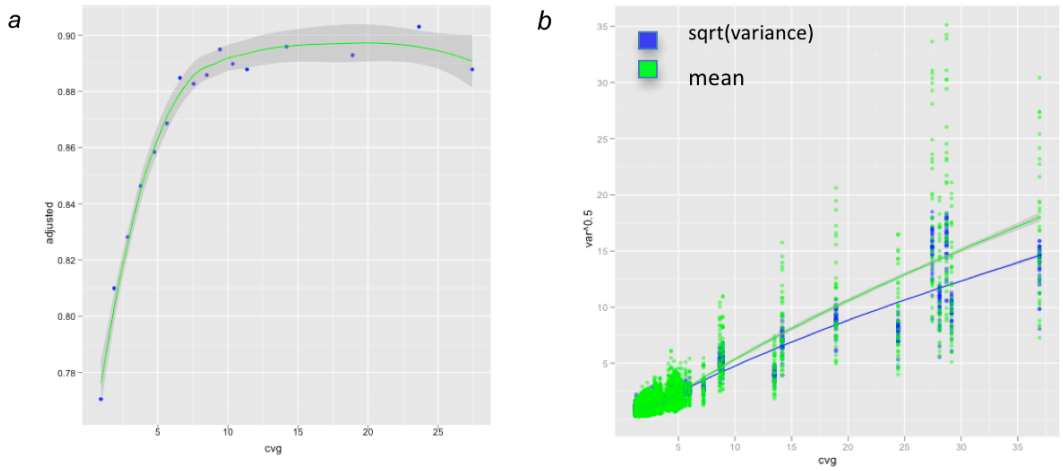
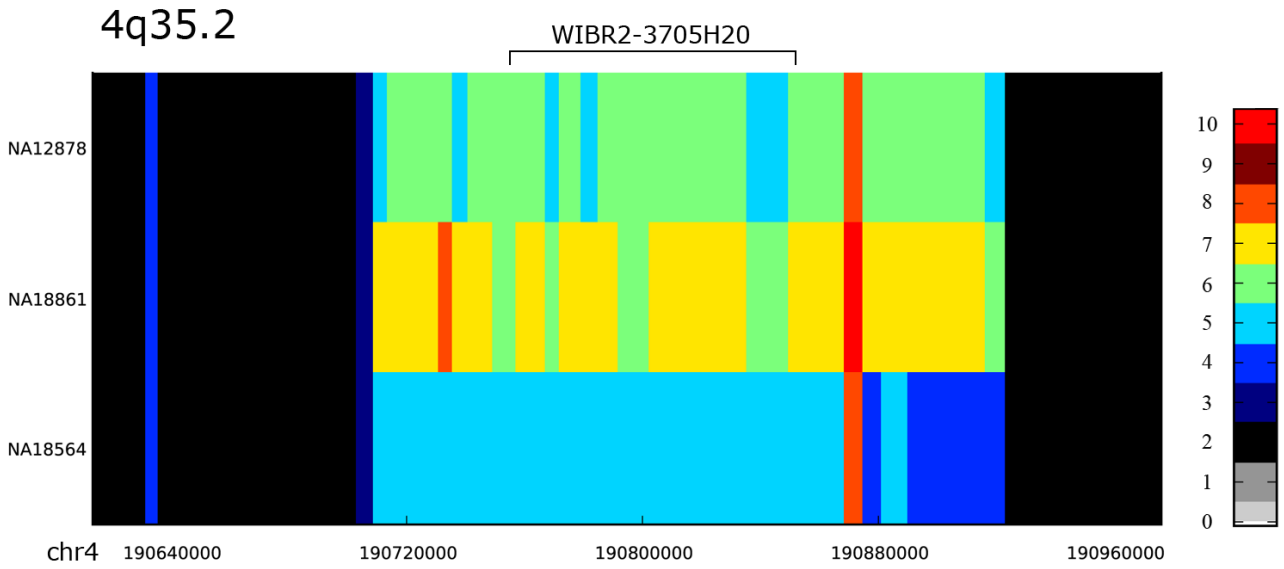
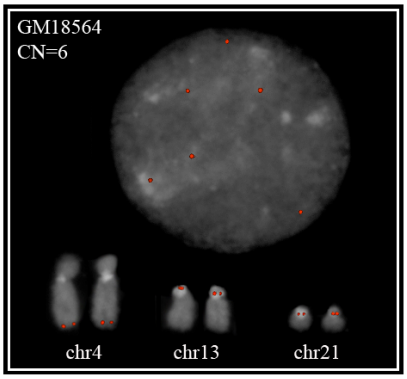


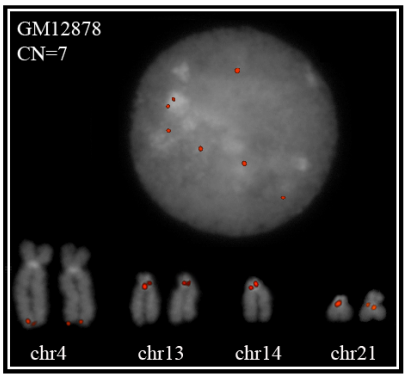
Figure S23. a) Concordance of genotypes with McCarroll *et al.* calls (*S21*) for NA18507 subsampled to various depths of coverage. b) Mean and standard deviation of read depth in 1-kb windows across copy number invariant (copy 2) region plotted as a function of coverage.l

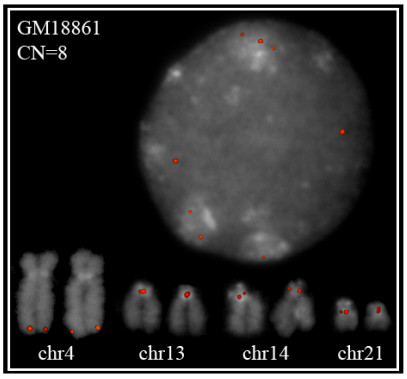
Figure S24. FISH validation (below). a) 12 examples of computational prediction and FISH validation of rare events seen in a single individual and duplications predicted to be duplicated in all individuals but represented as single copy in the human reference. Each line in the heat maps represents a sample for which the computational predictions were generated. b) Read depth-based predictions and FISH images of two highly copy number variable regions where individuals with the most extreme copy number difference were selected. In these cases FISH failed to accurately estimate copy number differences between individuals since the signals visualized on interphase nuclei were too numerous to provide an exact number of copies.

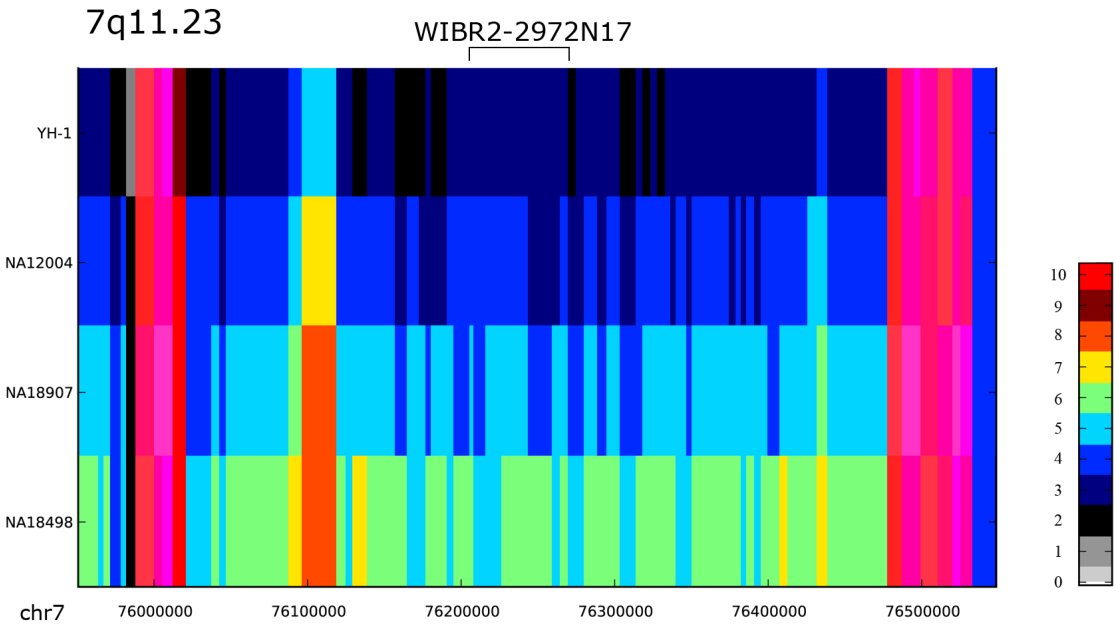
a)

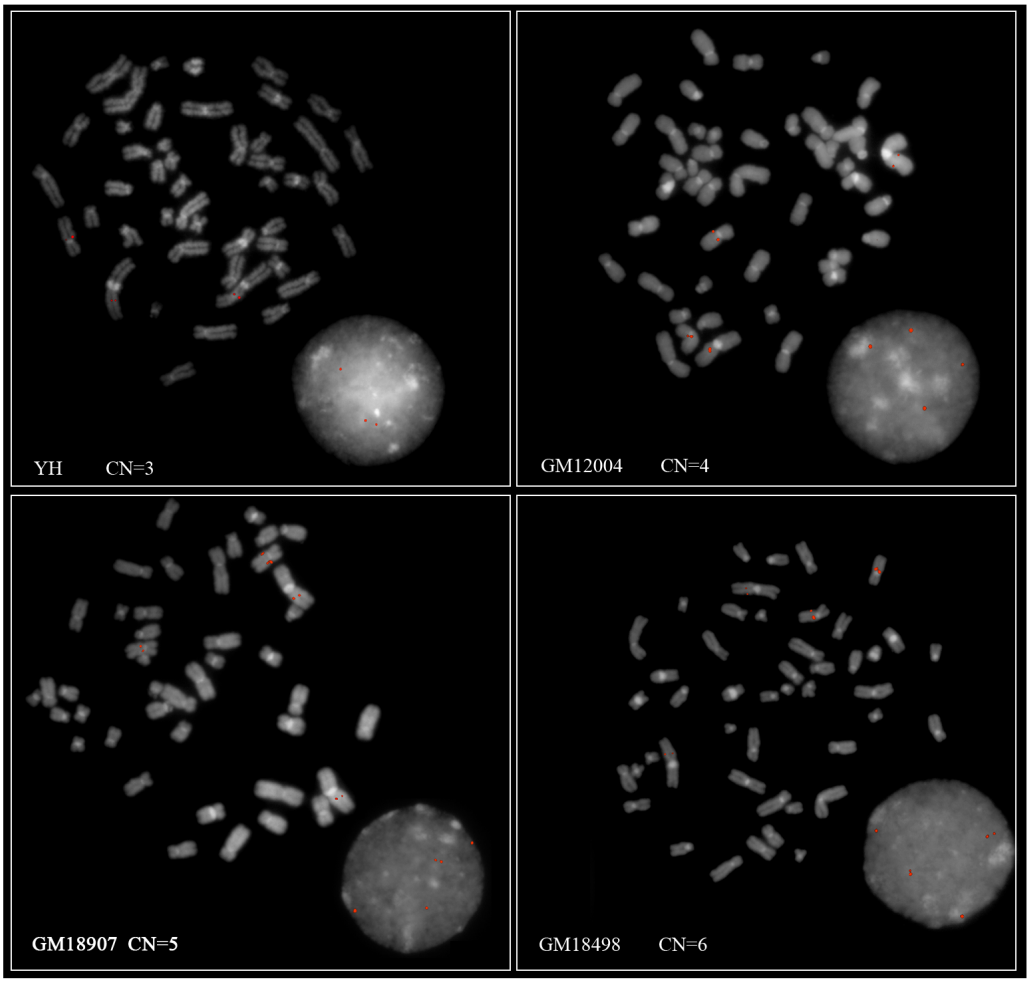


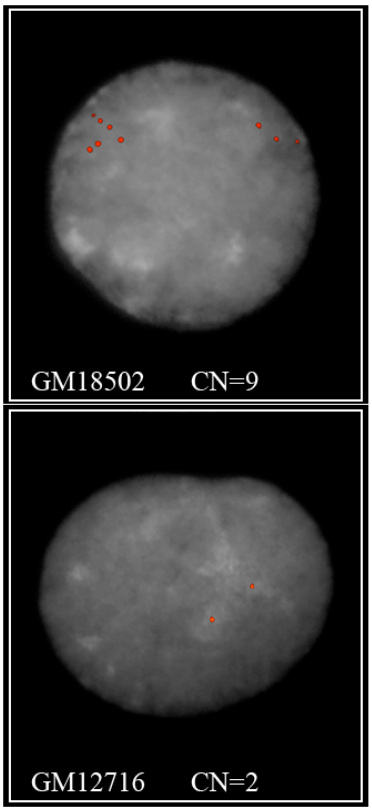


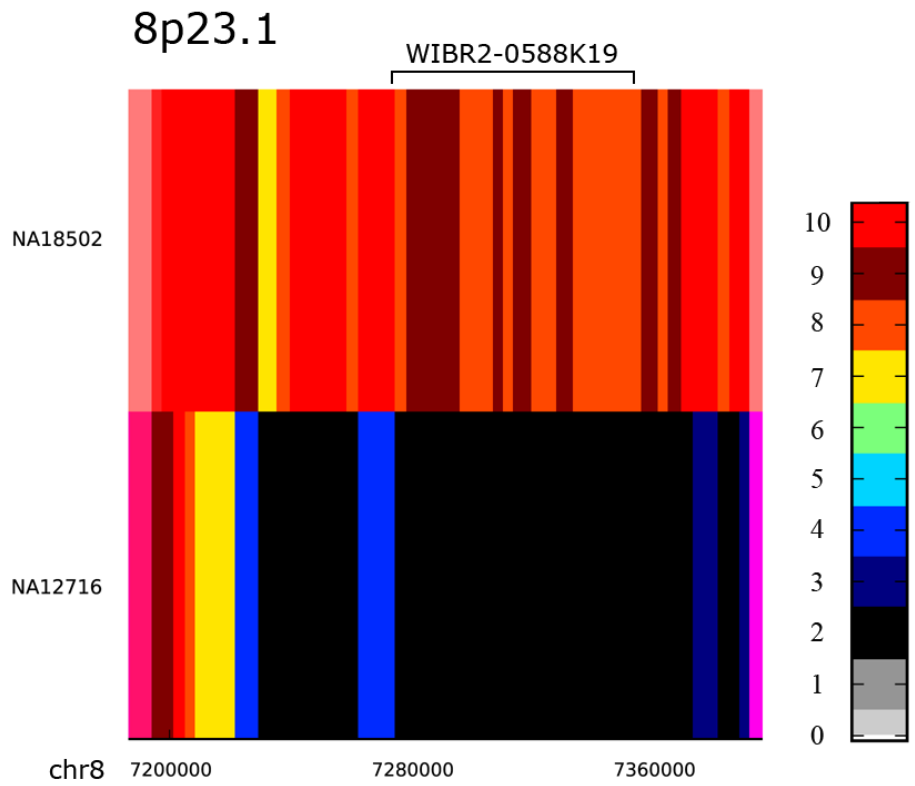


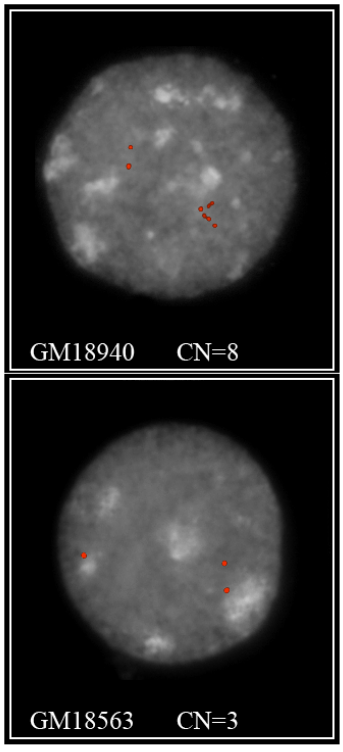


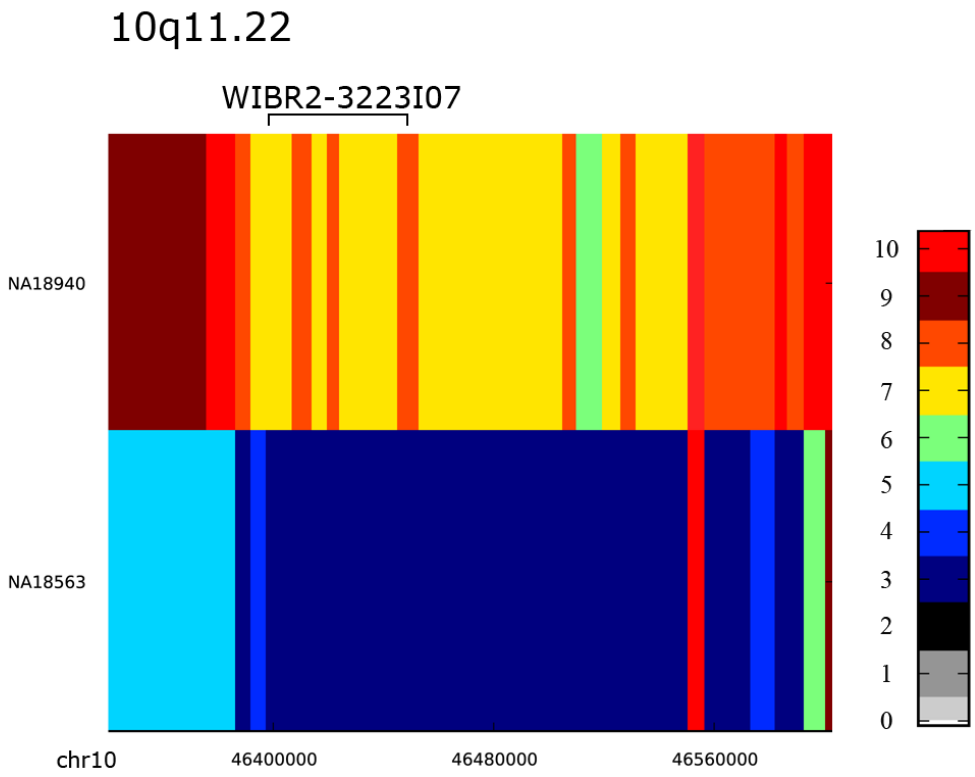

7q11.23
WIBR2-2972N17
YH-1
NA12004
NA18907
NA18498
chr7
76000000
76100000
76200000
76300000
76400000
76500000
10
9
8
7
6
5
4
3
2
1
0

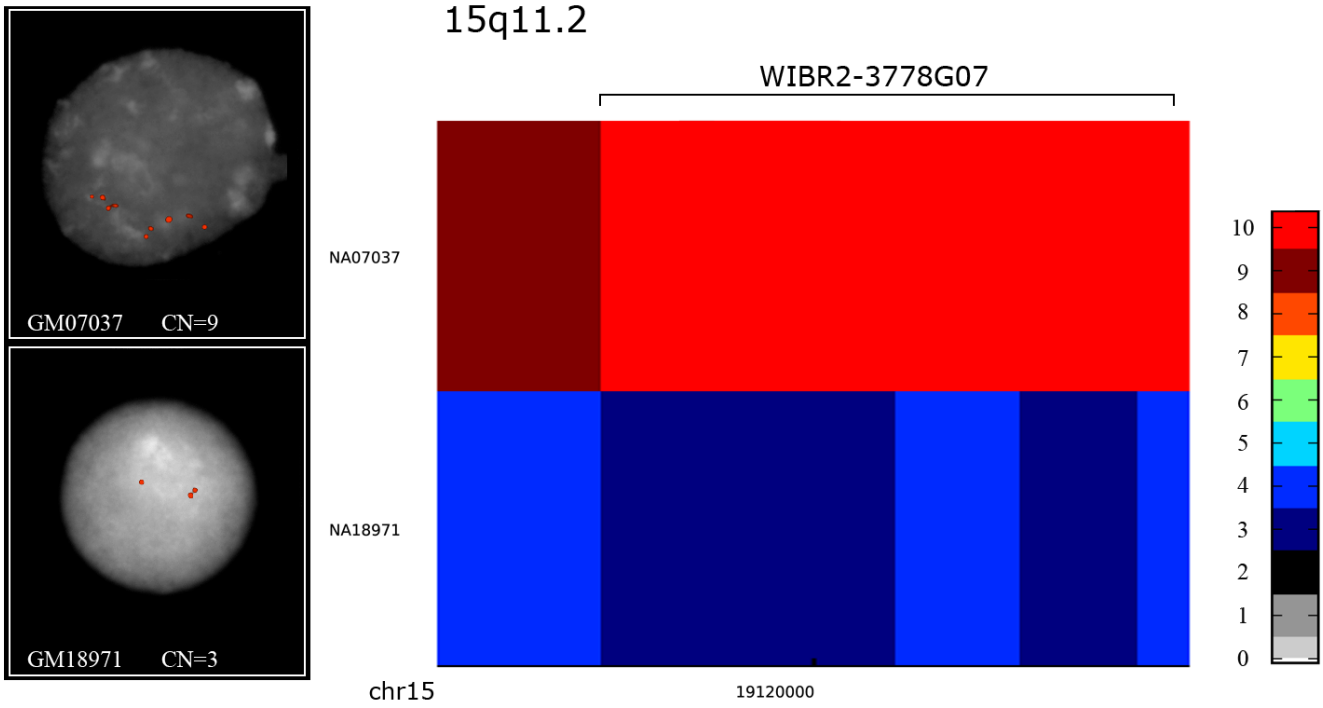

YH CN=3
GM12004 CN=4
GM18907 CN=5
GM18498 CN=6

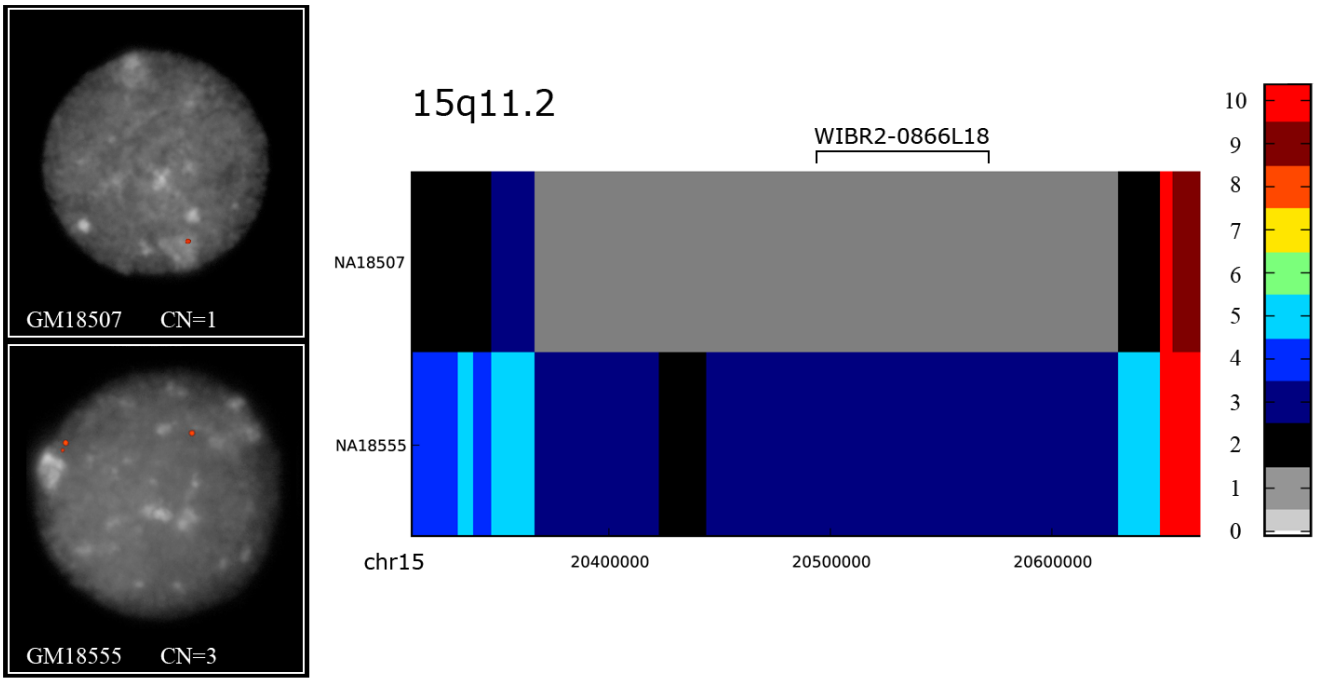

GM18502 CN=9
GM12716 CN=2

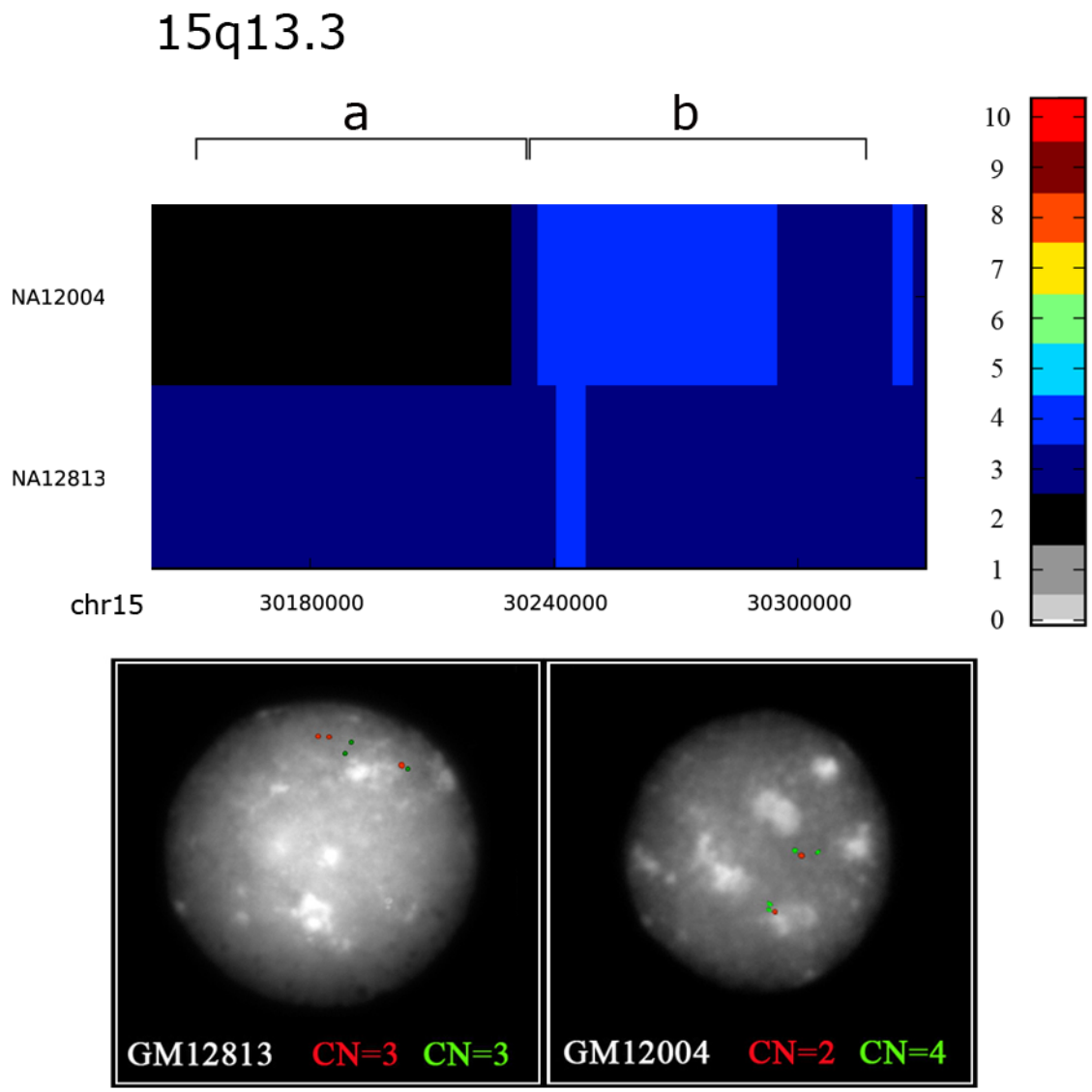

8p23.1
WIBR2-0588K19
NA18502
NA12716
chr8
7200000
7280000
7360000
10
9
8
7
6
5
4
3
2
1
0


GM18940 CN=8
GM18563 CN=3

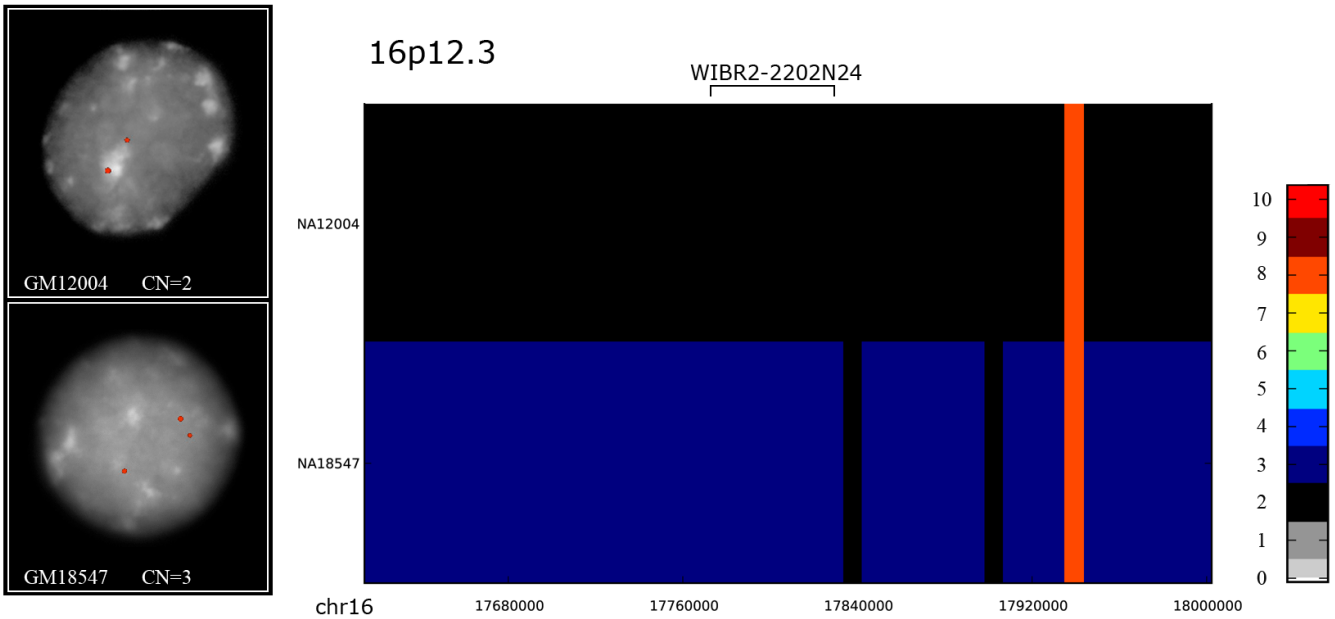

10q11.22
WIBR2-3223I07
NA18940
NA18563
chr10
46400000
46480000
46560000
10
9
8
7
6
5
4
3
2
1
0

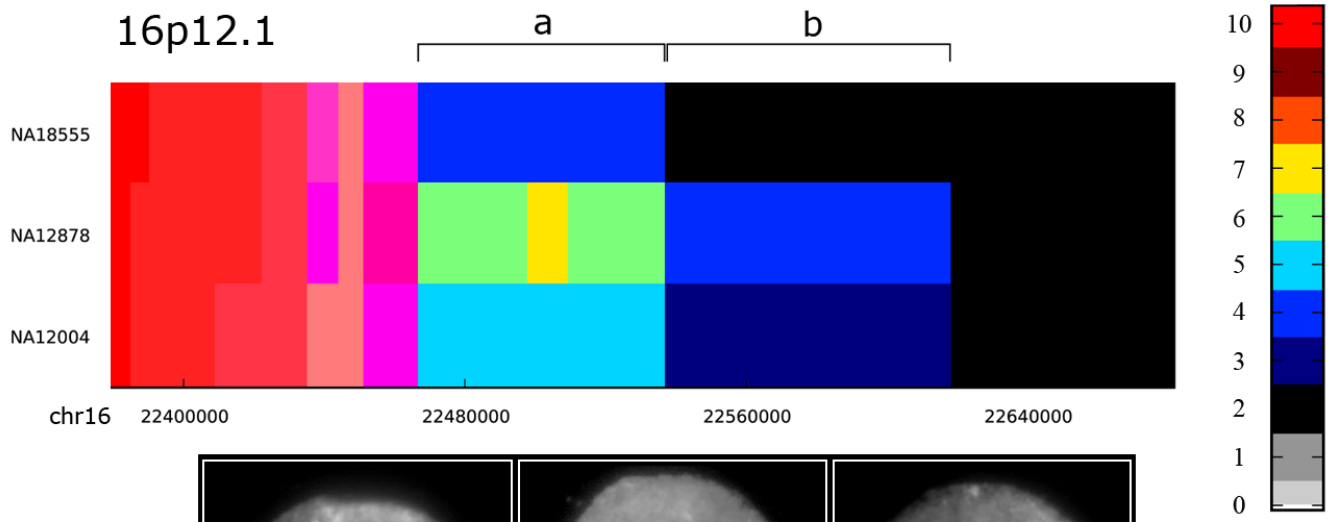

15q11.2
WIBR2-3778G07
GM07037 CN=9
GM18971 CN=3
NA07037
NA18971
chr15
19120000
10
9
8
7
6
5
4
3
2
1
0

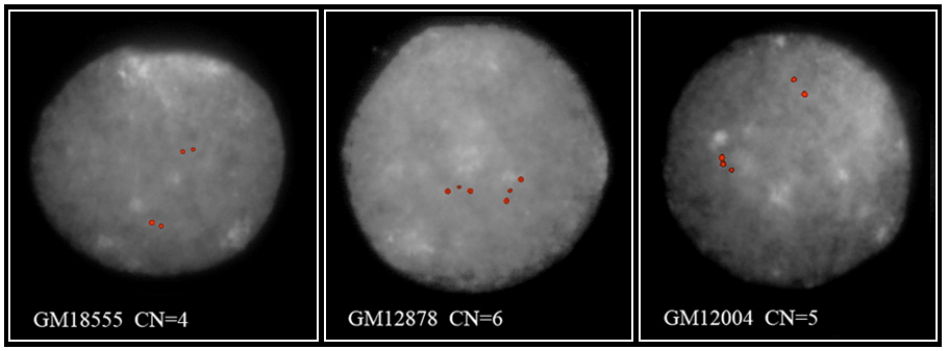

15q11.2
WIBR2-0866L18
GM18507 CN=1
GM18555 CN=3
NA18507
NA18555
chr15
20400000
20500000
20600000
10
9
8
7
6
5
4
3
2
1
0



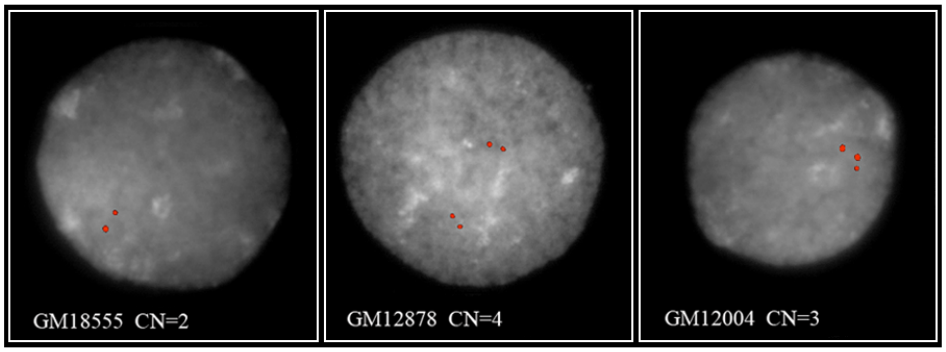
a) WIBR2-1228I16 (red)
b) WIBR2-2841A03 (green)


GM12004 CN=2
GM18547 CN=3
16p12.3
WIBR2-2202N24
NA12004
NA18547
chr16
17680000
17760000
17840000
17920000
18000000
10
9
8
7
6
5
4
3
2
1
0

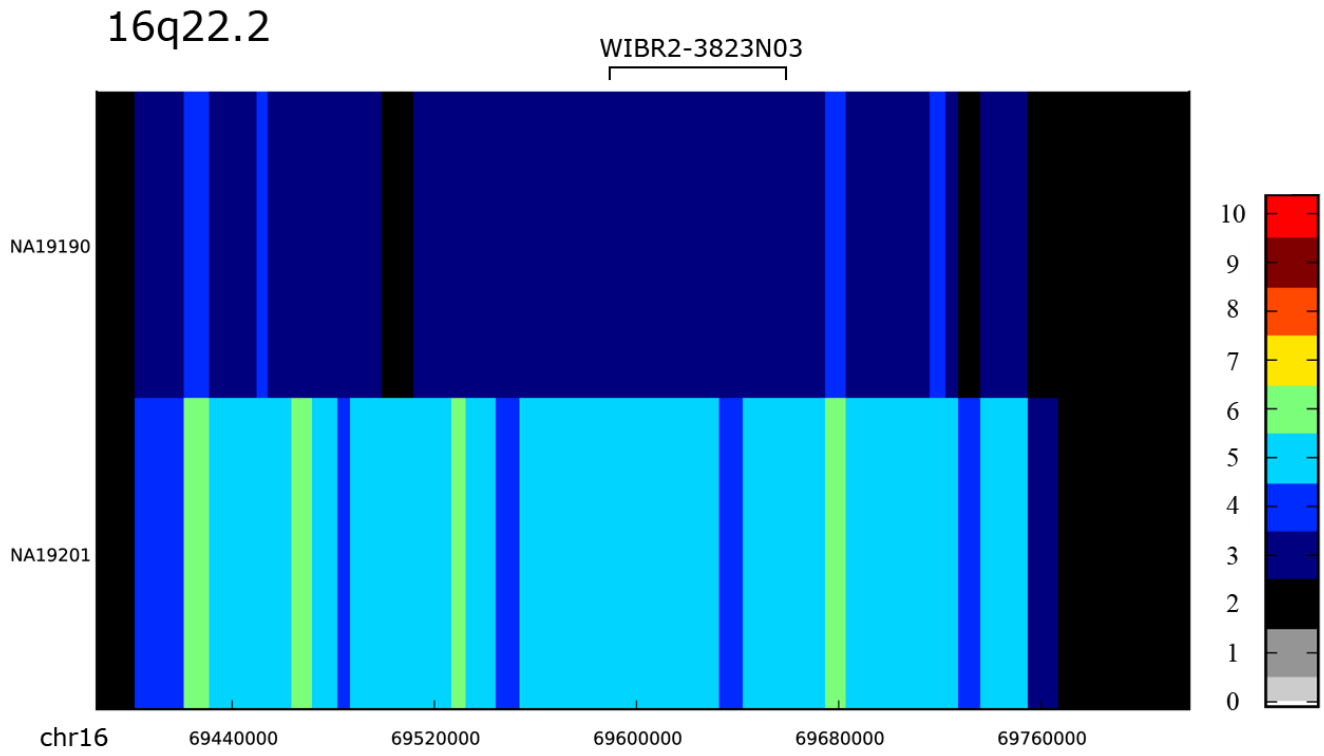


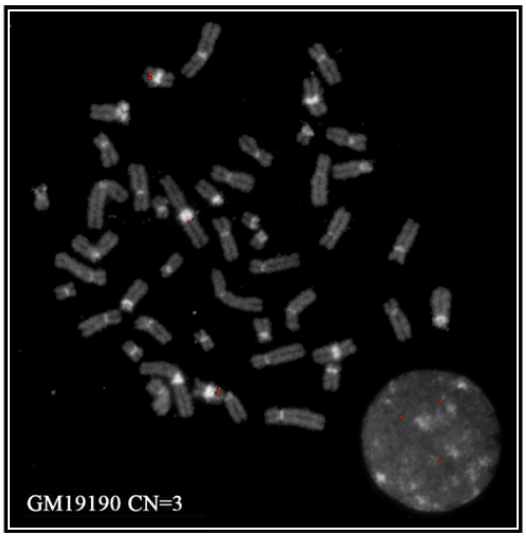


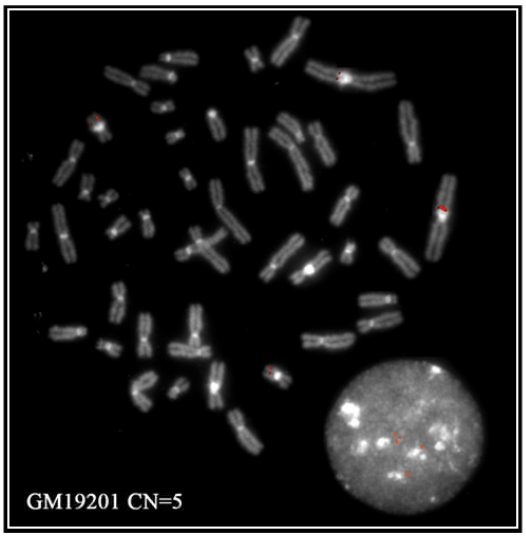
a)WIBR2-2031K01

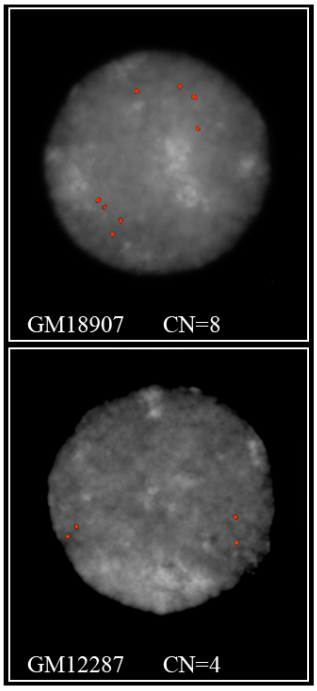


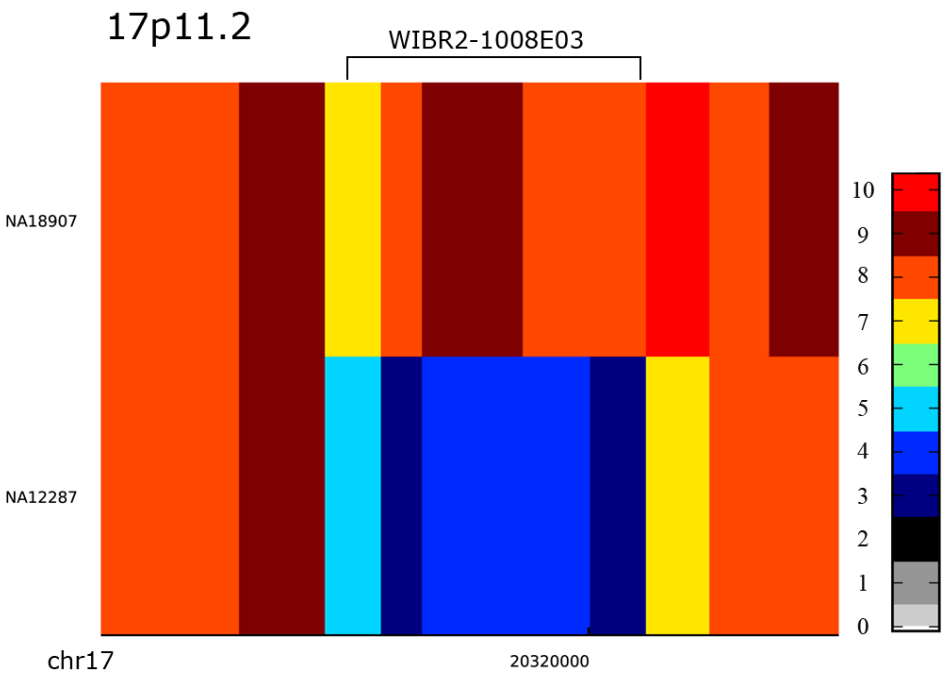
b)WIBR2-3608M06

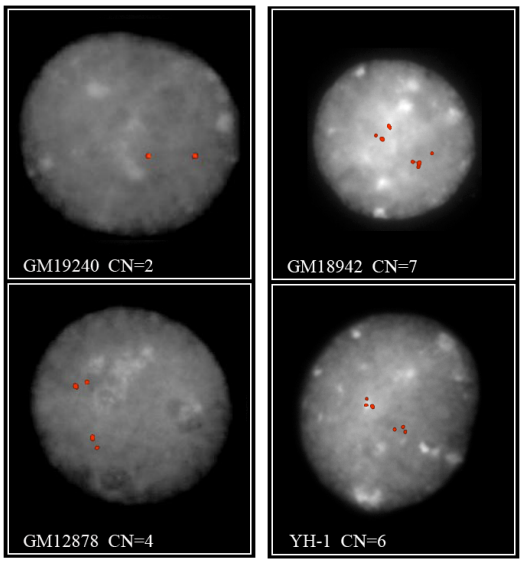

16q22.2
WIBR2-3823N03
NA19190
NA19201
10
9
8
7
6
5
4
3
2
1
0
chr16
69440000
69520000
69600000
69680000
69760000

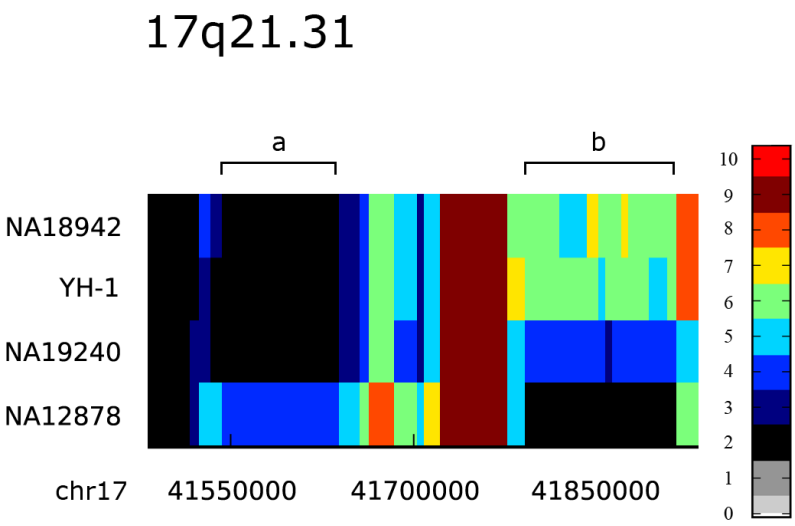

GM19190 CN=3


GM19201 CN=5



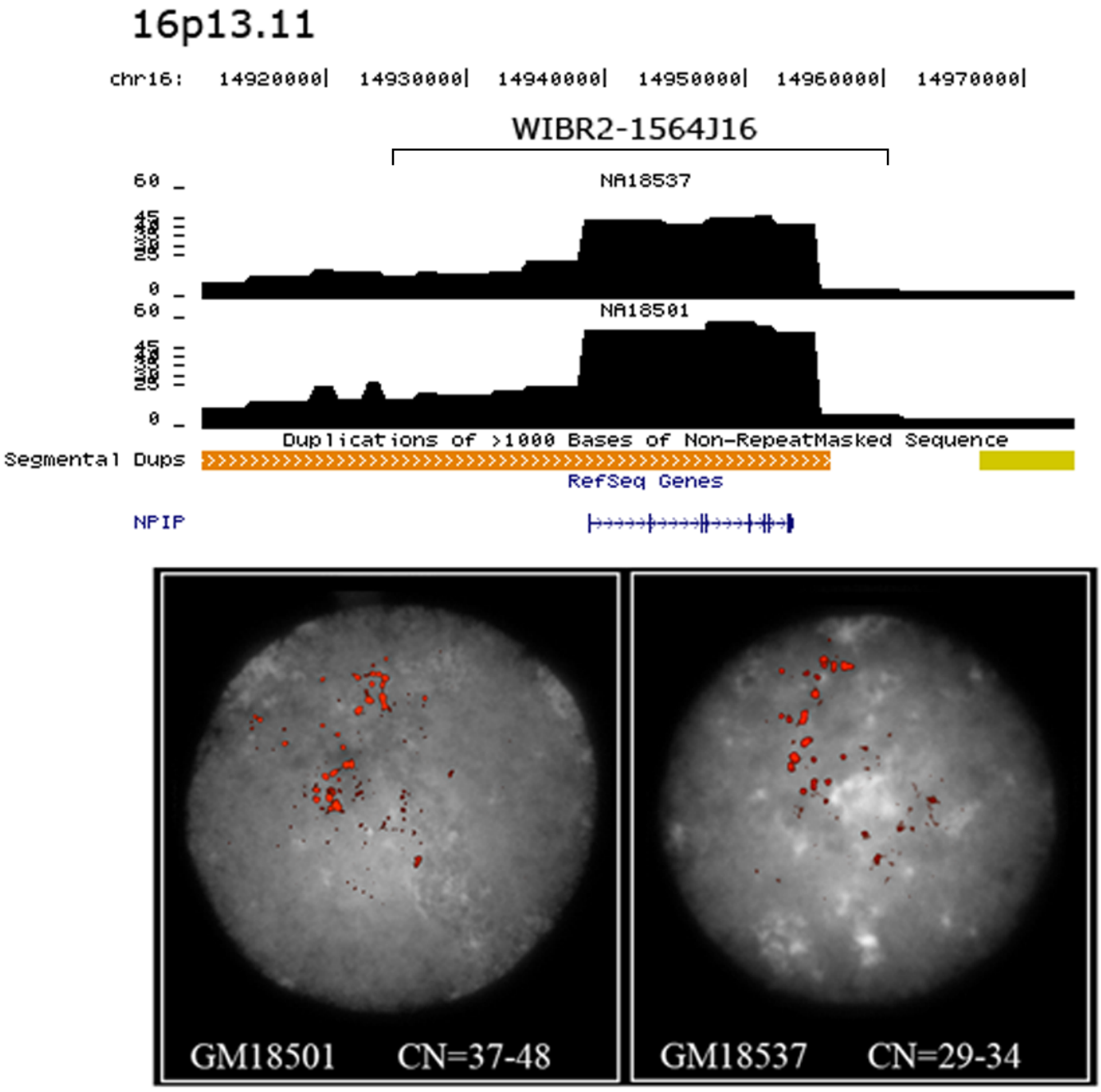


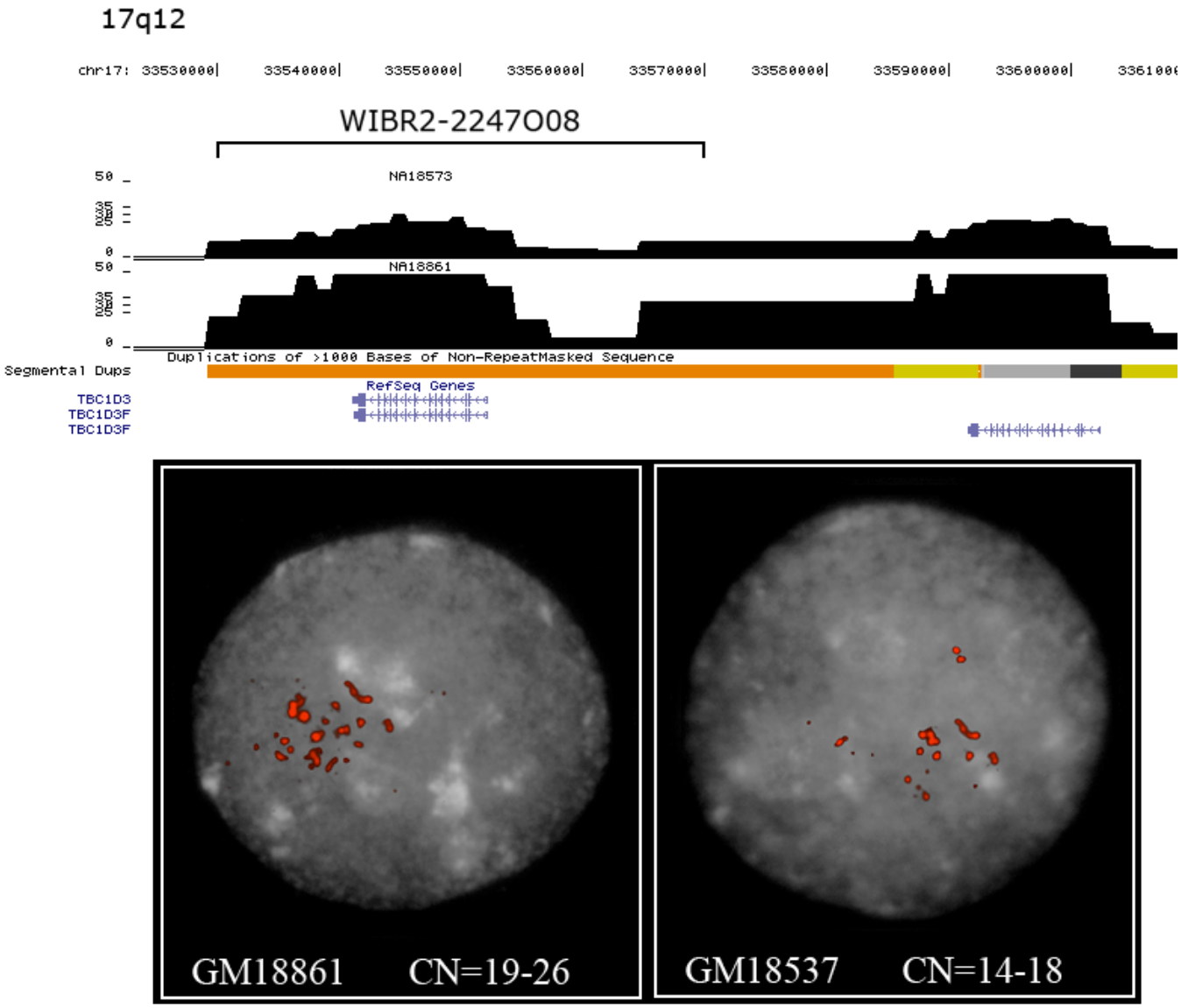




a) WIBR2-2004M23 b) WIBR2-1857K04

b)


16p13.11
chr16: 14920000| 14930000| 14940000| 14950000| 14960000| 14970000|
WIBR2-1564J16
NA18537
NA18501
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups
RefSeq Genes
NPIP
GM18501 CN=37-48
GM18537 CN=29-34


17q12
chr17: 33530000 33540000 33550000 33560000 33570000 33580000 33590000 33600000
WIBR2-2247O08
NA18573
NA18861
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups
RefSeq Genes
TBC1D3
TBC1D3F
TBC1D3F
GM18861 CN=19-26
GM18537 CN=14-18

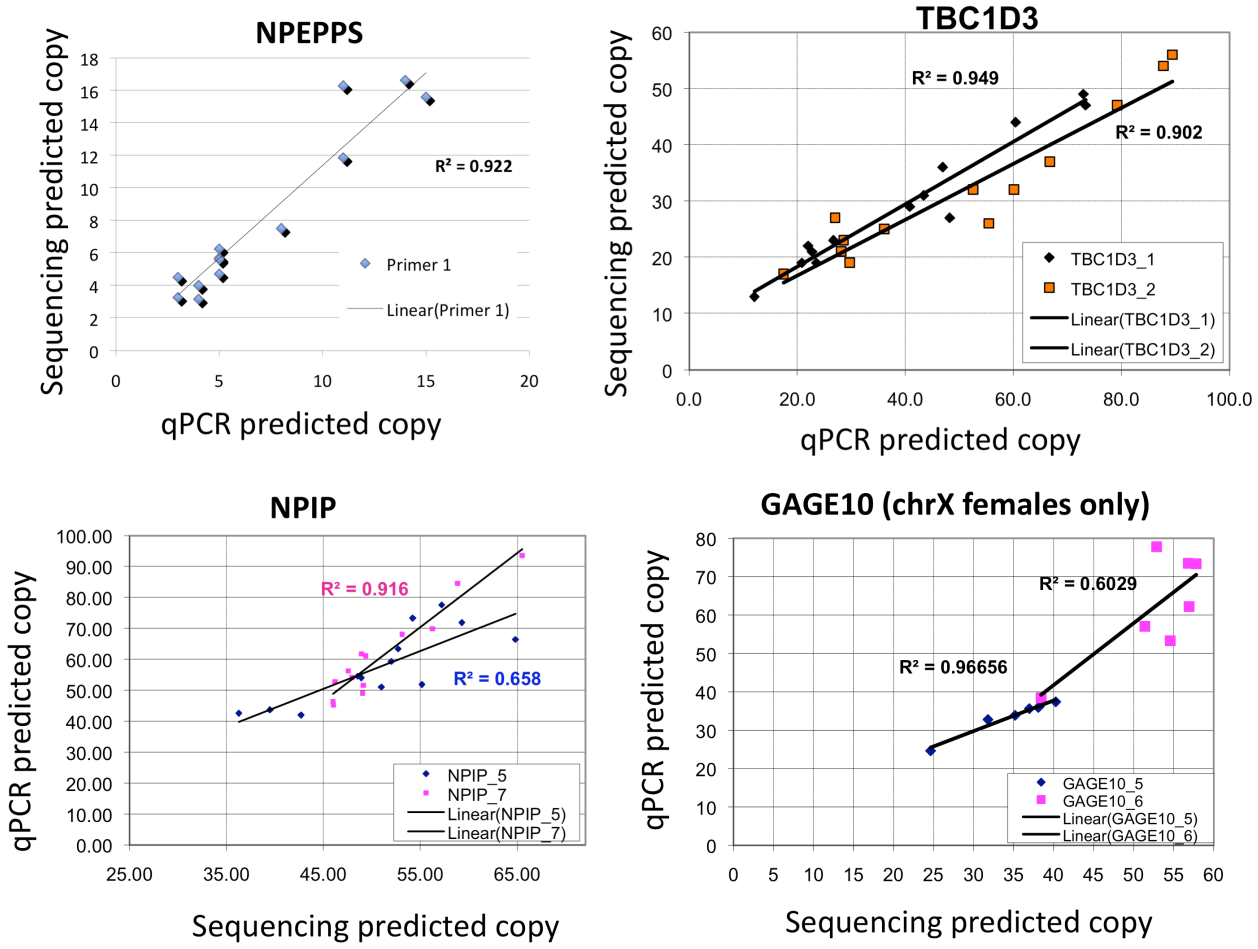


Figure S25. qPCR validation experiments for four genic loci showing strong correlation with sequencing-based copy number for a variety of ranges. Seven out of nine experiments had R^2 >0.84.

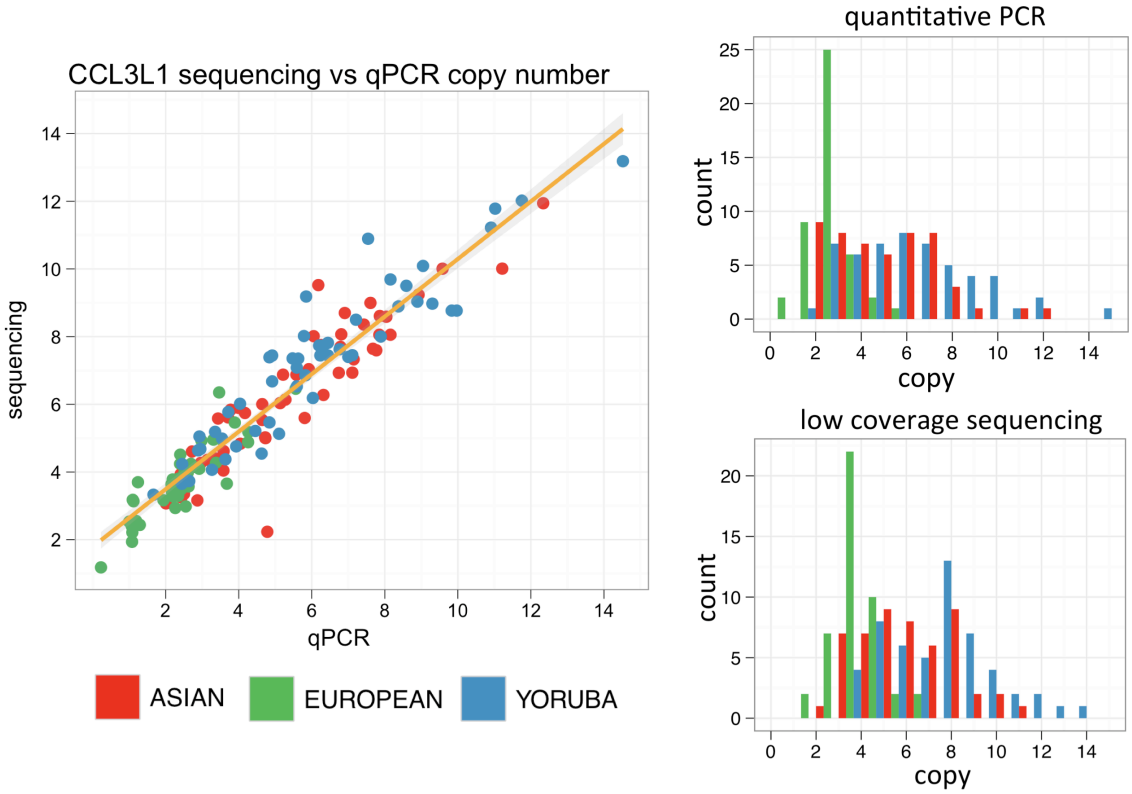


Figure S26. qPCR-based copy number genotyping is highly correlated with sequencing-based copy number estimates (r = 0.95). Sequencing-based copy number genotypes capture both the specific copy number range and population stratification of this locus.

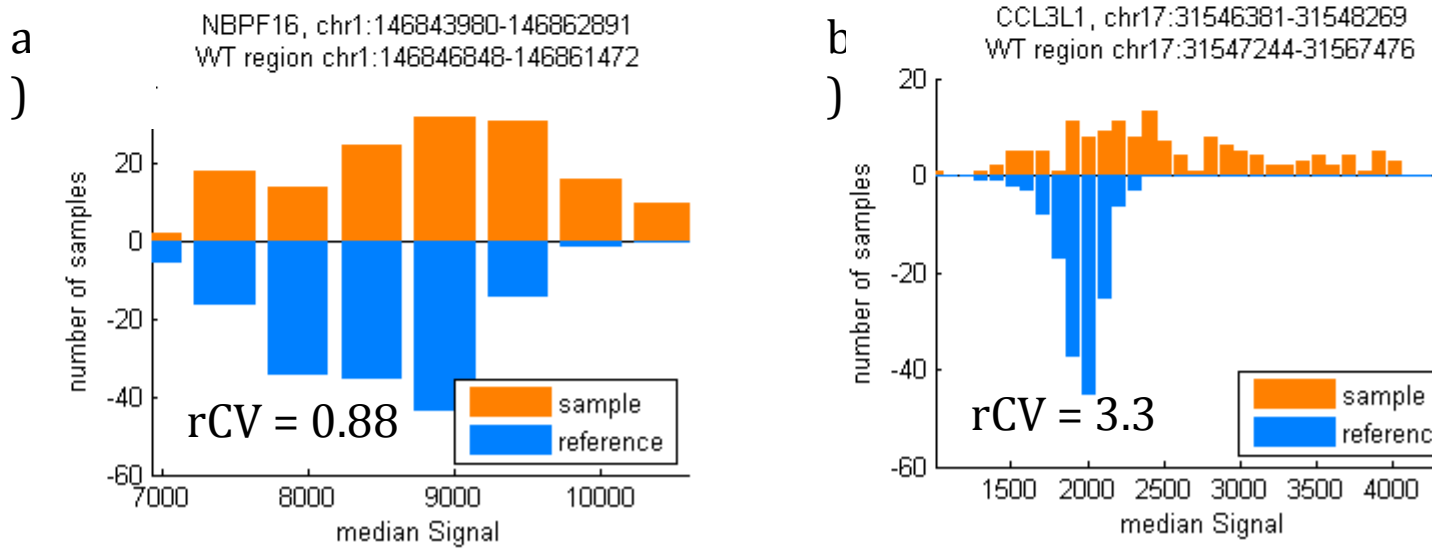


Figure S27. The median single-channel intensity signals for the sample and reference are shown for two loci. Copy number variation of an individual sample can only be detected if the signal variation between the sample and reference is detected. The rCV statistic summarizes the amount of variation observed between the sample and reference genomes.

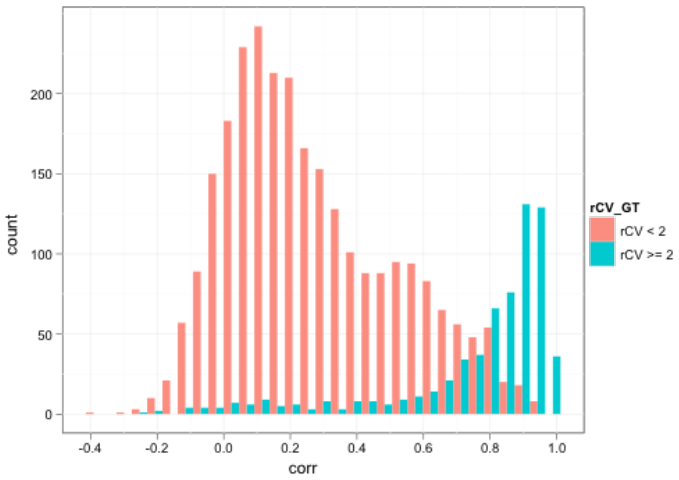


Figure S28. The rCV metric indicates if a region on the array can accurately assay the spectrum of copy number variation at that locus. Histogram of correlations shown between sequencing and array genotypes for rCV <2 and rCV >=2. Above rCV 2, 77% of loci assayed have a correlation >0.7.

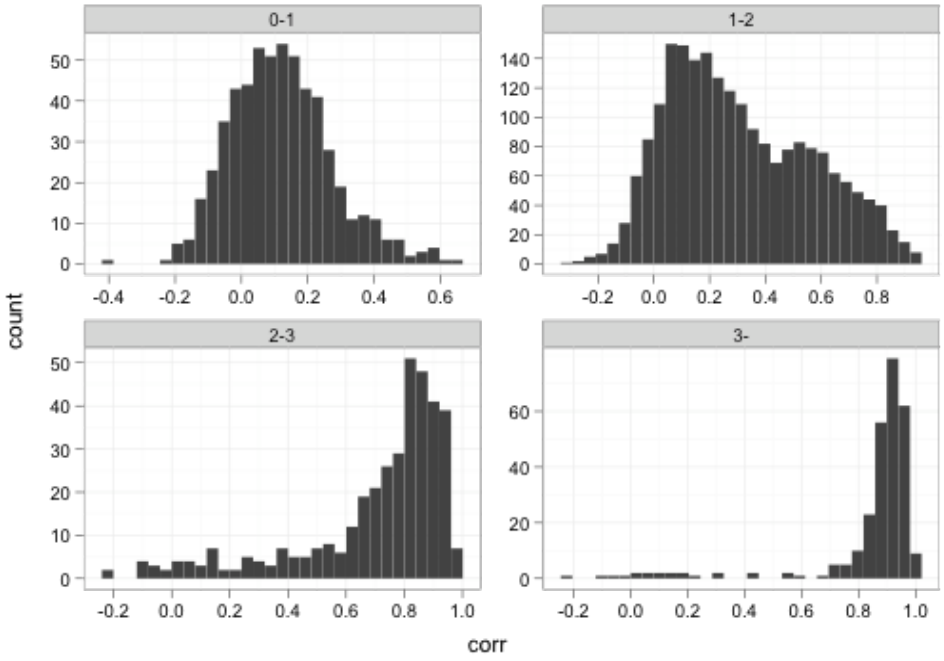


Figure S29. Histograms of correlation between sequencing-based copy number predictions and array-based copy number predictions for different ranges of rCV. At rCV >2 most regions have correlations of >=0.7.

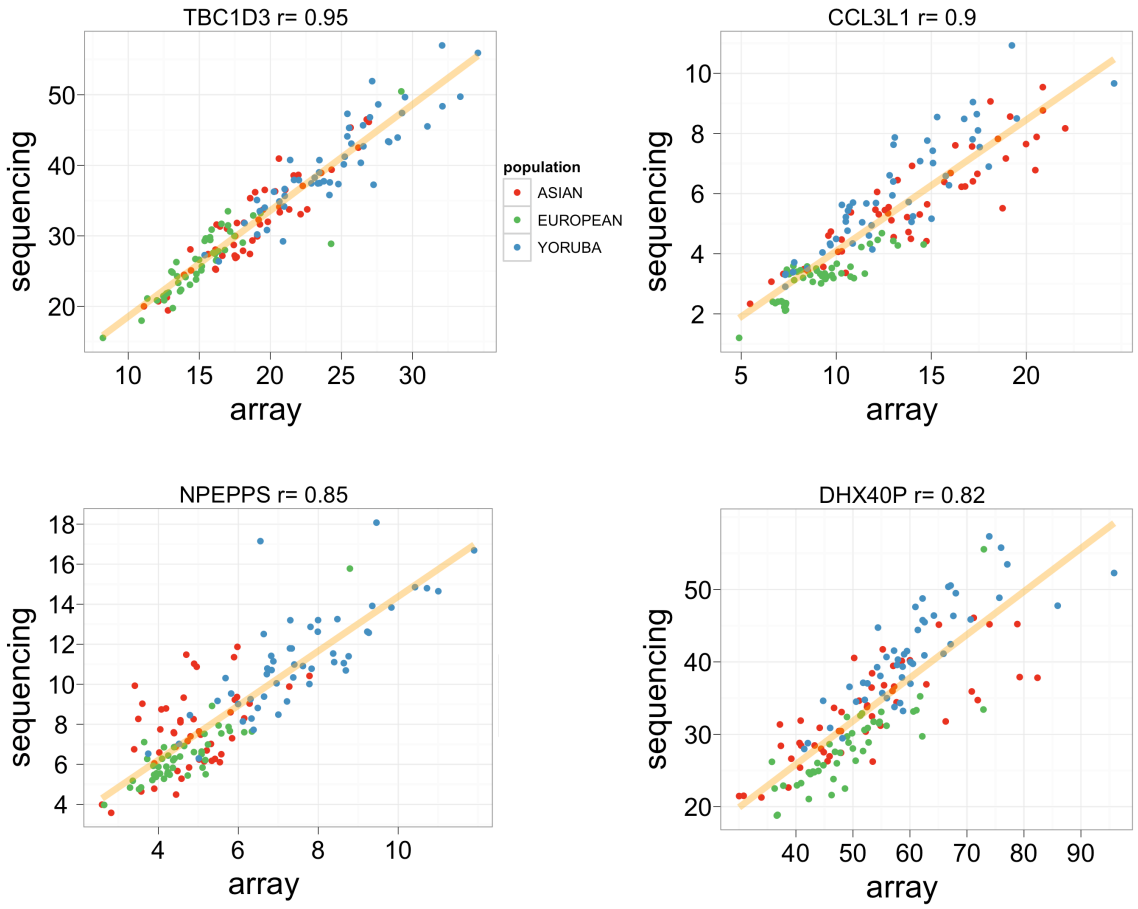


Figure S30. CGH-based copy number estimates in many regions are highly correlated with our read depth-based copy predictions. Read depth-based copy number estimates can then be used to calibrate CGH copy predictions on a per locus basis.

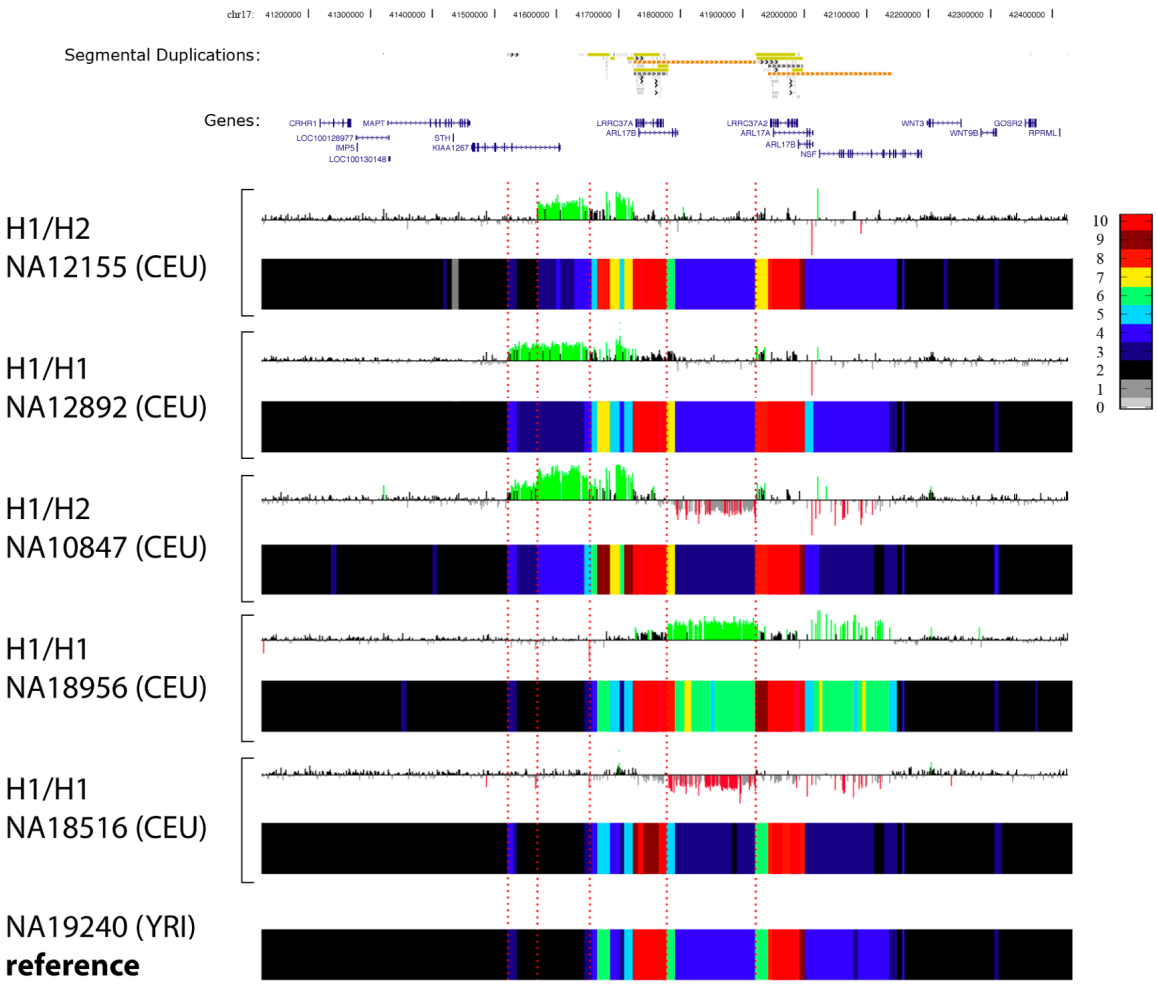


Figure S31.
Analysis of 159 individual genomes used to select a reference individual nearest the population-wide median copy number distribution to maximize discriminatory power in array CGH experiments. Array CGH using NA19240 as a reference sample is shown above read depth-based copy number estimates for chromosome 17q21.

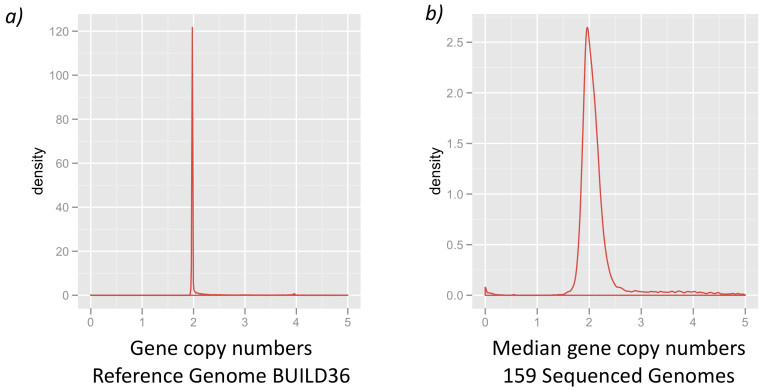


Figure S32. Density plots of the estimated copy number of all genes in the human genome for a) the reference genome Build36 versus b) the median over 159 sequenced individual genomes.

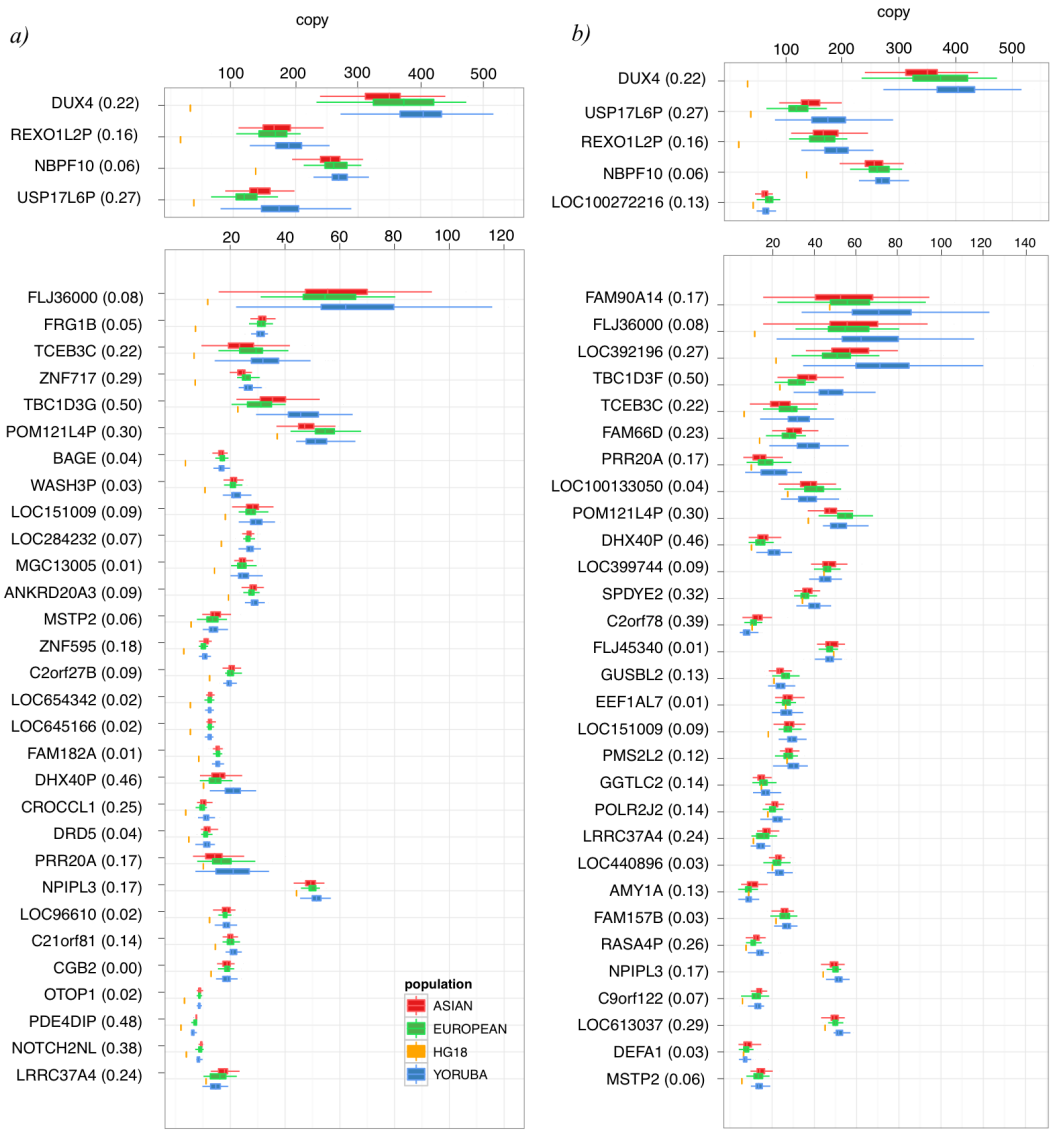


Figure S33. a) Genes missing from the human reference genome (Build36) and b) the most variable genes found among humans. Note that many of the genes with missing copies in the human reference are among the most variable genes within the human species.

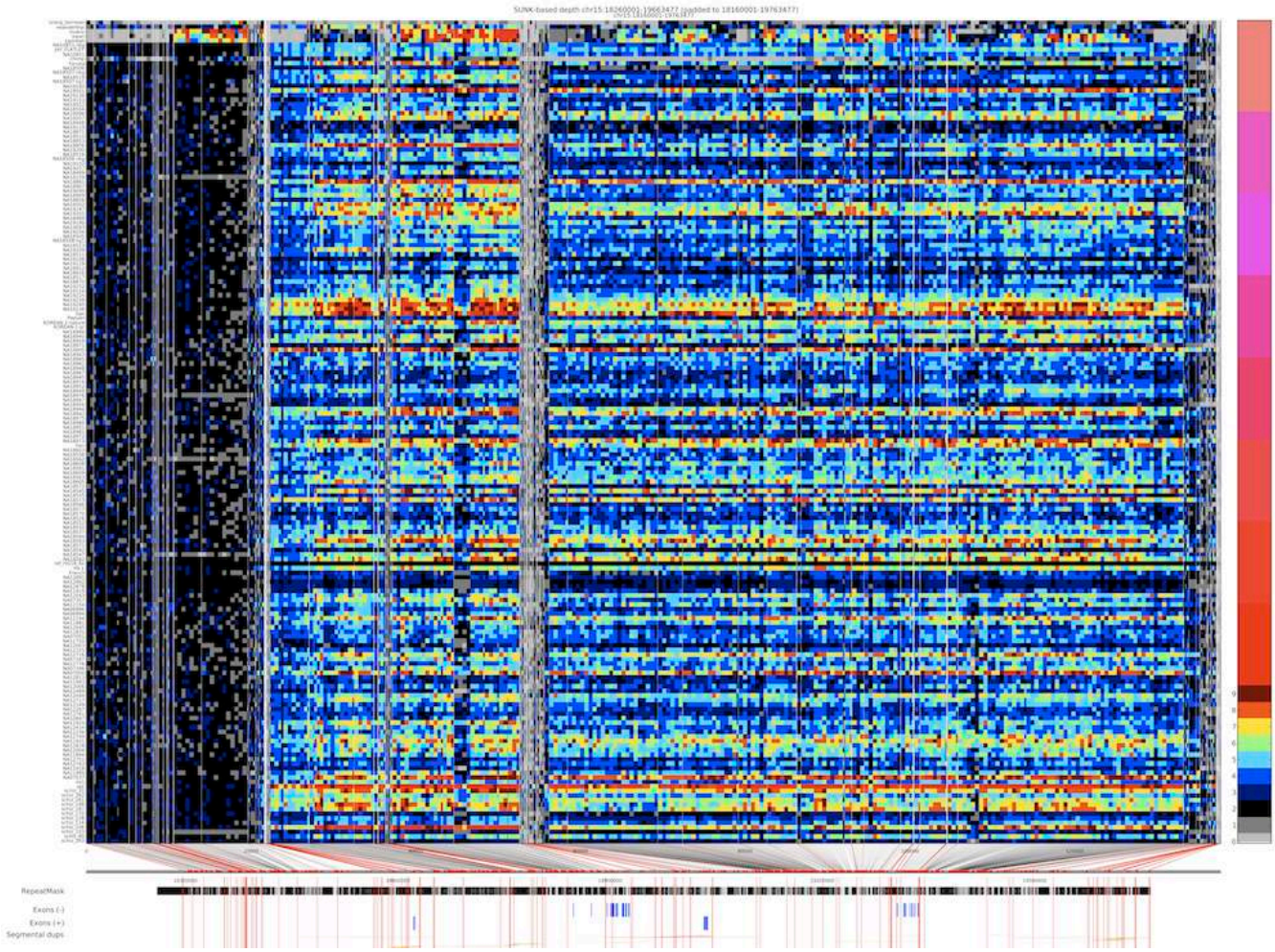


Figure S34. Complex patterns of variation in the copy number landscape of chromosome 15q11 and 15q12.

A highly copy-number polymorphic, ~900-kbp portion of this region (chr15:18160001-19763477) ranges from 2-11 copies among individuals sampled. The individuals at the extremes of this range thus differ by ~8.1 Mbp of euchromatic DNA harboring brain- and testis-expressed genes.

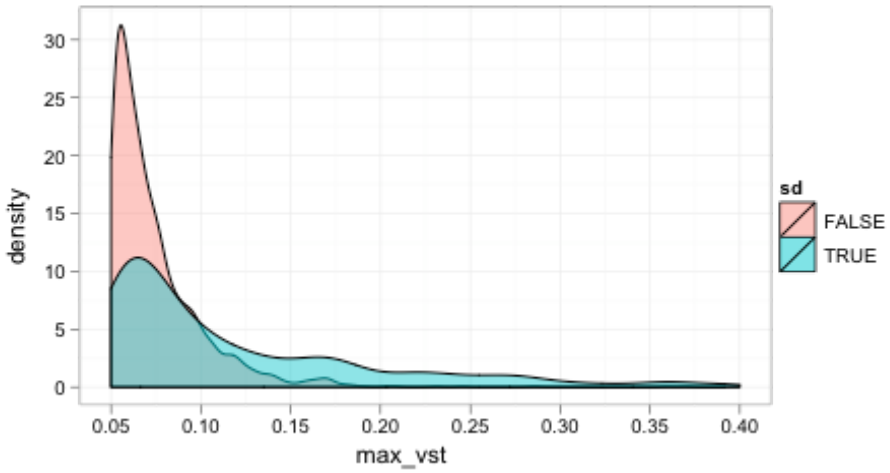


Figure S35. Distribution of the V_{st} statistic for genes with a V_{st} >0.05 classified by segmental duplication (sd) overlap. Genes of increased V_{st} are enriched for segmental duplications.

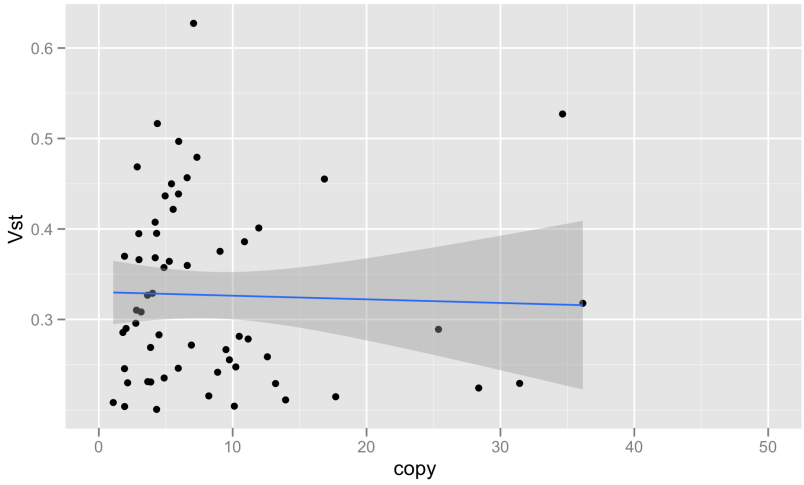


Figure S36. V_{st} plotted against copy. No correlation was observed between V_{st} and copy (r^2 = 0.01, P = 0.2).

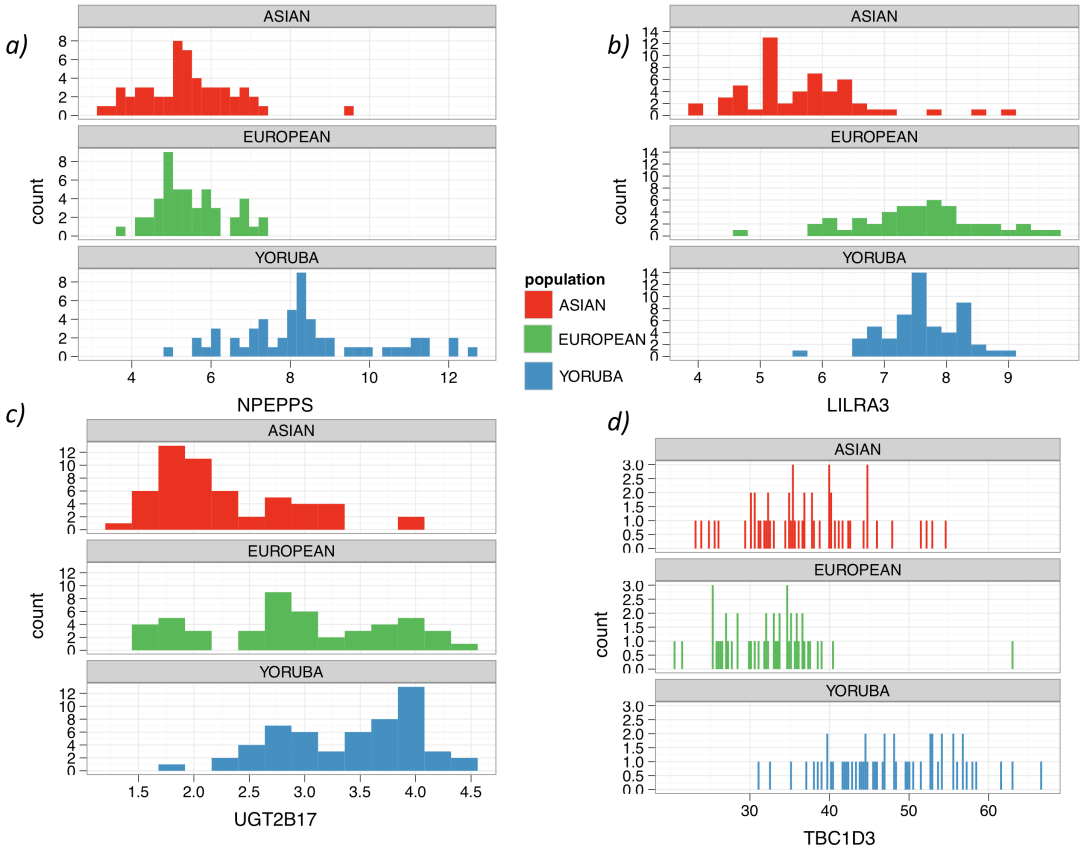


Figure S37. Population-specific copy number distributions for four genes with extreme population stratification of copy number.

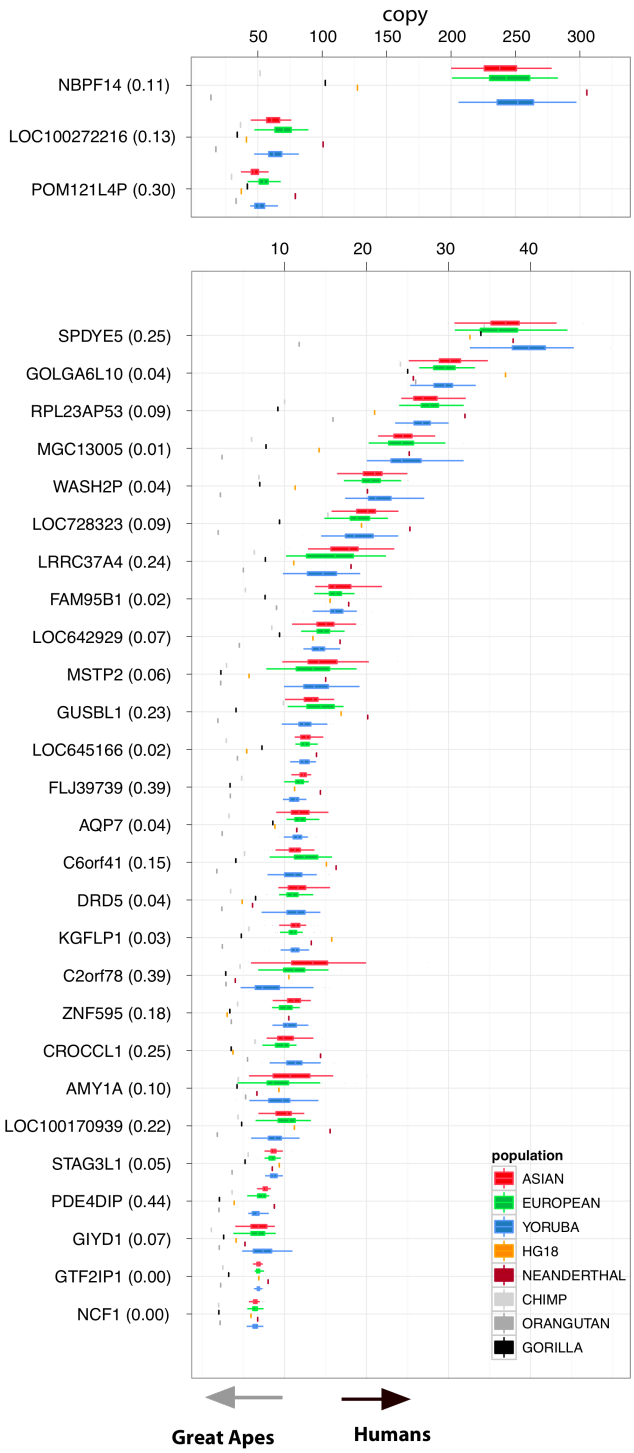


Figure S38. Box plots of 30 gene families identified as copy number expanded within the human lineage.

Among these genes we identified several implicated in brain function, including *DRD5* and *GTF2L1P1*. Additionally, we confirm known human-specific expansions such as those at the amylase gene cluster (*S34*).

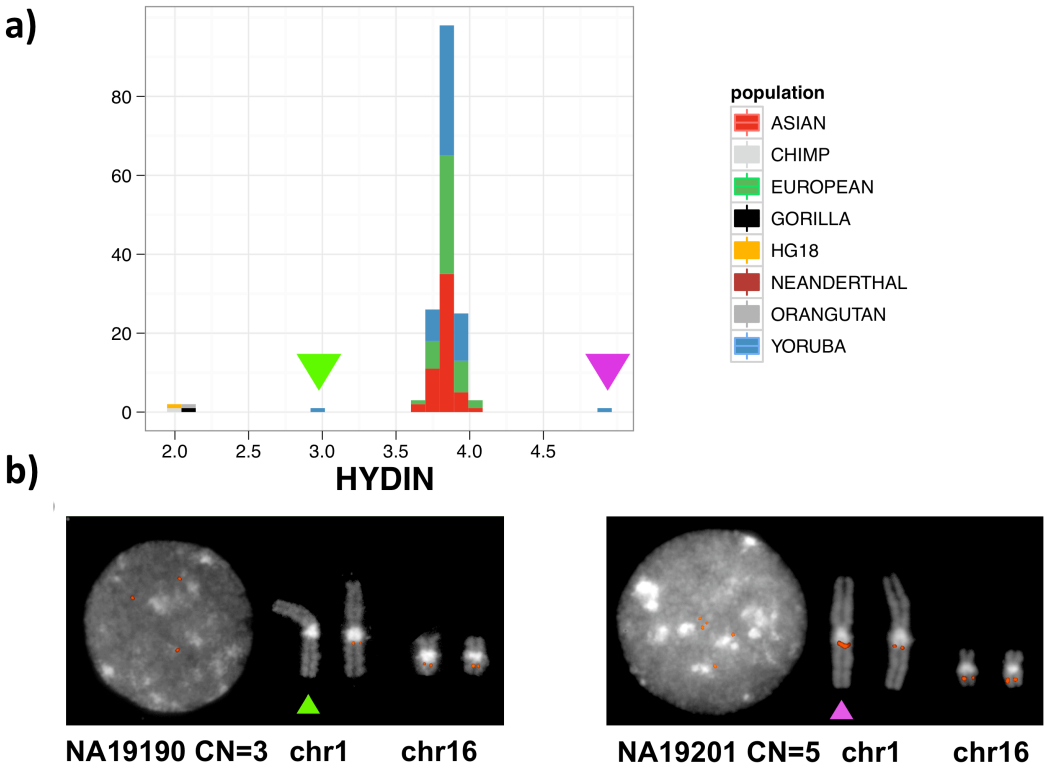


Figure S39. a) Copy number histogram of the hydin gene which is specifically duplicated in the human lineage. b) Two rare Hydin duplications/deletions are confirmed by FISH

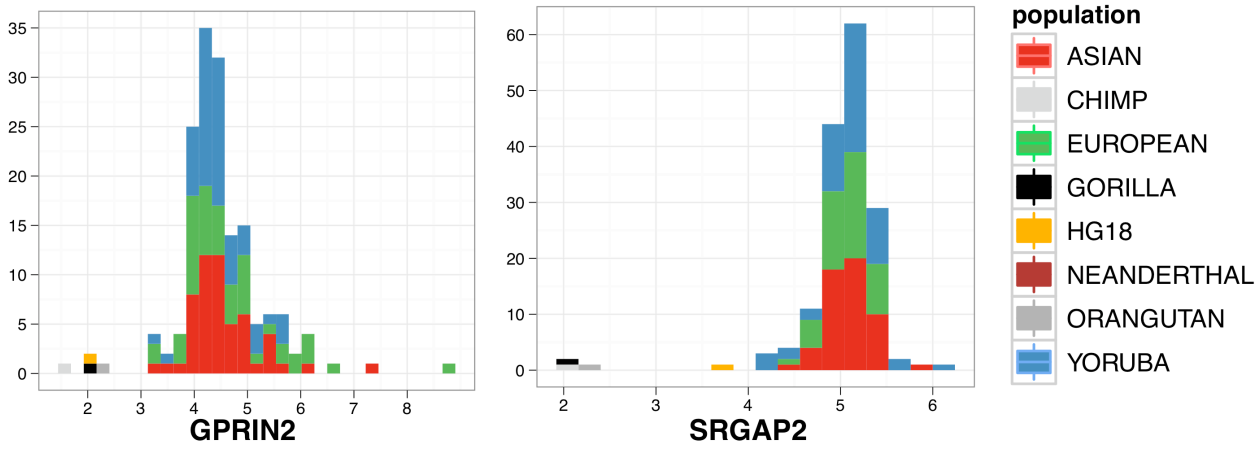


Figure S40. Copy number histograms of the human specific duplicated genes GPRIN2 and SRGAP2

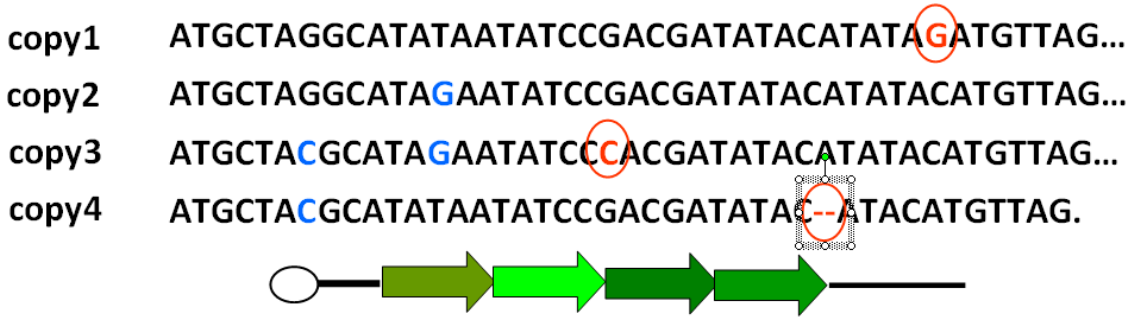


Figure S41. Schematic depicting singly unique nucleotide (SUN) positions used to distinguish one paralog among highly identical duplicates.

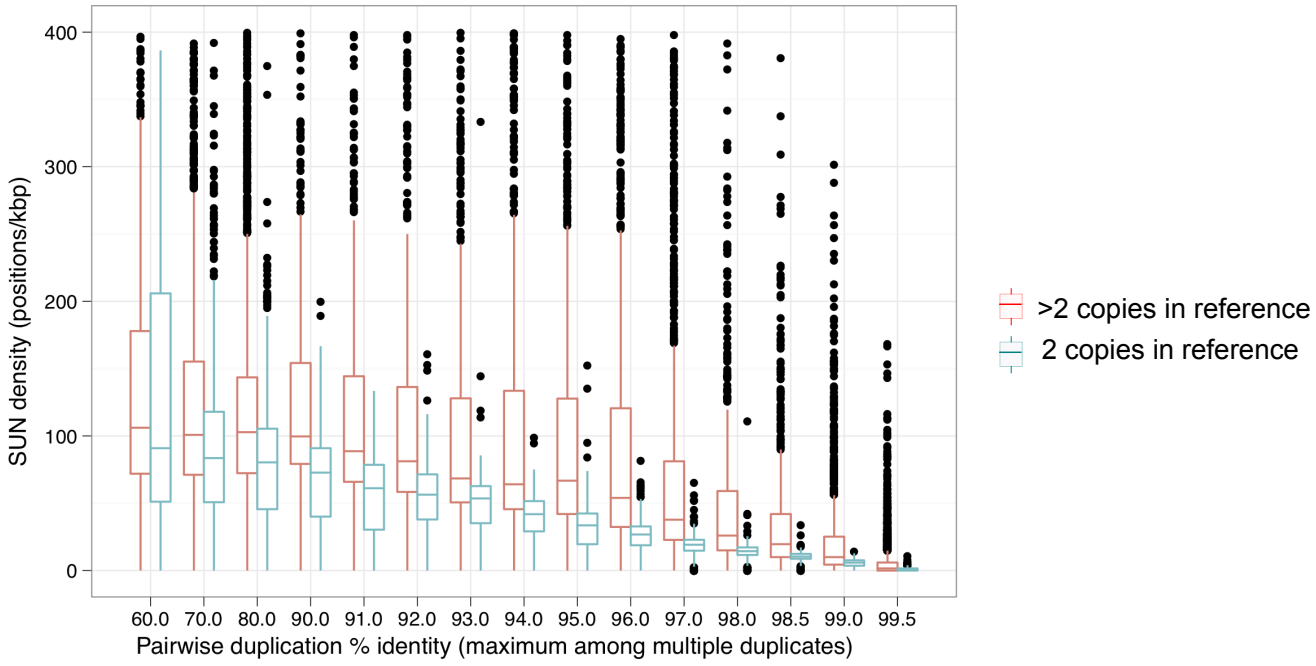


Figure S42. Density of SUN positions (average per 1 kbp) within segmental duplications as a function of duplication percent identity. Segmental duplications were divided into nonoverlapping windows of equal content (2500 bp unmasked sequence, variable physical width). Within each window, the number of SUNs was counted. Local percent identity was computed by extracting each window from the pairwise global alignment(s) between that window and the one or more duplicate copies elsewhere in the genome. When a given window has more than one duplicate locus, the highest percentage pairwise identity between it and its multiple paralogs is taken. Duplicated windows are stratified based on whether they are duplicated exactly once (blue) or at multiple copies in the reference genome (red).

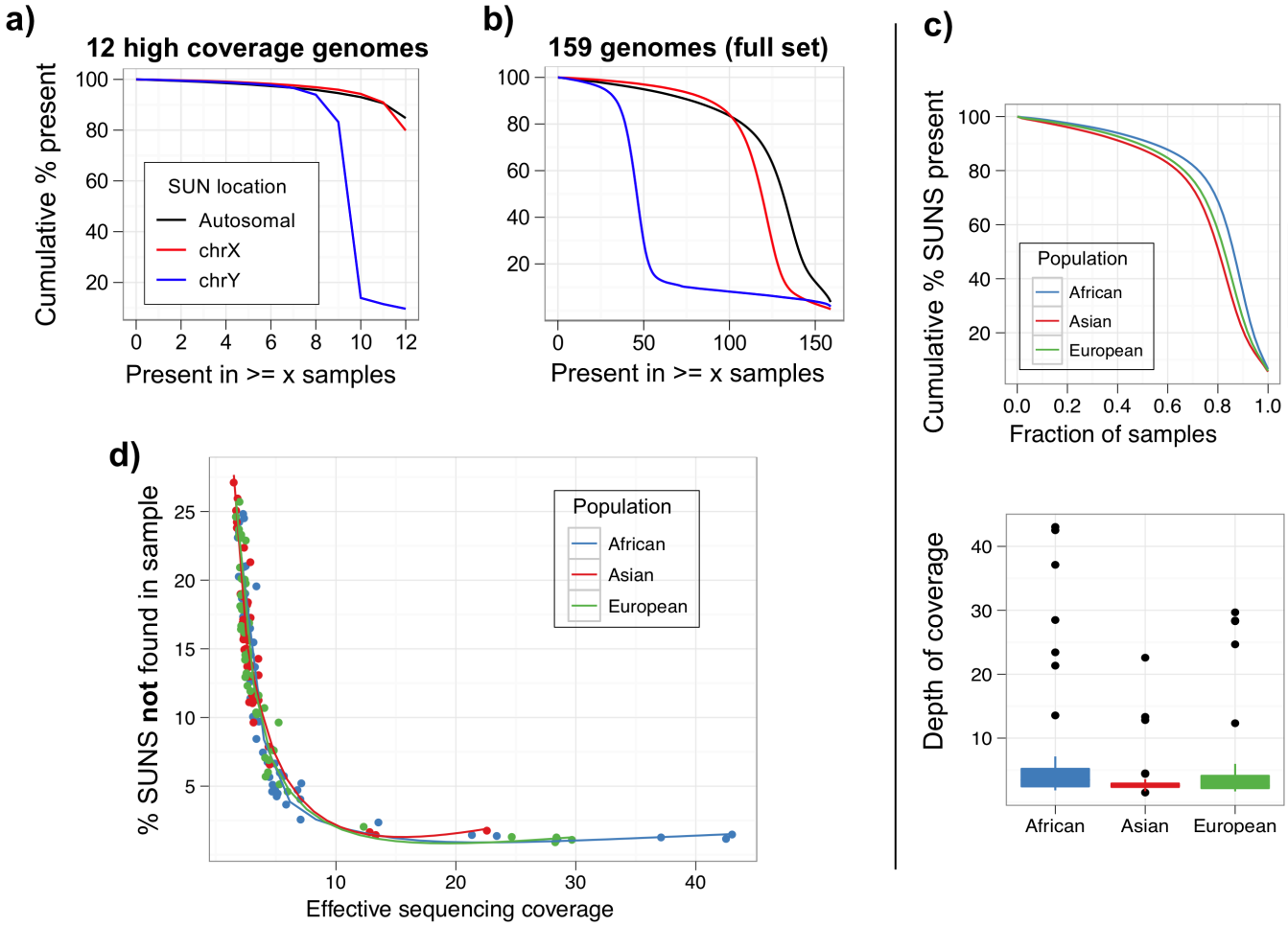


Figure S43. Fraction of SUN positions found within a) genomes of 12 unrelated humans sequenced to high coverage (each >10X, mean = 25.3X) and b) the full set of 159 genomes, including 144 of lower coverage (mean = 3.1X). c). Proportions of SUNs observed among each of the three populations are similar, with a slight excess of markers observed among the African genomes, which were the most deeply sequenced. d). Lower coverage genomes show higher rates of SUN absence to a similar extent among the three populations sampled.

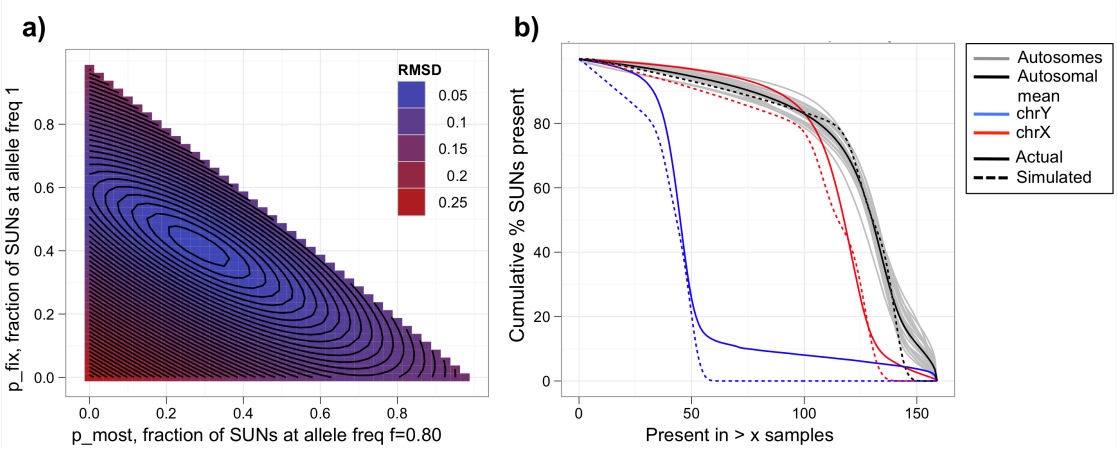


Figure S44. Simulation of SUN presence/absence. a) RMSD between simulated and observed rates of SUN presence, shown for varying p_{fix} and p_{most} with f=0.800. RMSD is minimized at (p_{fix},p_{most},f)=(0.425,0.275,0.800). b). Simulated and observed rates of SUN presence across 159 samples using best-fit parameters.

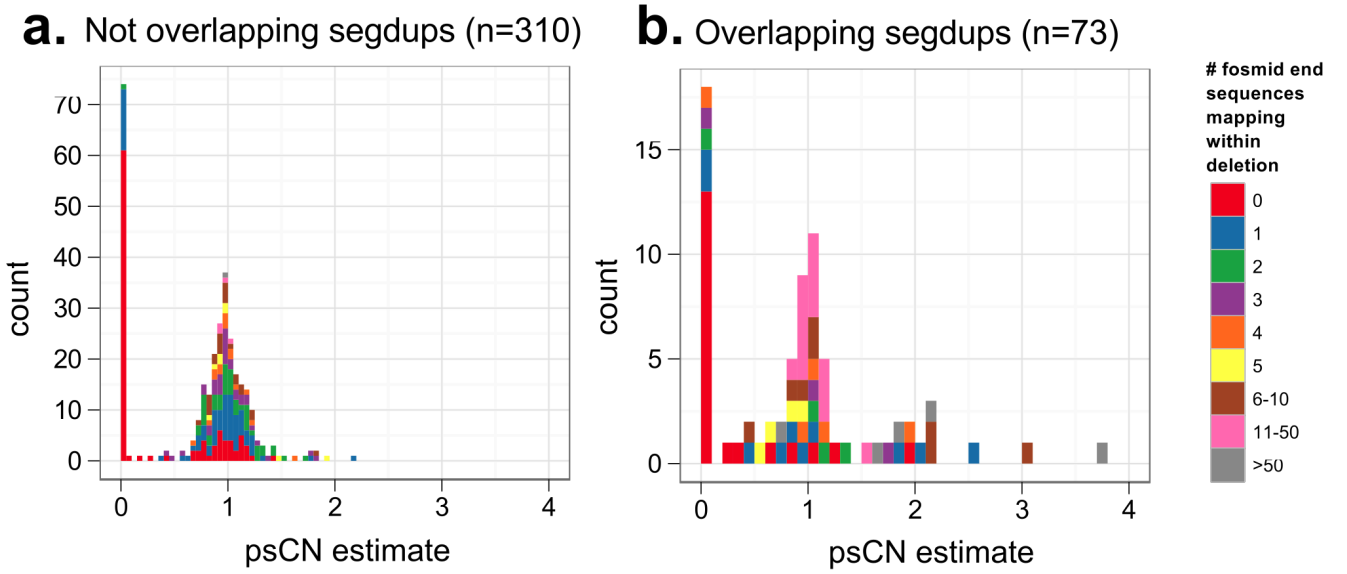


Figure S45. Predicted psCN of validated deletions within the same individual where each deletion was initially mapped. Bars are colored by the number of other fosmid end sequences mapping within each deletion interval (excluding the clone that was sequenced to resolve the deletion). The majority of deletions called as homozygous (psCN = 0) are overlapped by no other clone end sequences (red). a) Of the 310 validated deletions not overlapping segmental duplications, 81 (26.1%) are homozygous (psCN<0.5), 220 (71.0%) are hemizygous (0.5<=psCN<1.5), and 9 (2.9%) are called as diploid or greater (psCN>=1.5). b) Of the 73 deletions overlapping segmental duplications, 22 (30.1%) are homozygous deleted, 37 (50.7%) are hemizygous deleted, and 14 (19.2%) are diploid or greater, the latter possibly being nested within larger blocks of amplification.

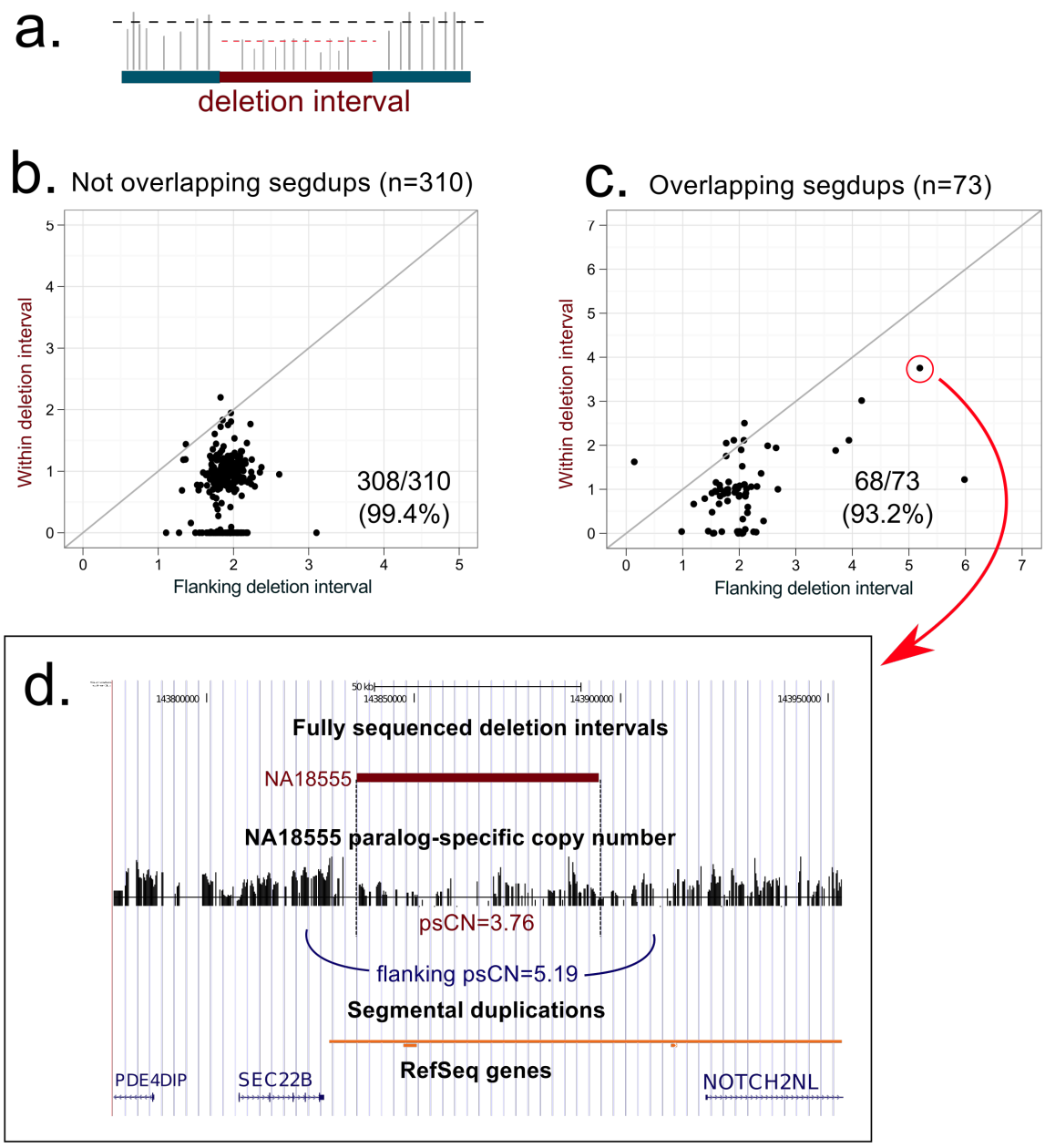


Figure S46. Comparison of psCN estimates within deletion intervals to flanking regions. a) psCN is compared between the deletion interval and for 10 kbp of flanking sequence. If genotyping is accurate, the psCN estimate in the deletion interval should be lower than that of the corresponding flanks. b) For 308/310 deletions that do not overlap segmental duplications, the psCN estimate within the deletion interval was less than that of the flanking regions (below the diagonal). c) For 68/73 sequenced deletions overlapping segmental duplications, the deletion interval shows lower psCN than the flanking regions. d) For some deletions, such as this example on 1q21.1, the psCN estimate is >=2 and depressed relative to flanking regions, indicating that the deletion is within an amplified region.

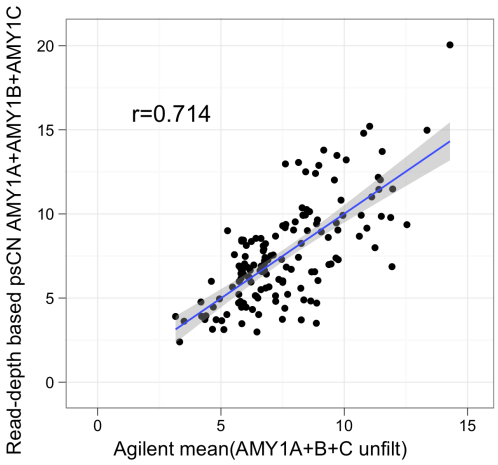


Figure S47. Comparison across 145 individuals of aggregate psCN genotypes of *AMY1* paralogs with estimates obtained from single-channel Agilent microarrays (*S6*).

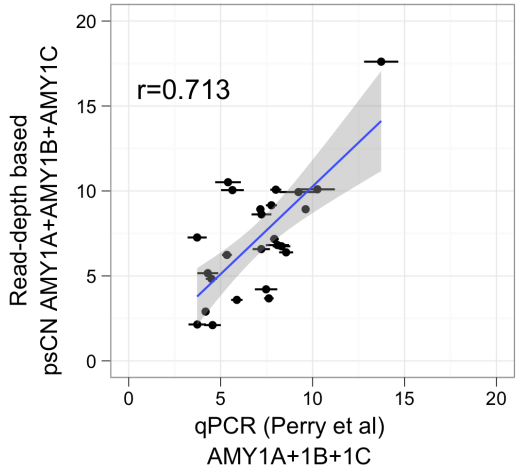


Figure S48. Comparison across 25 Japanese individuals of aggregate psCN genotypes of *AMY1* paralogs with estimates obtained by quantitative PCR directed at the three functional *AMY1* copies (*S34*). Error bars, s.e.m.

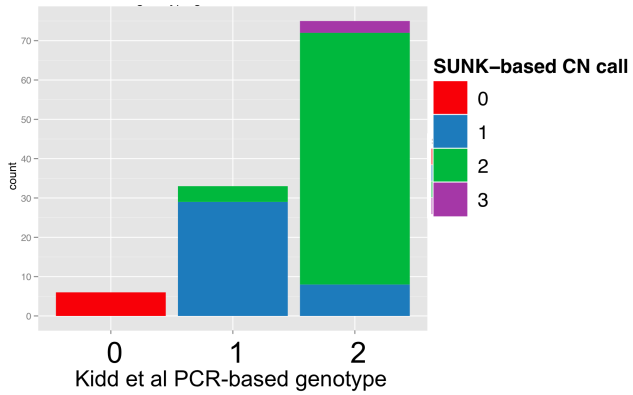


Figure S49. Comparison across 114 individuals of psCN genotypes and breakpoint PCR-based genotypes for *APOBEC3B* deletion.

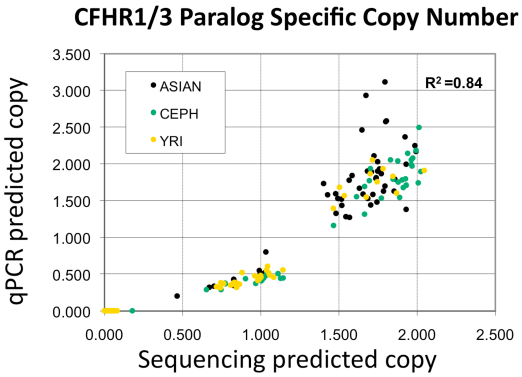


Figure S50. Confirmation of psCN genotyping of *CFHR3* by a paralog-specific qPCR assay.

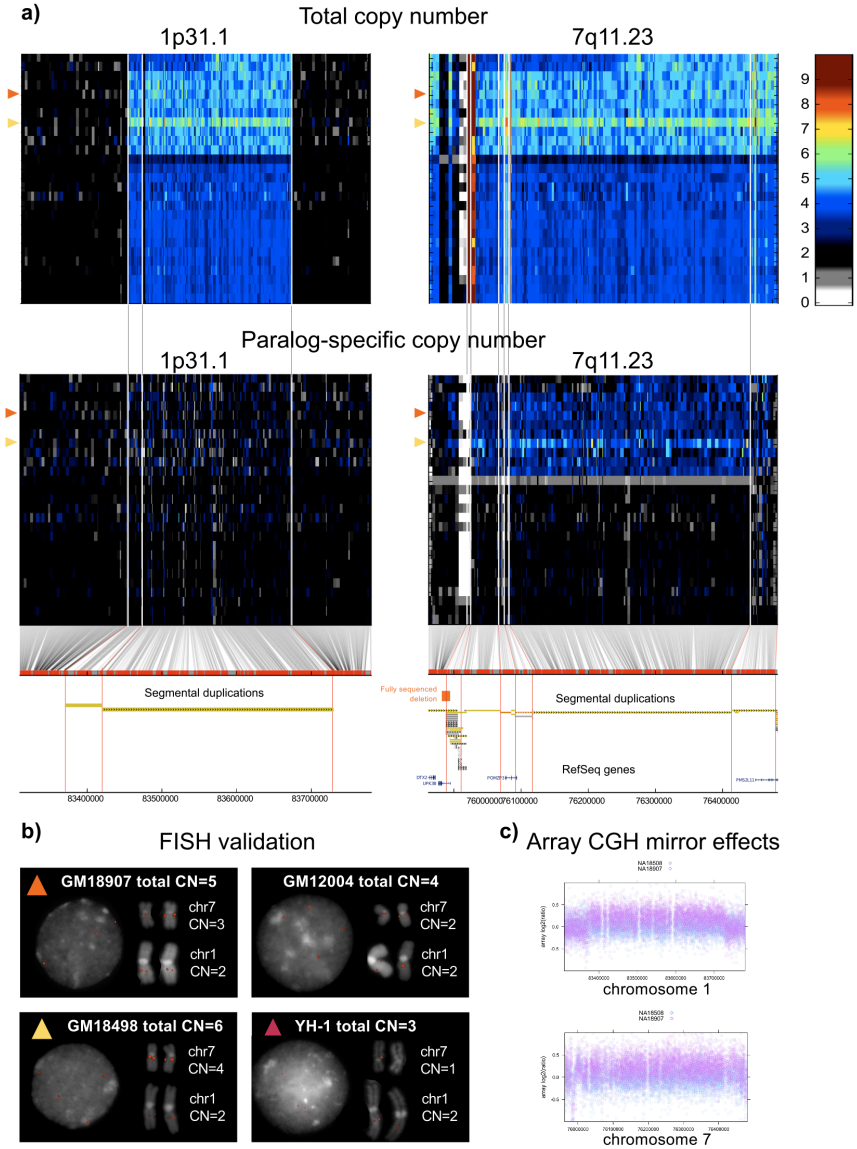


Figure S51. Resolving the true sites of copy number variation with paralog-specific genotyping. a) Apparent CNV within a ~360-kbp segmental duplication on chromosomes 1 and 7 detected by total copy number mapping. Paralog-specific copy

number mapping reveals this variation is exclusive to chromosome 7. b) FISH analysis of interphase nuclei confirms the total copy number change, and analysis of metaphase nuclei confirms the variable copy is on chromosome 7. c) Array CGH (*S5*) shows a similar effect at both loci in one of the copy-number amplified individuals (NA18907, purple relative to the individual without this amplification, NA18508, blue).

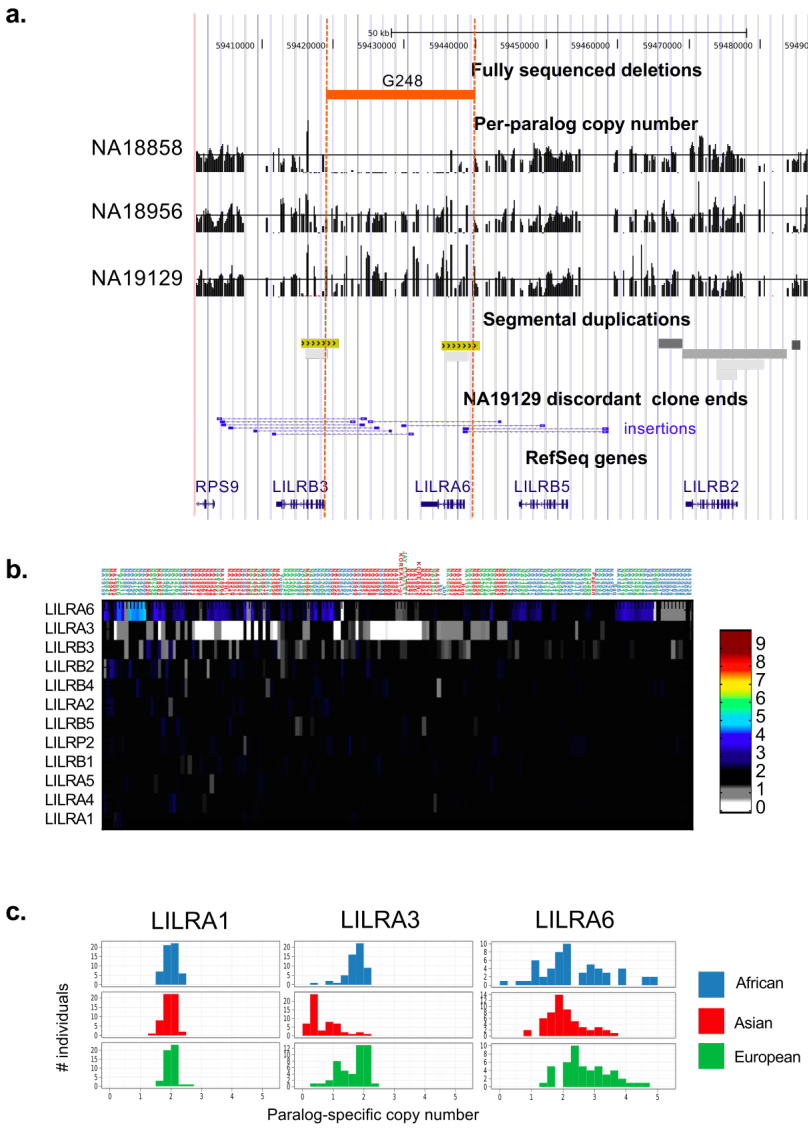


Figure S52. Paralog-specific copy number variation among *LILR* genes
a) Genomic view of paralog-specific copy number at one of the *LILR* gene clusters. Known, fully-sequenced deletions covering *LILRA6* and *LILRA3* are recapitulated, and a reciprocal, two-fold amplification of *LILRA6* supported by fosmid end sequence pair data is found in sample NA19129. b) Gene family heat map of paralog-specific copy number, with genes on the rows and hierarchically clustered individuals on the columns. c) Population comparison of *LILR* family CNV.

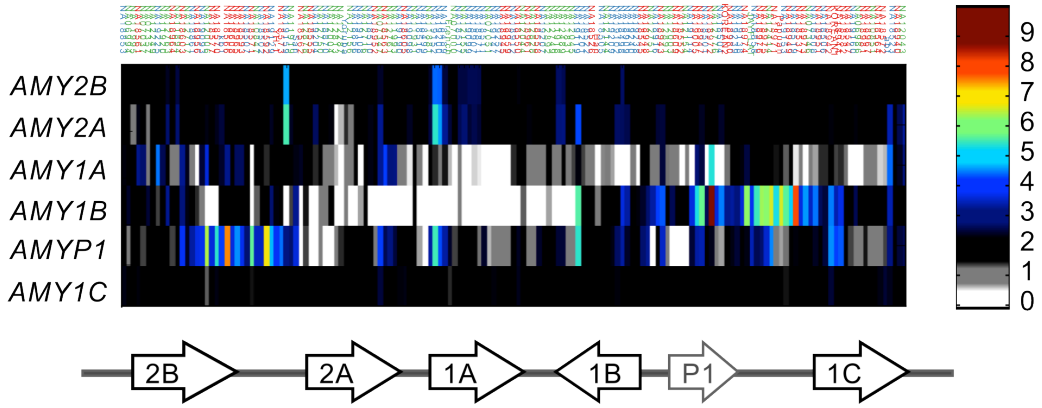


Figure S53. Copy number variation of the both pancreatic and salivary amylase genes shows that the pseudogene *AMY1P* and salivary genes *AMY1A* and *AMY1B* are much more variable in copy when compared to pancreatic amylase genes, *AMY2A* and *AMY2B*, especially among Asians. Expansion of these loci represents a potential adaptation to digestion of starch-rich diets among human populations (*S34*).

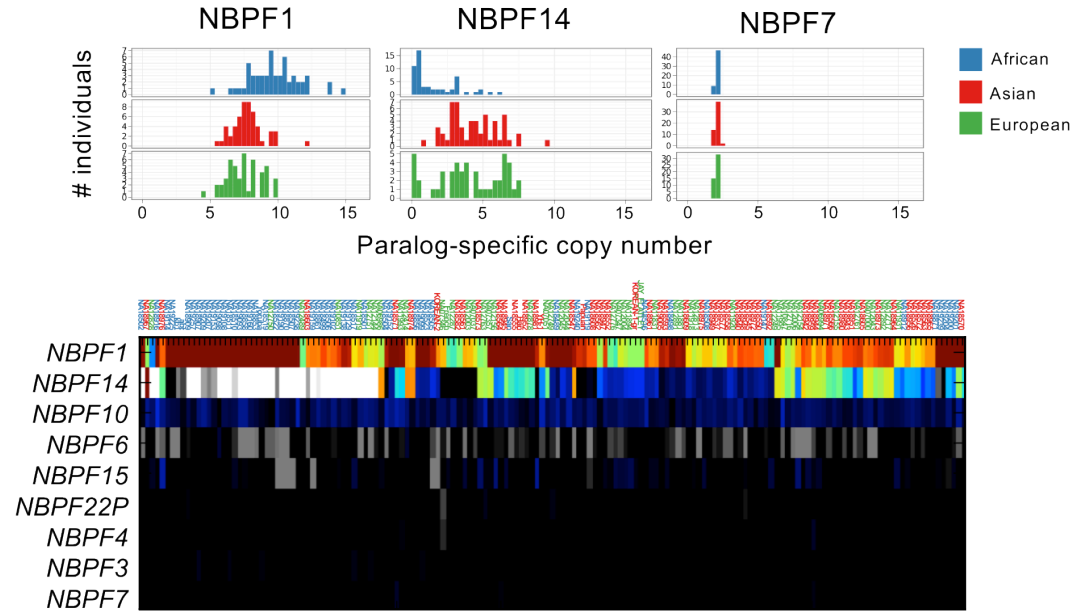


Figure S54. Population stratification and paralog-specific copy variability of a human expanded gene family of unknown function, *NBPF* (neuroblastoma breakpoint gene family). Certain paralogs (e.g., *NBPF1*) are highly amplified, extremely variable, and population stratified while others are nearly fixed and diploid (e.g., *NBPF7*).

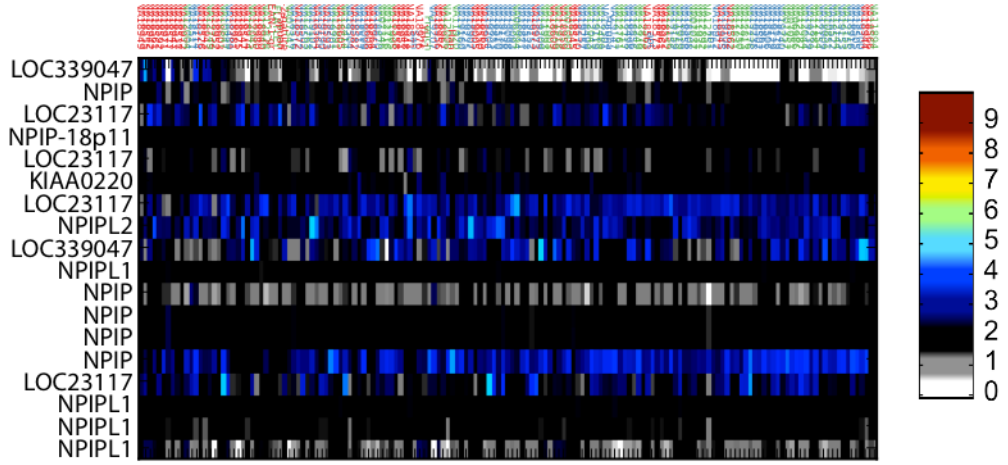


Figure S55. *NPIP* gene family. Each row represents a distinct paralog (e.g., *NPIP*, for which 5 distinct loci on chromosome 16 are shown). Individual genomes are hierarchically clustered (names shaded by population) and shown as columns.

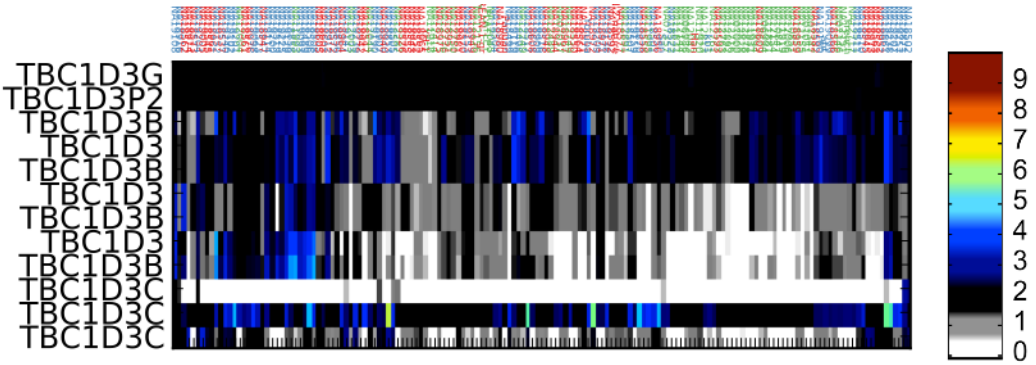


Figure S56. *TBC1D3* gene family.

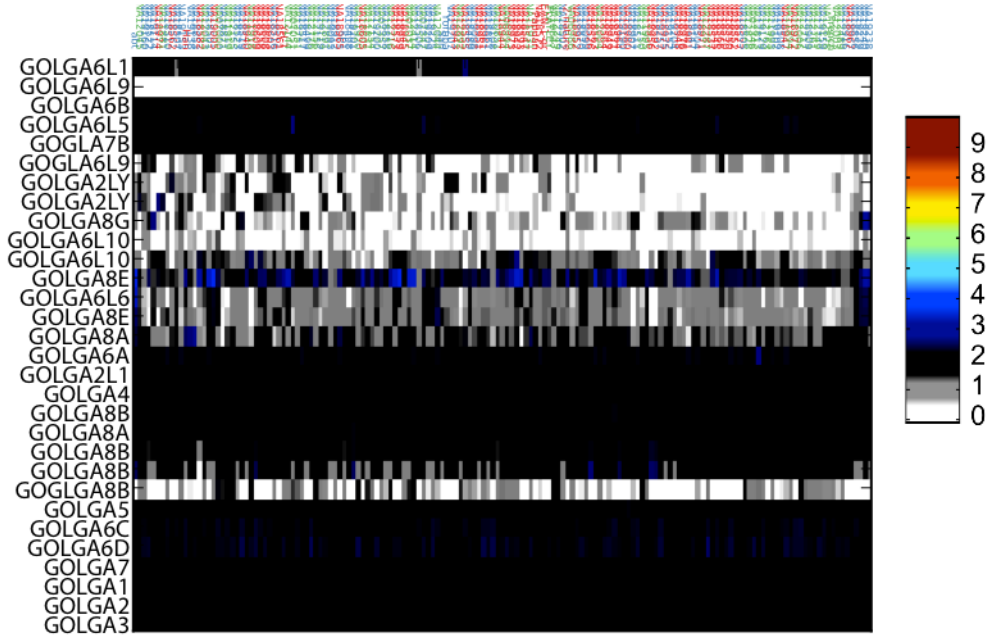


Figure S57. *GOLGA* gene family.

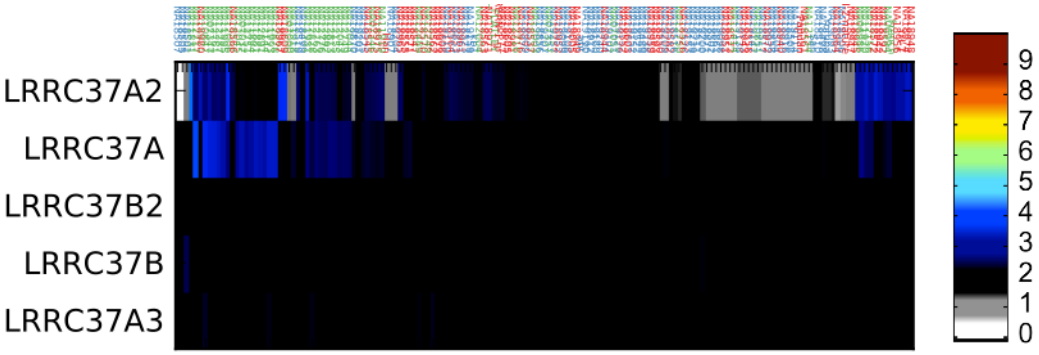


Figure S58. *LRRC37* gene family.

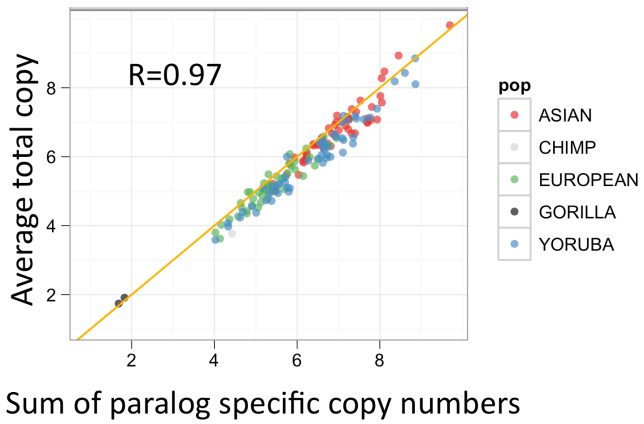


Figure S59. The duplicated gene *ESPN* and the paralogous pseudogene *ESPNP*. Average total copy number estimated is plotted against the sum of the paralog-specific copy numbers of each gene. Orange line denotes y=x.

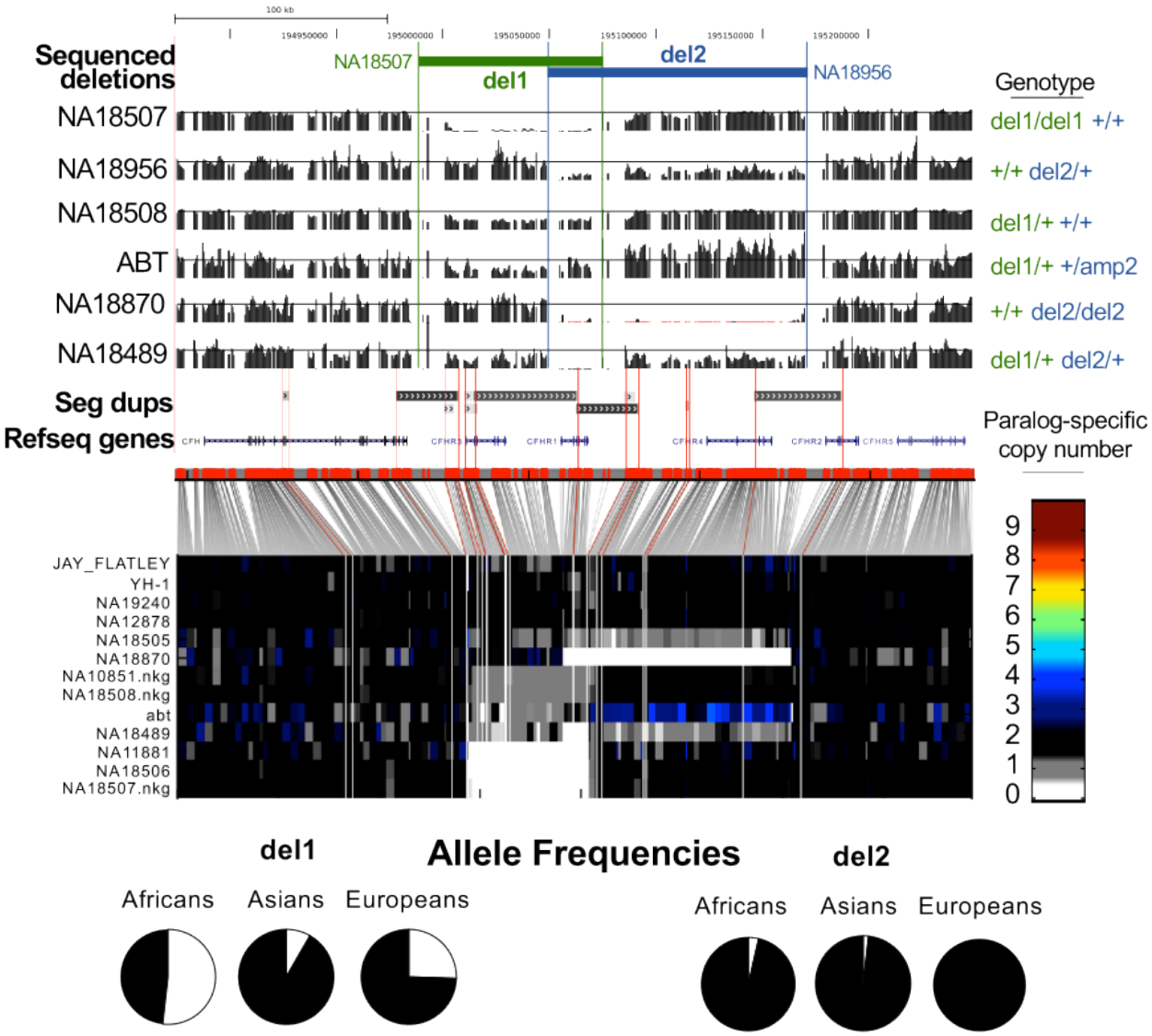


Figure S60. Paralog-specific genotyping of the complement factor H receptor (CFHR) gene family on chromosome 1q31.3. SUN read depth (histograms) across the region delineates the boundaries of two known deletions as defined by clone sequencing. Genotying paralog-specific copy number reveals population stratification in allele frequencies both for del1 (CFHR3/1) and the less common del2 (CFHR1/4). Rarer states, such as the reciprocal duplication of del2 found only in Archbishop Desmond Tutu (ABT) are also detected.

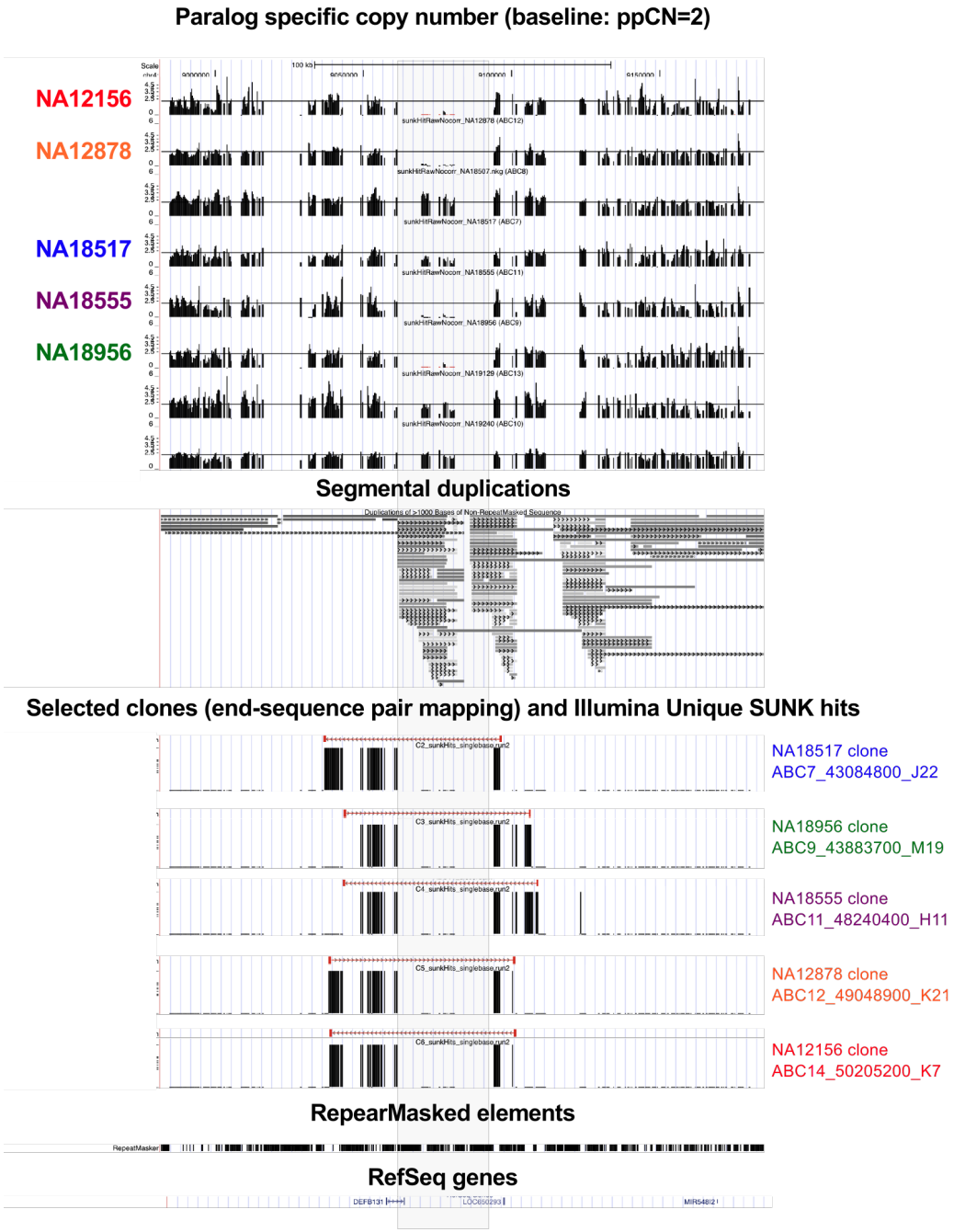


Figure S61. Deletion on chromosome 4p16 confirmed in 5/5 individuals based on fosmid resequencing.

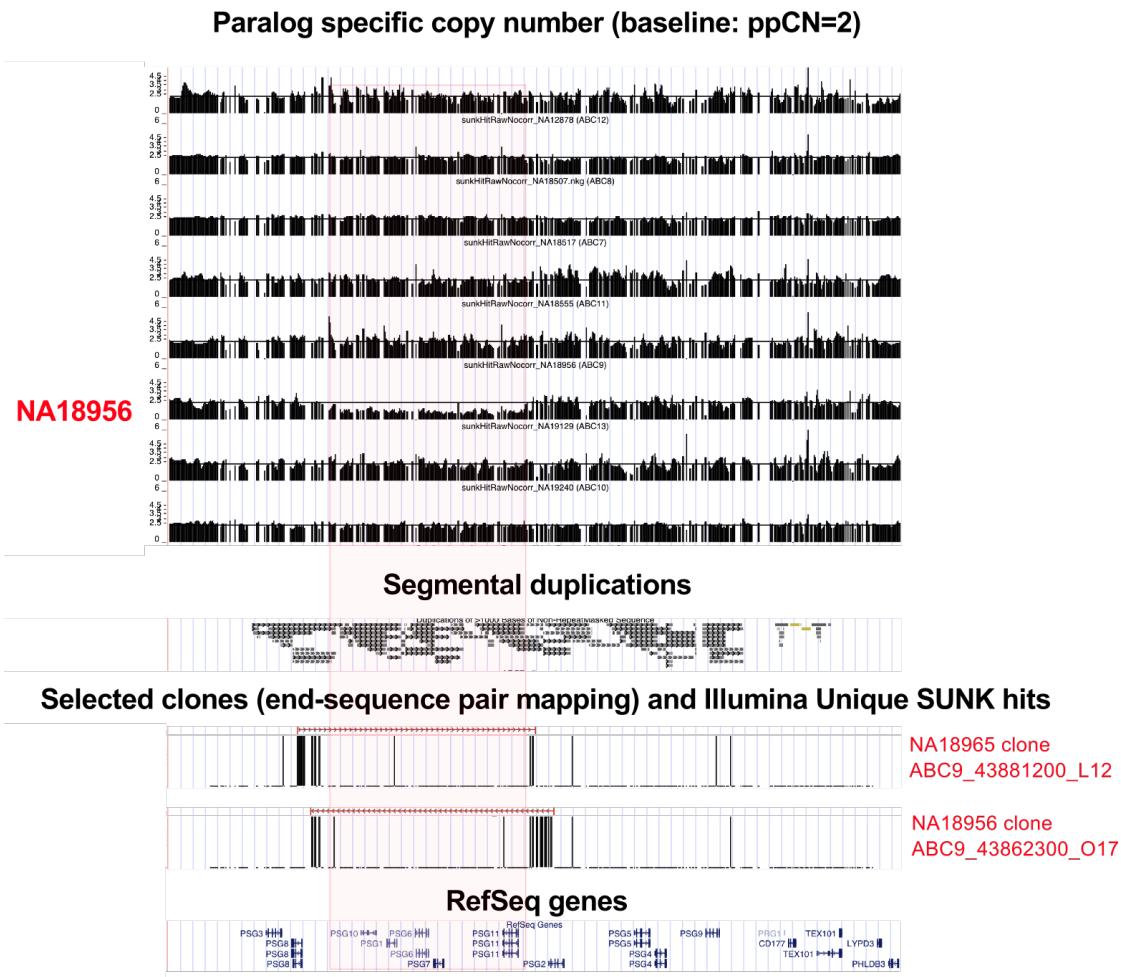


Figure S62. Deletion within the *PSG* gene cluster on chromosome 19q13.31 confirmed in NA18956 (2/2 clones).

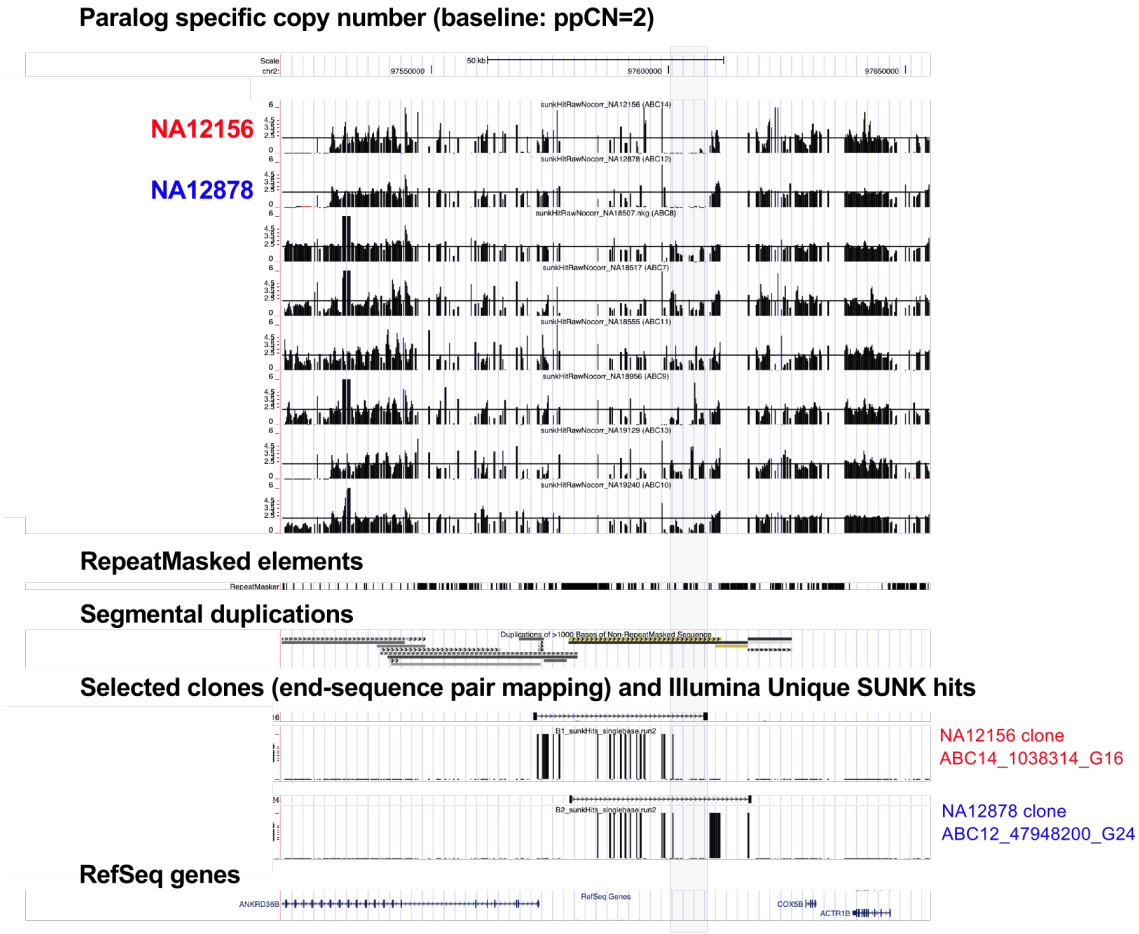


Figure S63. Small (~8-kbp) deletion upstream of *ANKRD36B* confirmed in 2/2 individuals.

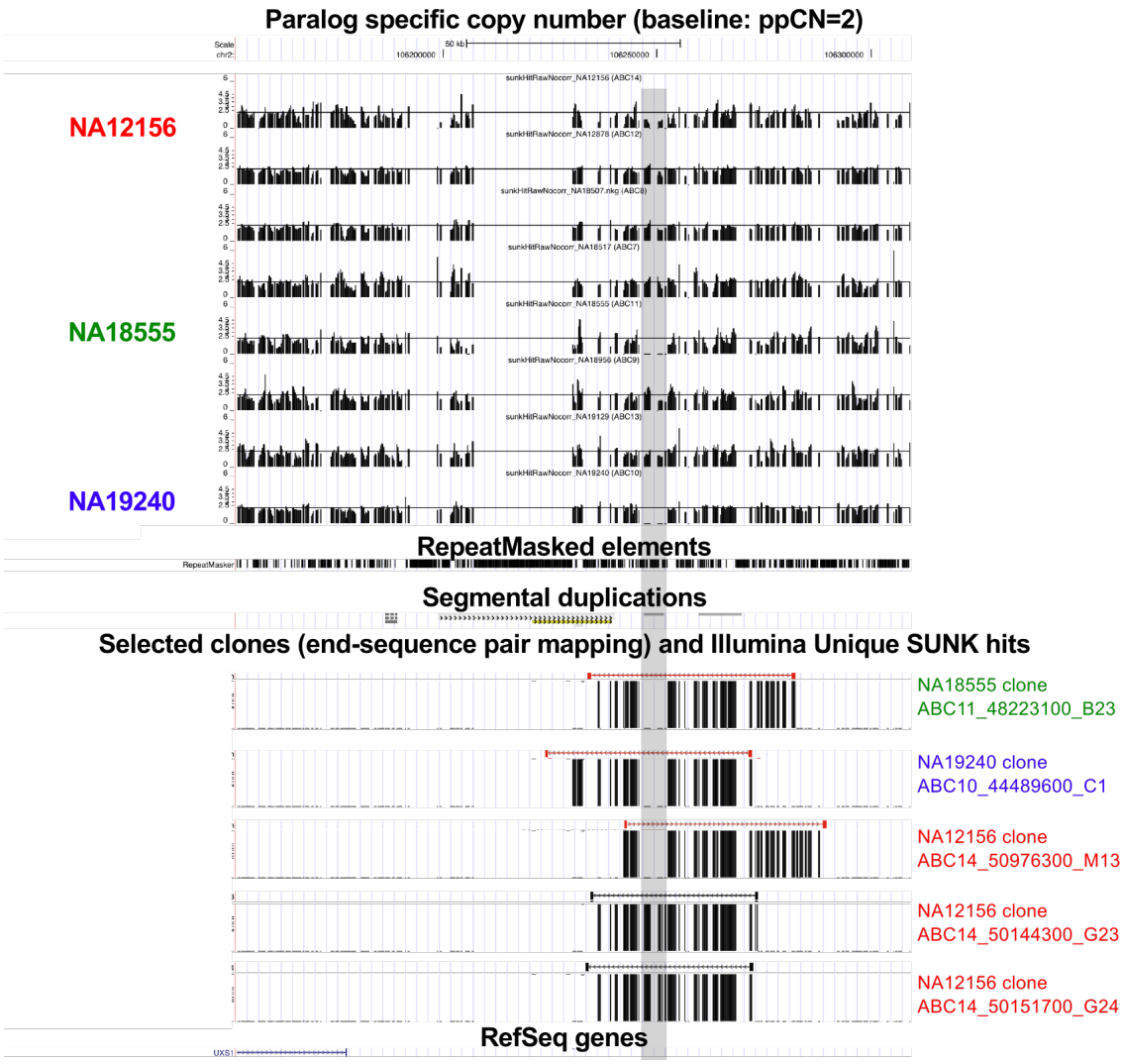


Figure S64. Deletion on chromosome 2q12.2 confirmed in 3/3 samples (deletion allele confirmed in 3/3 clones, nondeletion allele in 2/2).

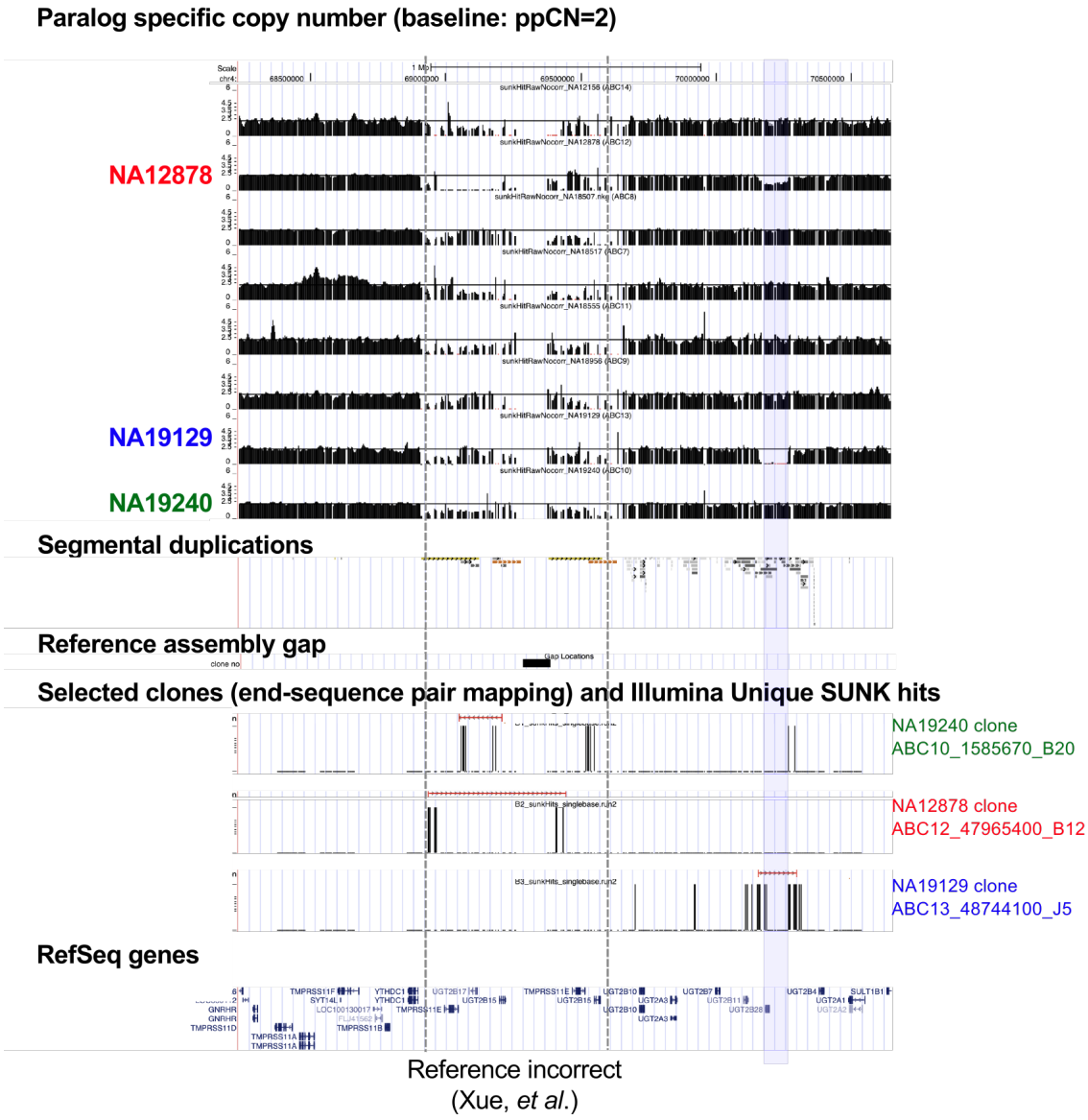


Figure S65. Deletion of *UGT2B28* on chromosome 4q13 confirmed in NA19129. Proximal deletion clones and whole-genome SUNK data reflect apparent deletion generated by spurious segmental duplication in the reference assembly (*S31*).

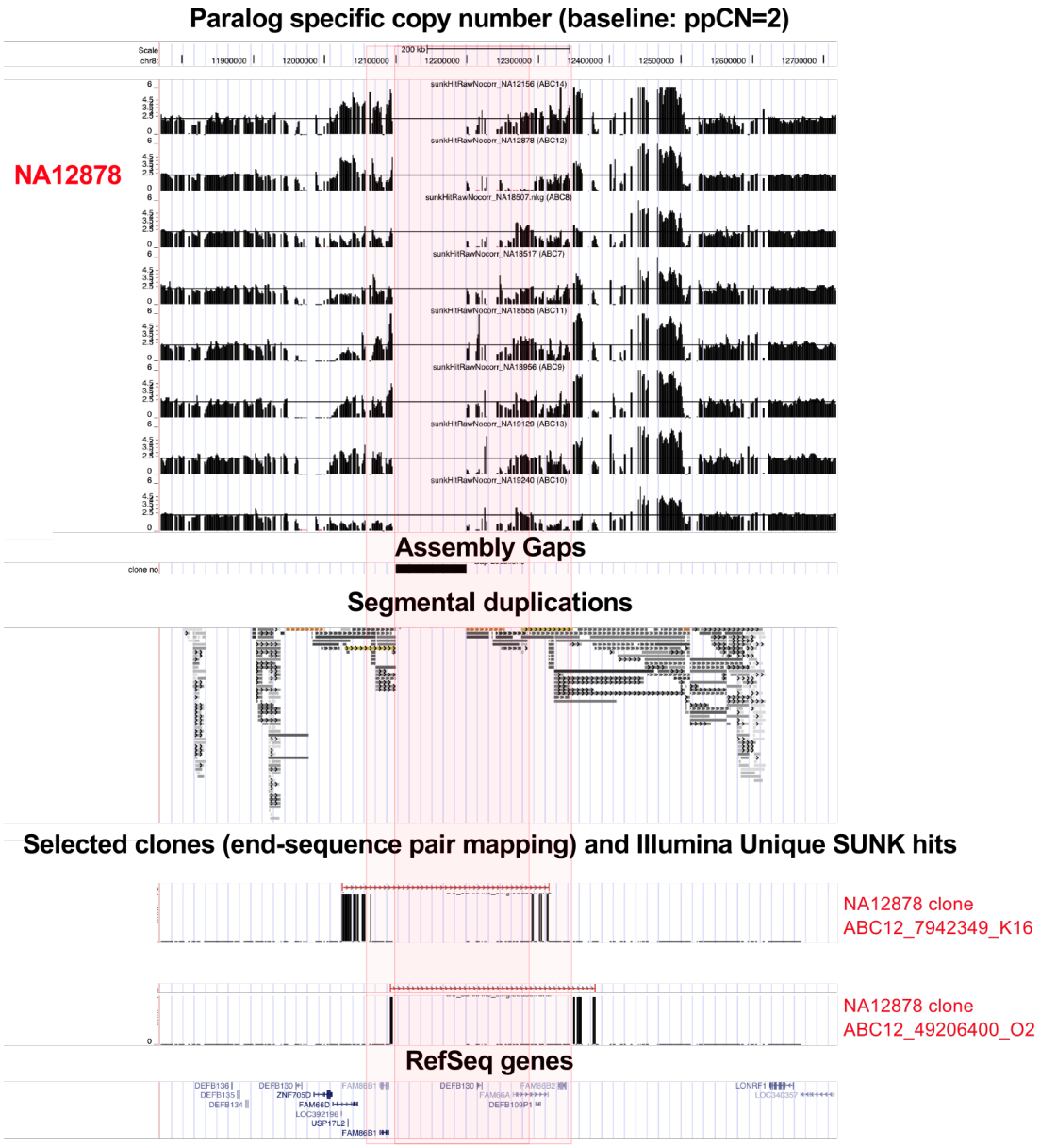


Figure S66. Nested deletions of defensin genes on chromosome 8p23.1 confirmed in NA12878.

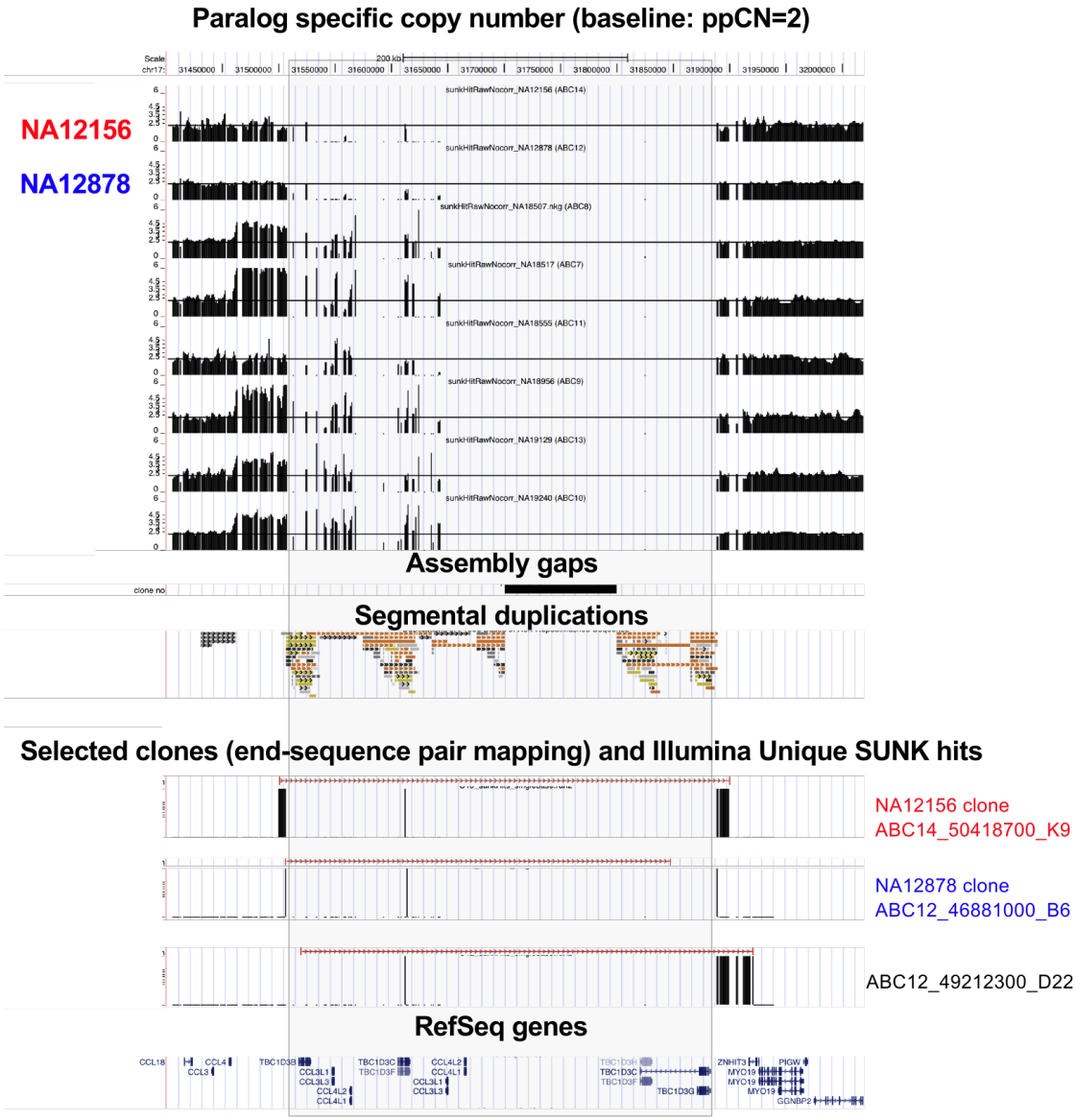


Figure S67. Nested deletions of chemokine ligand genes on chromosome 17q12 confirmed based on fosmid sequencing individuals NA12878 and NA12156.

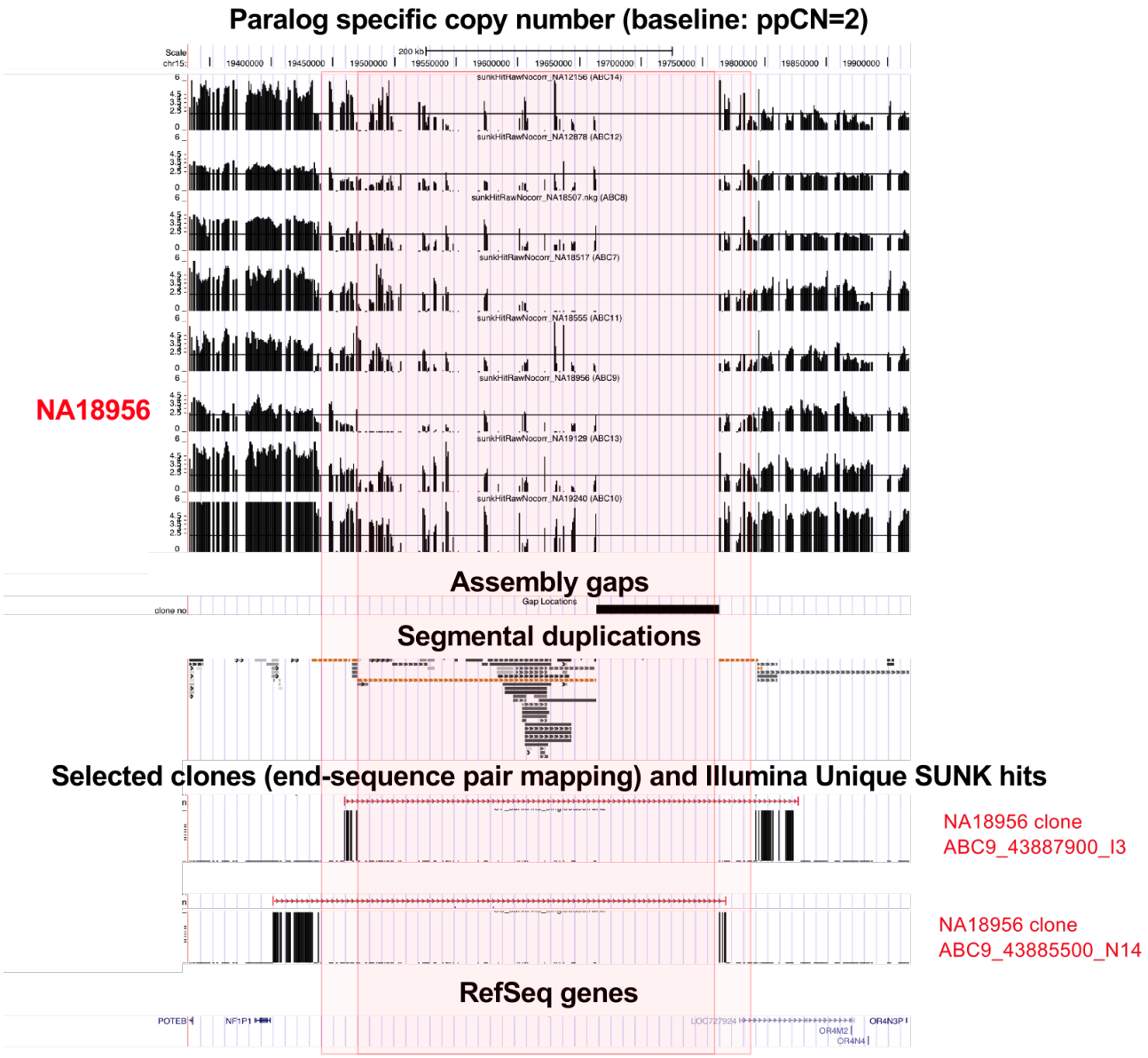


Figure S68. Nested deletions on chromosome 15q11.2 confirmed in NA18956.

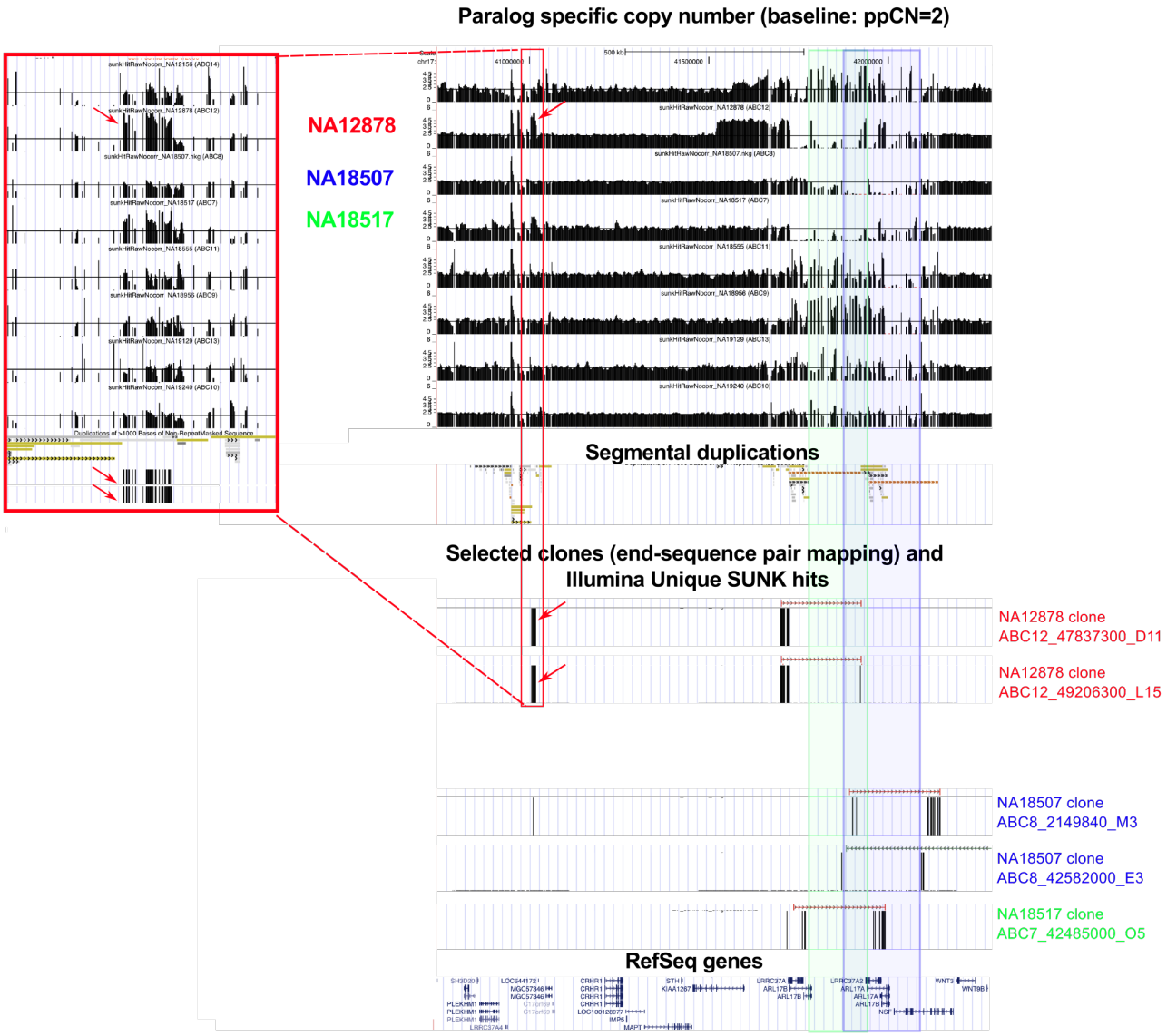


Figure S69. Deletions on 17q21.31 confirmed in 2/3 individuals. In the other individual, NA12878, 2/2 clones suggest this deletion is nested within an amplified, alternative structural configuration including a segmental duplication found in the reference genome ~600 kbp proximal to the end sequence mapping location (inset).

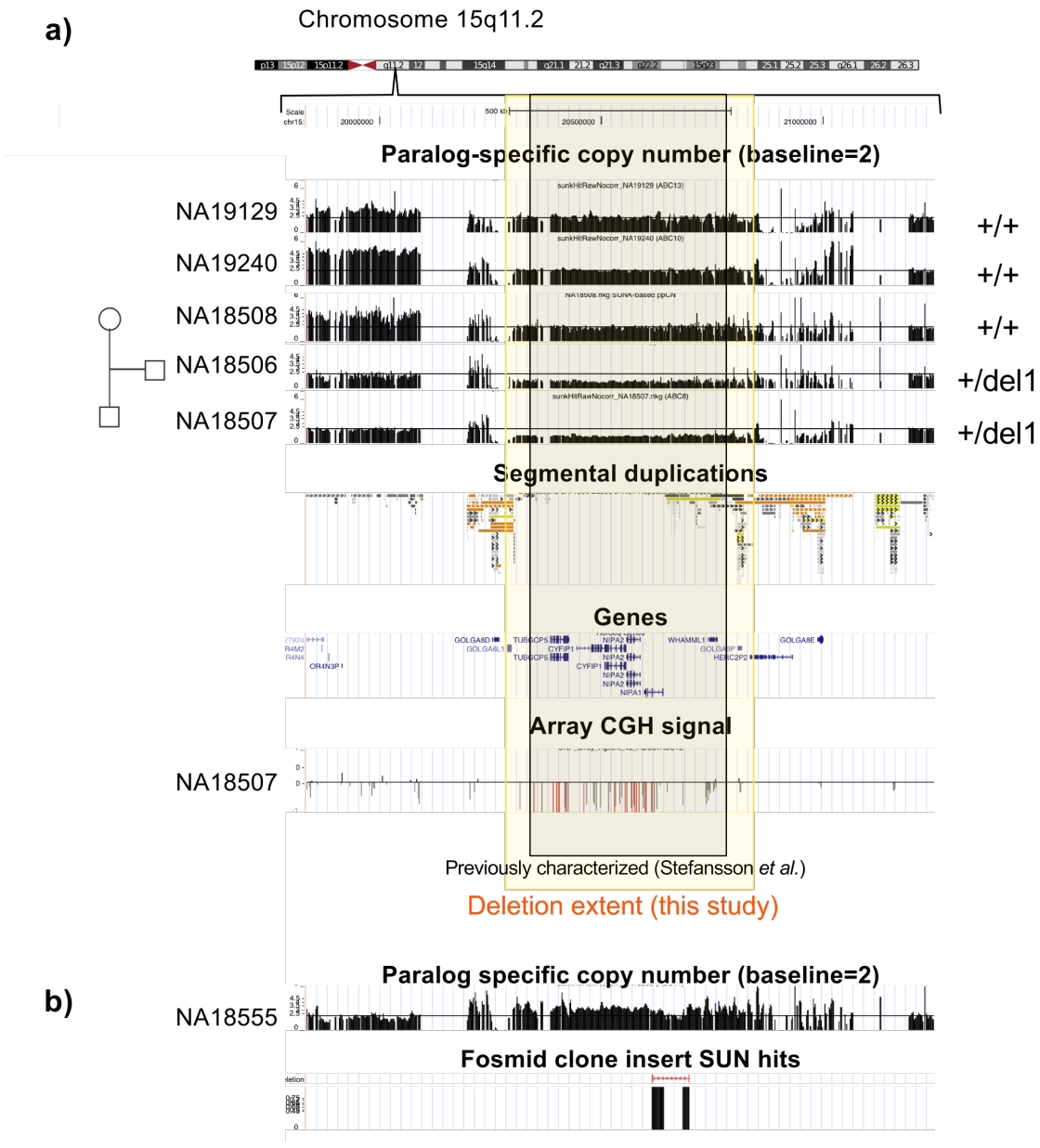


Figure S70. Breakpoint refinement at 15q11.2 schizophrenia-associated deletion

a) Two individuals carry a deletion previously associated with schizophrenia (*S36*). Paralog-specific copy number tracks show the true extent of this deletion into flanking segmental duplications, encompassing additional genes relative to its previously characterized boundaries. This deletion is inherited by NA18506 and is confirmed in the father (NA18507) by array CGH and FISH. b) A novel deletion within the reciprocal single-copy amplification is confirmed by fosmid resequencing in individual NA18555.

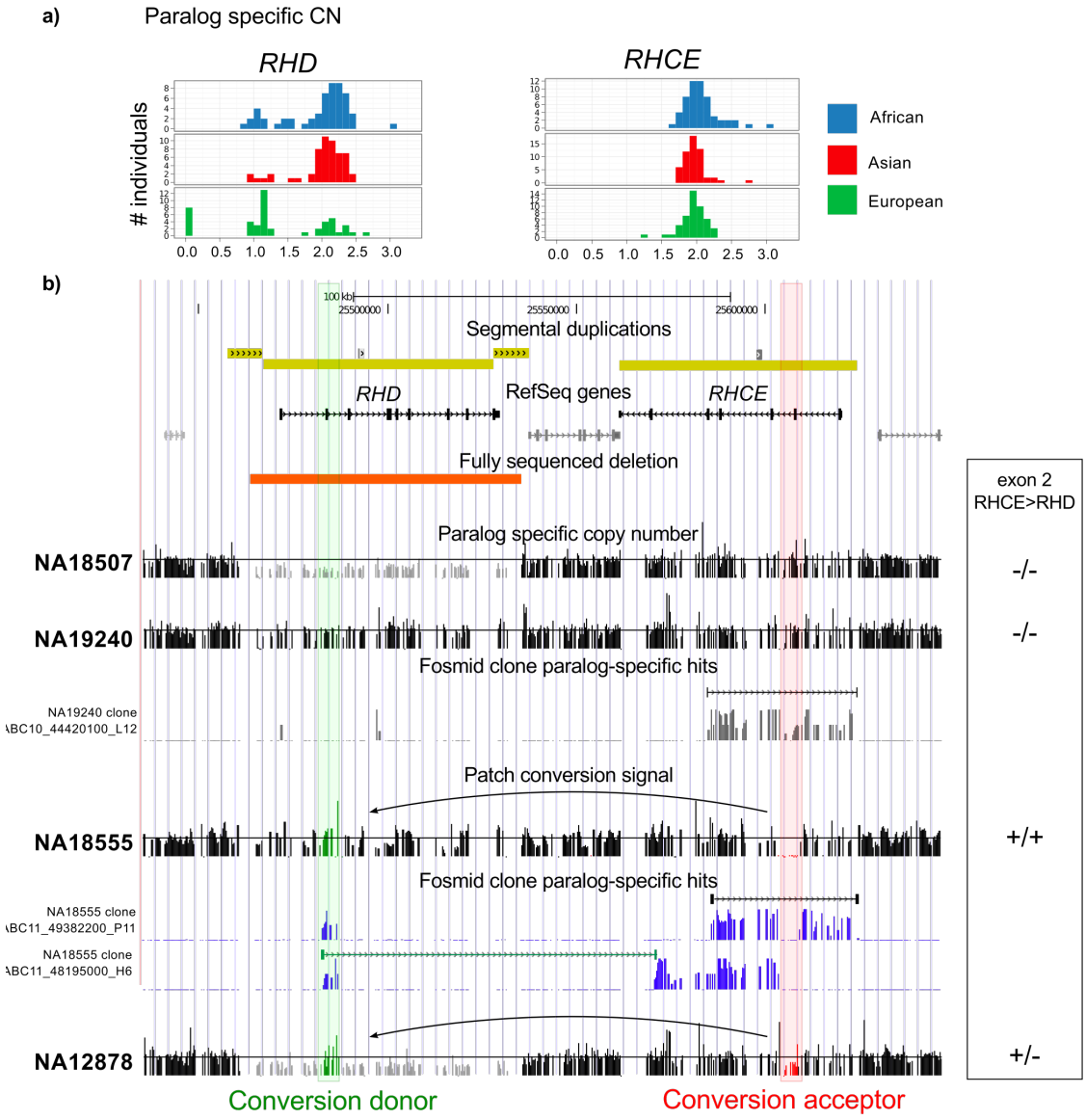


Figure S71. Signatures for gene conversion at the Rh blood group locus

Rhesus (Rh) blood group genes *RHD* and *RHCE* lie within tandem segmental duplications in inverted orientation on chromosome 1. a) *RHD*, but not *RHCE*, is commonly deleted among Europeans. b) Signatures of patch gene conversion detected using whole-genome psCN maps and validated by fosmid clone insert short read sequencing.

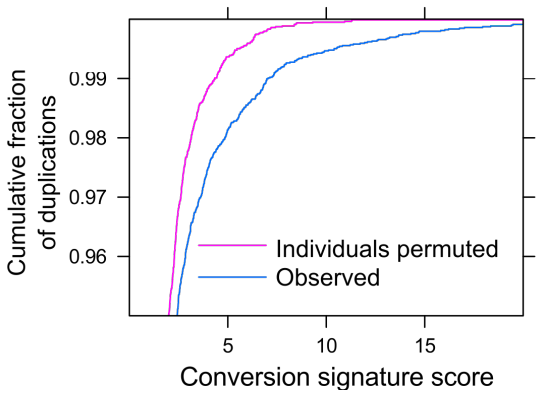


Figure S72. Significant excess of conversion score (P < 2.2x10^{-16}) relative to permuted controls suggests widespread signatures of recent interlocus gene among duplicate alleles segregating within the population.

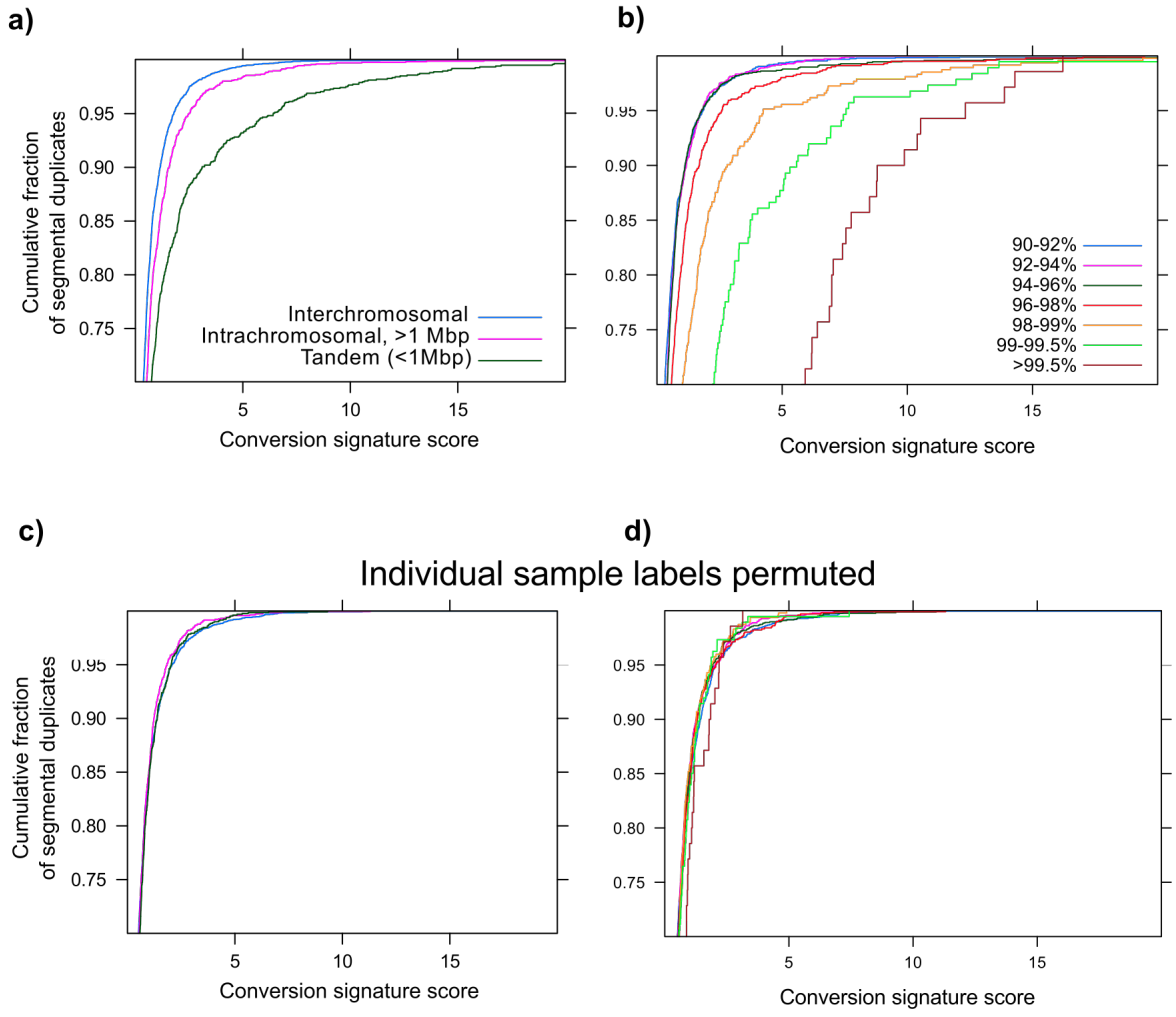


Figure S73. Conversion signature scores for segmental duplications binned by a) arrangement and b) homology show strong preferences for nearby, highly identical duplications. c,d) In permuted controls no significant difference is observed.

SUPPORTING TABLES

Table S1 – Summary of human genomes analyzed

	Population	# of individuals	Effective Coverage	Read Placements	Source
Low Coverage	CEU	43	1.7-6.4X	$5.37X10^9$	1000 Genomes Pilot1
	YRI	46	1.8-7.1X	$7.11X10^9$	1000 Genomes Pilot1
	ASN	50	1.5-3.6X	$5.40X10^9$	1000 Genomes Pilot1
High Coverage					
NA12878 trio	CEU	3	24.7-29.7X	$3.52X10^9$	1000 Genomes Pilot2
NA19239 trio	YRI	3	13.6-28.5X	$2.71X10^9$	1000 Genomes Pilot2
NA18507 trio	YRI	3	37.1-43.0X	$5.26X10^9$	(*S53*)
JF	European	1	12.3X	$5.00X10^8$	Illumina
NA10851	CEU	1	28.4X	$1.23X10^9$	(*S10*)
AK1	Korean	1	22.6X	$9.32X10^8$	(*S11*)
SJK	Korean	1	12.8X	$5.25X10^8$	(*S12*)
YH-1	CHB	1	13.3X	$6.00X10^8$	(*S14*)
South Africans	Bantu, Kalahari Bushman	2	7.0-23.4X	$1.36X10^9$	(*S13*)
HGDP	Han, Papuan, San, Yoruba	4	4.4-7.1X	$9.16X10^8$	(*S54*)
HUMAN TOTAL		**159**	**1145X**	$3.54x10^{10}$	
Chimpanzee		1	6.7X	$2.78x10^8$	(*S55*)
Orangutan		1	8.5X	$3.25x10^8$	(*S56*)
Gorilla		1	4.5X	$2.08x10^8$	GenBank Short Read Archive accession SRP002878
NON-HUMAN TOTAL		**162**	**1165X**	$3.62x10^{10}$	

Data were downloaded from the NCBI Sequence Read Archive (*http://www.ncbi.nlm.nih.gov/sra*) and mapped to a human reference genome (Build36) using the *mrsFAST* aligner.

Table S2 – Singly Unique Nucleotide (SUN) identifier summary

	All SD	Genic	Protein-coding
Total Size	$1.539x10^8$ bp	$4.080x10^7$ bp	$2.125x10^6$ bp
Non-repeat	$5.480x10^7$ bp	$1.726x10^7$ bp	$1.638x10^6$ bp
Indels	1,898,842	619,484	37,476
Transitions	1,283,671	398,268	37,852
Transversions	782,245	236,094	19,996
Multi-SUNs	111,502	28,400	2,205
Total bp	4,076,260	1,282,246	97,529
SUNs/kbp	74.4	74.3	59.5
Ti/Tv	1.64	1.69	1.89

Insertions were counted along their entire lengths; deletions were counted only at junction positions. Multi-SUNs are positions with distinct, uniquely identifying differences to more than one paralog, and, therefore, could be classified into more than one category (e.g., a transition with respect to one duplicate and a transversion with respect to another). Repeats as determined by RepeatMasker and TandemRepeatsFinder, padded +/- 36bp. Genic and protein-coding as defined by RefSeq. SD segmental duplications Ti/Tv, transition to transversion ratio.

Table S3: Summary of FISH validaton

Region	Prediction	FISH result	Cell Line	Clone used as probe	Clone location (Build 36)
4q35.2	5	6 (chr4; chr13; chr21)	GM18564	WIBR2-3705H20_G248P801591D10	chr4:190,776,606-190,812,852
	7	8 (chr4; chr13; chr14; chr21)	YH-1		
	7	8 (chr4; chr13; chr14; chr21)	GM18861		
	6	7 (chr4; chr13; one chr14; chr21)	GM12878		
7q11.23	6	6 (chr1; duplicated on both chr7)	GM18498	WIBR2-2972N17_G248P88301G9	chr7:76,229,616-76,269,483
	5	5 (chr1; chr7 and duplicated on one chr7)	GM18907		
	3	3 (chr1; one chr7)	YH-1		
	4	4 (chr1; chr7)	GM12004		
7q11.32	6	6 (chr1; duplicated on both chr7)	GM18498	WIBR2-2854G20_G248P88510D10	chr7:76,144,718-76,182,710
	4	4 (chr1; chr7)	GM12004		
7q35	8	5	GM07347	WIBR2-3730J21_G248P801433E11	chr7:143,542,919-143,583,088
	2	2	GM18853		
8p23.1	2	2	GM12716	WIBR2-0588K19_G248P80424F10	chr8:7,286,320-7,324,309
	9	9	GM18502		
10q11.22	3	3	GM18563	WIBR2-3223I07_G248P802188E4	chr10:46,393,134-46,431,399
	8	8	GM18940		
15q11.2	3	3	GM18555	WIBR2-0866L18_G248P8587F9	chr15:20,506,689-20,550,176
	1	1	GM18507		
15q11.2	3	3	GM18971	WIBR2-3778G07_G248P802620D4	chr15:19,107,938-19,144,015
	9	9	GM07037		
15q13.3	3	3	GM12813	WIBR2-1228I16_G248P83186E8	chr15:30,068,513-30,106,795
	2	2	GM12004		
	3	3	GM12813	WIBR2-2841A03_G248P88552A2	chr15:30,235,939-30,278,876
	4	4	GM12004		
15q13.3	3	3	GM12813	WIBR2-0682I02_G248P80054E1	chr15:29,881,242-29,924,151
	2	2	GM12004		
16p12.3	3	3	GM18547	WIBR2-2202N24_G248P86159G12	chr16:17,788,765-17,827,587
	2	2	GM12004		
16p12.1	5	5	GM12004	WIBR2-2031K01_G248P86836F1	chr16:22,477,340-22,523,335
	6	6	GM12878		
	6	6	GM18861		
	4	4	GM18555		
	5	5	GM18947		
	5	5	GM18956		
16p12.1	3	3	GM12004	WIBR2-3608M06_G248P801202G3	chr16:22,542,172-22,579,566
	4	4	GM12878		
	4	4	GM18861		
	2	2	GM18555		
	3	3	GM18947		
	3	3	GM18956		
16p12.1	6	7	GM12004	WIBR2-0590C03_G248P80492B2	chr16:21,423,413-21,466,802
	8	8	GM12878		
	7	8	GM18861		
	6	6	GM18555		
	7	7	GM18947		
	6	7	GM18956		
16q22.2	5	5 (duplicated on one chr1)	GM19201	WIBR2-3823N03_G248P802141G2	chr16:69,580,922-69,624,156
	3	3 (missing copy on one chr1)	GM19190		
17p11.2	4	4	GM12287	WIBR2-1008E03_G248P84612C2	chr17:20,284,001-20,322,710
	8	8	GM18907		
17q21.31	2	2	GM19240	WIBR2-2004M23_G248P85784G12	chr17:41,545,746-41,586,094
	4	4	GM12878		
17q21.32	6	7	GM18942	WIBR2-1857K04_G248P85426F2	chr17:42,055,759-42,093,331
	6	6	YH-1		
	4	4	GM19129		
	4	4	GM12892		
17q21.32	3	5	GM18555	WIBR2-0946N09_G248P801829G5	chr17:42,986,667-43,025,934
	4	6	GM12813		
	10	12	GM18502		
5q31.3*	3	2	GM07347	WIBR2-3662F08_G248P801703C4	chr5:140,151,286-140,188,584
	6	2	GM18545		
	4	2	GM07347	WIBR2-2544K17_G248P86728F9	chr5:140,526,836-140,569,682
	8	2	GM18545		
7q11.23**	19	18-20	GM18552	WIBR2-0458K07_G248P8624F4	chr7:73,942,214-73,981,865
	19	16-20	GM12004		
	30	22-26	GM18861		
16p13.11**	33	29-34	GM18537	WIBR2-1564J16_G248P83059E8	chr16:14,921,937-14,963,735
	49	37-48	GM18501		
17q12**	18	14-18	GM18573	WIBR2-1684A05_G248P81904A3	chr17:33,536,637-33,572,498
	48	19-26	GM18861		

*prediction not validated

**FISH failed to accurately estimate copy-number

Table S4. Quantitative PCR validation targets and assay selection

TARGET	Primer-set #1 r^2	Primer-set #2 r^2	on X chromosome	Primer-set 1 product size	Primer-set 2 product size
OPN1MW2	0.76	NA	No	100	NA
NPEPPS	0.92	NA	no	120	NA
TBC1D3	0.95	0.9	no	103	103
CCL3L1	0.93	0.90	no	119	117
NBPF16	0.93	0.66	no	108	112
GAGE10 (males and females)	0.98	0.91	yes	130	204
GAGE10 (males)	0.37	0.63	yes	130	204
GAGE10 (females)	0.97	0.6	yes	130	204
PRSS1	0.14	0.59	no	201	115
NPIP	0.66	0.92	no	231	374
CFHR3 (paralog-specific assay)	0.84	NA	no	159	NA

Table S5. Quantitative PCR primer sets

Primer Target	Primer Name	Primer Sequence
Albumin (diploid control)	alb_HSA_3F	TTGTGGGCTGTAATCATCG
	alb_HSA_3R	TGCTGGTTCTCTTTCACTGAC
CCL3L1	CCL3L1_2F	GGGTCCAGAAATACGTCAGT
	CCL3L1_2R	CATGTTCCCAAGGCTCAG
	CCL3L1_4F	TGGGGTCTGTTCTTCACTCT
	CCL3L1_4R	CGGTTCAAGAAGTCATACCC
NBPF16	NBPF16_1F	AAACGTCAGCATGGTGGTAT
	NBPF16_1R	TGTTTCTTCTCTGCCAACTG
	NBPF16_3F	TGCAGGACTCACTGGATAGA
	NBPF16_3R	CCAAGGTACTGTTCCTCCAA
NPEPPS	NPEPPS_2F	CCGCACACCTGTTATGTCTA
	NPEPPS_2R	CAGGAGTGTAAACACGGACA
	NPEPPS_3F	AGAGCATCCACCAGTACCTC
	NPEPPS_3R	TTCGGTAGCTCCACCTTATC
OPN1MW2	OPN1MW2_1F	ACCATGAAGTTCAAGAAGCTG
	OPN1MW2_1R	CTGGTTCACAACGCTGATAG
	OPN1MW2_4F	TGATGCTTCAGTGCACTCTT
	OPN1MW2_4R	ATGGGTACCGGAGTTCATTA
TBC1D3	TBC1D3_1F	TAGCACCTGTCTGCTCTCTG
	TBC1D3_1R	TCAGGGAGAAAACCTTTGAG

	TBC1D3_2F	CATAGATCGAGCGTACAAGG
	TBC1D3_2R	GTATCTTCCGGGGTTTTTC
GAGE10	GAGE10_5F	GAGTTGGAGAAATGTCTTTAGGC
	GAGE10_5R	AATACGTGTGGGTGTGCAAAT
	GAGE10_6F	ATTGTCACATAGGAGGAAGAGG
	GAGE10_6R	TTTTATCATAGTGAGGGATTTGC
NPIP	NPIP_5F	GAAGCTGTCTGAACTTACTCAGG
	NPIP_5R	TGAATAGCGTGGGATTTCTC
	NPIP_7F	TTTTCAGATCACCCCTTCTG
	NPIP_7R	AGCCGTAGGGAGAAAAATGT
CFHR3	CFHR3.2F*	TGATTCTGGACGTTTTGCTG
	CFHR3.2R*	AAGGGAACGAAAGGCTTCTG
* CFHR3.2 primers were designed to overlap SUNs specific to the *CFHR1/3* deletion		

Table S6: Missing Paralogous Genes in Reference Genome

Gene families were idenditified which are missing from the human reference genome if the reference genome had 5 or more fewer copies than the median number of copies among 159 individuals. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

gene name*	location	median copy number (159 individuals)	BUILD36 reference genome copy	Vst
WASH3P	chr15:100318538-100334331	21.23	10.75	0.03
NBPF10	chr1:144004727-144080039	262.88	135.86	0.06
CROCCL1	chr1:16817339-16829988	10.23	3.72	0.25
MSTP2	chr1:16844655-16849501	13.85	5.67	0.06
PDE4DIP	chr1:143663117-143787436	7.33	1.98	0.48
NOTCH2NL	chr1:143920467-143997269	9.06	3.92	0.38
LOC645166	chr1:147194909-147218219	12.5	5.4	0.02
DUX4	chr10:135330357-135335265	367.72	31.34	0.22
LOC284232	chr13:18306542-18344109	26.81	16.72	0.07
PRR20A	chr13:56619622-56622644	16.42	10.13	0.17
NPIPL3	chr16:22432384-22455362	50.29	44.18	0.17
FLJ36000	chr17:21828188-21837197	58.28	11.76	0.08
TBC1D3G	chr17:31871287-31882216	37.3	22.75	0.5
LRRC37A4	chr17:40939892-40948305	16.22	11.13	0.24
DHX40P	chr17:55408166-55451118	16.85	10.22	0.46
TCEB3C	chr18:42808570-42810447	28.38	6.74	0.22
CGB2	chr19:54226941-54228307	18.62	12.96	0
LOC654342	chr2:91188435-91211702	12.52	5.4	0.02
ANKRD20A3	chr9:67516580-67560113	28.15	19.31	0.09
LOC151009	chr2:110567223-110576650	28.33	18.16	0.09
MGC13005	chr2:114073074-114075623	24.27	14.22	0.01
C2orf27B	chr2:132269003-132275704	19.98	12.44	0.09
FAM182A	chr20:25983249-26015552	15.48	8.47	0.01
FRG1B	chr20:28225539-28247668	31.33	7.28	0.05
BAGE	chr21:10079666-10120808	16.89	3.59	0.04
C21orf81	chr21:14237966-14274636	20.21	14.52	0.14
POM121L4P	chr22:19373842-19376009	50.71	37.14	0.3
LOC96610	chr22:20982462-21007324	18.33	12.42	0.02
ZNF717	chr3:75868718-75916945	25.37	7.18	0.29

ZNF595	chr4:43226-78099	10.63	3.01	0.18
OTOP1	chr4:4241430-4279522	8.73	3.25	0.02
USP17L6P	chr4:8978697-8979892	142.64	37.04	0.27
DRD5	chr4:9392355-9394731	11.34	4.84	0.04
LOC100133050	chr5:99743108-99751857	38.23	27.39	0.04
FLJ35390	chr7:44045591-44048606	13.16	7.75	0.21
SPDYE5	chr7:74962234-74971564	37.77	32.6	0.25
MUC12	chr7:100399623-100448949	12.21	5.82	0.04
RPL23AP53	chr8:148346-172318	27.08	20.99	0.09
FAM90A14	chr8:7124701-7127711	59.59	47.29	0.17
FAM66D	chr8:12010699-12046107	31.43	14.06	0.23
REXO1L2P	chr8:86884283-86885361	171.62	15.68	0.16
C9orf122	chr9:38611084-38613275	13.11	5.96	0.07
LOC442421	chr9:66234088-66242849	15.42	6.1	0.16
CCDC29	chr9:68715487-68738681	32.2	22.03	0.03

Table S7: Most Variable Copy Number Human Genes

The most variable gene familes among all individuals were identified by selecting for those with a variance >3 among all 159 individuals. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

gene name*	location	variance	median copy	Vst	segmentally duplicated base-pairs
DUX4	chr10:135330357-135335265	3858	367.72	0.22	4908
USP17L6P	chr4:8978697-8979892	2031.71	142.64	0.27	1195
REXO1L2P	chr8:86884283-86885361	1055.8	171.62	0.16	1078
NBPF10	chr1:144004727-144080039	629.65	262.88	0.06	75312
FAM90A14	chr8:7124701-7127711	488.38	59.59	0.17	3010
FLJ36000	chr17:21828188-21837197	334.11	58.28	0.08	2218
LOC392196	chr8:12022775-12024213	331.86	58.27	0.27	1438
TBC1D3F	chr17:33591517-33602455	93.2	38.13	0.5	10938
TCEB3C	chr18:42808570-42810447	86.57	28.38	0.22	1877
LOC100272216	chr5:68962735-68964784	82.32	63.64	0.13	2049
FAM66D	chr8:12010699-12046107	79.24	31.43	0.23	35408
PRR20A	chr13:56619622-56622644	45.67	16.42	0.17	3022
LOC100133050	chr5:99743108-99751857	44.5	38.23	0.04	8749
POM121L4P	chr22:19373842-19376009	33.18	50.71	0.3	2167
DHX40P	chr17:55408166-55451118	18.4	16.85	0.46	42952
LOC399744	chr10:38757079-38781086	16.09	46.03	0.09	24007
SPDYE2	chr7:101983351-101989853	14.05	37.2	0.32	6502
C2orf78	chr2:73864823-73897782	14.03	10.88	0.39	14288
FLJ45340	chr7:128068530-128088288	12.1	47.2	0.01	18612
GUSBL2	chr6:58354117-58395683	9.95	24.47	0.13	41566
EEF1AL7	chr4:106625311-106626956	9.69	26.85	0.01	1645
LOC151009	chr2:110567223-110576650	9.53	28.33	0.09	9427
PMS2L2	chr7:72114615-72152902	9.49	28.15	0.12	38287
GGTLC2	chr22:21318781-21320368	9.36	15.62	0.14	1587
POLR2J2	chr7:102064709-102099418	9.34	21.26	0.14	34709
LRRC37A4	chr17:40939892-40948305	9.3	16.22	0.24	8413
LOC440896	chr9:68464033-68470861	8.99	22.6	0.03	6828
AMY1A	chr1:104031562-104040435	8.31	9.49	0.13	8873

FAM157B	chr9:140226457-140253993	8.04	26.46	0.03	27536
RASA4P	chr7:44035010-44046747	7.94	12.59	0.26	11737
NPIPL3	chr16:22432384-22455362	7.57	50.29	0.17	22978
C9orf122	chr9:38611084-38613275	7.08	13.11	0.07	2191
LOC613037	chr16:30141850-30164433	6.61	50.59	0.29	22583
DEFA1	chr8:6822580-6825012	6.55	7.58	0.03	2432
MSTP2	chr1:16844655-16849501	6.45	13.85	0.06	4846
UPLP	chr7:102064709-102070474	6.38	9.5	0.27	5765
GOLGA9P	chr15:20806682-20814184	6.15	34.92	0.11	7502
RPL23AP82	chr22:49569022-49584930	5.67	31.94	0.09	15908
MGC13005	chr2:114073074-114075623	5.51	24.27	0.01	2549
WASH1	chr9:4510-19739	5.09	22.42	0.05	15229
MUC12	chr7:100399623-100448949	5.02	12.21	0.04	7765
MBD3L3	chr19:7007216-7009645	5.01	10.03	0	2429
ZNF717	chr3:75868718-75916945	4.94	25.37	0.29	36176
CCL3L1	chr17:31546381-31548269	4.87	4.95	0.44	1888
DGCR9	chr22:17385346-17387760	4.81	17.06	0.1	0
TP53TG3	chr16:32592349-32595554	4.78	8.55	0.21	3205
POTEH	chr22:14636331-14667937	4.36	19.26	0.07	31606
CCDC29	chr9:68715487-68738681	4.39	32.2	0.03	23194
FRG1B	chr20:28225539-28247668	4.39	31.33	0.05	22129
PRAMEF15	chr1:13514559-13521574	4.39	17.83	0.01	7015
LOC728323	chr2:242679516-242751142	4	19.18	0.09	40432
OR7E91P	chr2:71104712-71110568	3.82	34.64	0.12	5856
ANKRD20A1	chr9:67516580-67559660	3.51	28.11	0.08	43080
NPEPPS	chr17:42963442-43055641	3.5	5.97	0.5	62993
FKSG73	chr2:91492885-91494221	3.35	14.23	0.03	1336
CGB2	chr19:54226941-54228307	3.32	18.62	0	1366

Table S8: Population Stratified Human Gene Families

The most population stratified genes among 159 individuals in three representative populations (European, Yoruba and Asian) were determined using the Vst metric. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

gene name*	location	Vst	median copy	variance	segmentally duplicated basepairs	mean European copy	mean Asian copy	mean Yoruba copy
TBC1D3C	chr17:31820231-31882204	0.53	34.64	83.69	61973	29.28	34.17	43.86
TRY6	chr7:142158331-142161973	0.52	4.37	1.18	3642	4.39	3.51	5.17
NPEPPS	chr17:42963442-43055641	0.5	5.97	3.5	62993	5.5	5.42	8.27
PDE4DIP	chr1:143663117-143787436	0.48	7.33	0.47	34281	7.08	7.49	6.51
UGT2B17	chr4:69085497-69116840	0.47	2.87	0.69	31343	3.03	2.25	3.41
ESPNP	chr1:16890299-16919239	0.46	6.6	1.95	28940	5.62	7.47	6.6
DHX40P	chr17:55408166-55451118	0.46	16.85	18.4	42952	14.81	15.79	21.02
LOC644172	chr17:41033278-41035531	0.45	5.43	0.76	2253	6.56	5.36	5.26
LGALS9B	chr17:20293767-20311440	0.44	5.95	1.11	17673	5.07	6.04	6.68
CCL3L1	chr17:31546381-31548269	0.44	4.95	4.87	1888	3.44	5.93	6.44
PRSS3	chr9:33785558-33789229	0.42	5.55	0.51	3671	5.58	5.02	5.97
LILRA3	chr19:59491666-59496050	0.63	7.08	1.61	3358	7.62	5.62	7.62
DHFRL1	chr3:95259455-95264350	0.41	4.21	0.32	3434	4.55	4.37	3.79

LOC100286793	chr1:142510035-142536110	0.4	11.95	0.64	26075	11.79	12.31	11.25
CCL4	chr17:31455332-31457127	0.4	4.32	1.63	1795	3.49	4.92	5.13
CFHR1	chr1:195055483-195067942	0.39	2.99	0.48	7459	3.09	3.32	2.43
C2orf78	chr2:73864823-73897782	0.39	10.88	14.03	14288	11.07	13.52	8.24
NOTCH2NL	chr1:143920467-143997269	0.38	9.06	0.53	76802	8.96	9.38	8.5
HLA-DPA1	chr6:33140771-33149356	0.37	1.92	0.05	0	2.06	1.91	1.79
RDM1	chr17:31269199-31281416	0.37	4.21	0.1	1741	4.43	4.26	4.02
OCLN	chr5:70405174-70424776	0.37	3	0.68	19602	3.06	2.53	3.56
KRT14	chr17:36992058-36996673	0.36	5.26	0.57	4615	5.74	5.18	4.75
PGA5	chr11:60765244-60775491	0.36	6.61	2.26	0	5.62	7.52	6.18
PDXDC1	chr16:14976333-15039053	0.36	4.87	0.53	55935	5.31	4.38	5.23
PCDHB13	chr5:140573692-140577177	0.33	4.01	1.13	2510	4.94	4.42	3.77
MGC57346	chr17:41053494-41071110	0.33	3.64	0.48	7454	4.4	3.52	3.49
SPDYE6	chr7:101772912-101783609	0.32	36.15	12.86	10697	34.67	35.78	39.05
ACOT2	chr14:73105524-73112112	0.31	2.81	0.64	6588	2.92	2.38	3.31
C17orf58	chr17:63417678-63420227	0.31	3.16	0.3	2549	3.76	3.11	3.07
POM121L4P	chr22:19373842-19376009	0.3	50.71	33.18	2167	54.28	47.69	51.89
PRH1	chr12:10924826-11215477	0.3	2.76	0.03	29610	2.75	2.9	2.7
KIAA1267	chr17:41463128-41605371	0.29	2.04	0.17	18638	2.55	2.02	2.02
ZNF717	chr3:75868718-75916945	0.29	25.37	4.94	36176	25.78	24.08	26.64
ADAM5P	chr8:39291338-39379532	0.29	1.8	0.13	0	1.4	1.68	1.79
GTF2H2	chr5:68891809-68924334	0.28	4.49	0.84	32525	4.76	3.88	4.86
USP17	chr4:8969206-8970799	0.28	139.97	1990.24	1593	120.79	137.72	175.41
FAM157A	chr3:199363633-199392125	0.28	10.49	0.84	28492	10.55	10.17	11.16
LRRC37A3	chr17:60280949-60345365	0.28	11.15	1.36	64416	11.48	11.56	10.4
UBE2QP2	chr15:80820827-80881396	0.27	6.92	0.55	60569	6.45	7.2	7.01
LOC392196	chr8:12022775-12024213	0.27	58.27	331.86	1438	51.8	57.94	73.51
RHD	chr1:25471567-25529523	0.27	3.86	0.39	57956	3.15	3.85	3.81
UPLP	chr7:102064709-102070474	0.27	9.5	6.38	5765	8.29	9.53	11.08
NPIPL3	chr16:21320950-21344159	0.26	50.36	6.58	23209	49.78	49.33	52.09
RASA4P	chr7:44035010-44046747	0.26	12.59	7.94	11737	11.22	12.33	14.37
POLR2J4	chr7:43947018-44025273	0.26	9.75	5.46	57818	8.62	9.53	11.18
CROCCL1	chr1:16817339-16829988	0.25	10.23	2.57	12649	9.64	10.23	11.34
ARHGEF5	chr7:143683421-143708658	0.25	5.94	2.44	21888	6.09	6.71	5.15
ERC2	chr3:55517375-56477431	0.25	1.92	0	0	1.94	1.9	1.92
PLA2G10	chr16:14673905-14696027	0.24	8.87	2.09	7521	8.15	9.17	9.64
RFPL4A	chr19:60962318-60966351	0.24	4.88	1.75	4033	4.86	5.93	4.59
FAM103A1	chr15:81445998-81450427	0.23	3.66	0.26	0	3.87	3.81	3.36
CES1	chr16:54394264-54424576	0.23	3.88	0.1	5672	3.86	4.07	3.76
HEATR4	chr14:73014944-73095404	0.23	2.15	0.02	3215	2.2	2.08	2.22
FAM66D	chr8:12010699-12046107	0.23	31.43	79.24	35408	28.55	30.6	38.22
GUSBL1	chr6:26947244-27032312	0.23	13.2	2.62	85068	14.25	13.27	12.53
TCEB3C	chr18:42808570-42810447	0.22	28.38	86.57	1877	29.01	24.38	33.91
DUX4	chr10:135330357-135335265	0.22	367.72	3858	4908	368.19	341.29	401.37
TP53TG3	chr16:33112480-33115680	0.22	8.22	4.31	3200	8.34	7.32	9.35
ANKRD36	chr2:97142959-97279633	0.21	17.69	2.2	136674	17.22	17.6	18.69
NBPF3	chr1:21639217-21683980	0.21	13.95	1.49	44763	13.52	14.13	14.74
GSTT1	chr22:22706138-22714284	0.21	1.08	0.76	0	1.61	0.79	0.86
ZNF705D	chr8:11984255-12010434	0.2	10.12	1.31	26179	10.15	9.87	10.91

SMARCA2	chr9:2005341-2183623	0.2	1.93	0	0	1.96	1.92	1.93
KRTAP9-9	chr17:36642277-36666142	0.2	4.32	0.32	17293	4.38	4.57	4

Table S9: Most Copy-number Stratified Genomic Regions
The most population stratified genomic regions were detected by computing the Vst statistic in 1kb unmasked sequence windows across the entire genome and merging contigous windows of increased signal (Vst >0.2).

chr	start	end	size	average Vst	segmentally duplicated base-pairs	genes
chr1	16855269	16994786	139517	0.23697321	137937	ESPNP MSTP9
chr1	143012474	143113897	101423	0.223041759	100627	LOC728855 LOC728855 LOC728875 PPIAL4B PPIAL4C PPIAL4A
chr1	143660533	143789398	128865	0.306949902	36865	PDE4DIP PDE4DIP PDE4DIP PDE4DIP
chr1	143899256	144016942	117686	0.21152537	117294	NOTCH2NL NBPF10
chr1	147777596	147882589	104993	0.239820262	104993	PPIAL4A PPIAL4C LOC728855
chr1	194978578	195093170	114592	0.333558968	109991	CFH CFHR3 CFHR1
chr12	11061265	11184215	122950	0.210171532	66557	PRR4 PRH1 TAS2R19 TAS2R31 TAS2R46 TAS2R43 TAS2R30
chr16	14938245	15039130	100885	0.221789233	83352	NPIP PDXDC1
chr17	31427937	31700688	272751	0.426615843	229816	CCL3 CCL4 TBC1D3B CCL3L3 CCL3L1 CCL4L1 CCL4L2 TBC1D3C TBC1D3F CCL3L3 CCL3L1 CCL4L1 CCL4L2
chr17	33287994	33430126	142132	0.389299833	139484	LOC284100
chr17	33529126	33664621	135495	0.481327103	135390	TBC1D3 TBC1D3F TBC1D3F
chr17	40927933	41058228	130295	0.264685464	129472	LRRC37A4 LOC644172 MGC57346 MGC57346
chr17	41521057	42155415	634358	0.22378867	510129	KIAA1267 LRRC37A ARL17 ARL17 LRRC37A2 ARL17P1 ARL17P1 ARL17 NSF
chr19	15254	135375	120121	0.205305338	120121	FAM138A FAM138C FAM138F OR4F17
chr22	17035971	17269339	233368	0.223062515	233368	USP18 GGT3P
chr22	19794051	20040977	246926	0.224252184	245294	POM121L8P
chr3	163996489	164111761	115272	0.418407193	0	
chr4	8932857	9033353	100496	0.208320107	98308	USP17 USP17 USP17 USP17 USP17 USP17 USP17 USP17 USP17L6P
chr4	69049455	69237384	187929	0.429892723	135770	UGT2B17 UGT2B15
chr7	101901561	102120157	218596	0.224432098	218596	POLR2J POLR2J3 SPDYE2 RASA4 RASA4 UPLP POLR2J2 CTAGE4 ARHGEF5L OR2A42 OR2A1 OR2A9P OR2A20P OR2A7 CTAGE4
chr7	143508006	143705630	197624	0.21453918	195288	OR2A20P OR2A9P OR2A1 OR2A42 ARHGEF5
chr8	39345036	39508331	163295	0.243867462	2922	ADAM5P ADAM3A ADAM3A ADAM3A

Table S10: Human Specific Duplications
Genes duplicated specifically in the human lineage (diploid in great apes). Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

gene name*	locus	variance	median copy	Vst
FCGR1A	chr1:148020873-148030698	0.28	5.64	0.03
RBM8A	chr1:144218994-144222801	0.12	3.52	0.04
HIST2H2BF	chr1:148020868-148050552	0.21	5.49	0.02
SRGAP2	chr1:204582822-204696128	0.08	5.36	0.01
PTPN20A	chr10:45970128-46040064	0.07	3.68	0.01

FRMPD2	chr10:49034611-49053173	0.08	3.94	0
GPRIN2	chr10:46413551-46420574	0.53	4.4	0.02
C10orf57	chr10:81828405-81842287	0.38	4.06	0.09
CHRFAM7A	chr15:28440734-28473156	0.42	4.27	0.05
ARHGAP11A	chr15:30694982-30715674	0.08	3.55	0
ARHGEF5	chr7:143683421-143708658	2.44	5.94	0.25
HYDIN	chr16:69398789-69822070	0.02	3.85	0.01
NCRNA00152	chr2:87536088-87602143	0.06	3.88	0.03
GTF2H2	chr5:68891809-68924334	0.84	4.49	0.28
SERF1A	chr5:69356827-69364281	0.51	3.6	0.01
SMN1	chr5:69381105-69409172	0.36	3.58	0.02
NAIP	chr5:70300065-70356697	0.66	5.04	0.1
DUSP22	chr6:237100-296355	0.33	3.98	0.1
NCF1	chr7:73826244-73841595	0.18	6.43	0
GTF2IRD2	chr7:73848419-73905777	0.09	5.5	0.02
FAM115C	chr7:143028166-143053109	0.61	3.93	0.07
LOC154761	chr7:143139993-143164743	0.63	3.94	0.06
ZNF322B	chr9:98999357-99001731	0.16	3.85	0.04

Table S11: Gene Families Expanded in the Human Lineage
Genes which are duplicated in the primate lineage and have continued to expand along the human lineage. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

gene name*	location	variance	median copy number	Vst
WASH2P	chr2:114057699-114073081	3.97	20.81	0.04
CROCCL1	chr1:16817339-16829988	2.57	10.23	0.25
MSTP2	chr1:16844655-16849501	6.45	13.85	0.06
AMY1A	chr1:103999663-104008696	7.74	9.91	0.1
FLJ39739	chr1:142510035-142535980	0.61	11.83	0.39
PDE4DIP	chr1:143663117-143706379	0.52	7.2	0.44
NBPF14	chr1:146470265-146492472	521.77	244.66	0.11
NCF1	chr7:73826244-73841595	0.18	6.43	0
LOC645166	chr1:147194909-147218219	0.56	12.5	0.02
GOLGA6L10	chr15:80420179-80428761	3.91	29.81	0.04
GIYD1	chr16:29373375-29377041	1.83	6.9	0.07
LRRC37A4	chr17:40939892-40948305	9.3	16.22	0.24
C2orf78	chr2:73864823-73897782	14.03	10.88	0.39
MGC13005	chr2:114073074-114075623	5.51	24.27	0.01
RPL23AP53	chr8:148346-172318	3.15	27.08	0.09
LOC728323	chr2:242679516-242751142	4	19.18	0.09
POM121L4P	chr22:19373842-19376009	33.18	50.71	0.3
ZNF595	chr4:43226-78099	1.27	10.63	0.18
DRD5	chr4:9392355-9394731	2.13	11.34	0.04
LOC100272216	chr5:68962735-68964784	82.32	63.64	0.13
GUSBL1	chr6:26947244-27032312	2.62	13.2	0.23
LOC100170939	chr5:69459044-69557378	2.1	9.73	0.22
C6orf41	chr6:27032750-27099731	2.99	11.48	0.15
STAG3L1	chr7:74826382-74834926	0.34	8.61	0.05
GTF2IP1	chr7:72206961-72259270	0.09	6.81	0

SPDYE5	chr7:74962234-74971564	12.4	37.77	0.25
AQP7	chr9:33374947-33392517	1.19	11.7	0.04
KGFLP1	chr9:44185511-44246438	1.16	11.29	0.03
FAM95B1	chr9:42458584-42464233	2.26	16.19	0.02
LOC642929	chr9:43130534-43135480	1.94	14.62	0.07

Table S12. Summary of SUN markers found among:

a) 12 high coverage genomes

Chromosome(s)	# SUNs	Found in at least one sample		Found in >=66.7% of samples		Found in >=83% of samples		Found in all, or all but one		Found in all samples	
chrY (of males)	163025	162715	99.81%	ND*		ND*		152857	93.76%	134327	82.40%
chrX (of females)	257838	257350	99.81%	ND*		ND*		234449	90.93%	205982	79.89%
Autosomal	3606033	3594803	99.69%	3458149	95.90%	3410578	94.58%	3267192	90.60%	3058924	84.83%
TOTAL	4026896	4014868	99.70%	3861166	95.88%	3793594	94.21%	3654498	90.75%	3399233	84.41%

b) full set of 159 genomes

		Found in at least one sample		Found in >=66.7% of samples		Found in >=80% of samples		Found in >=90% of samples		Found in all samples	
chrY (of males)	163025	162838	99.89%	59245 (in >= 48/70 males)	36.34%	16835 (in >= 57/70 males)	10.33%	10456 (in >= 64/70 males)	6.41%	4586 (in 70/70 males)	2.81%
chrX (of females)	257838	257623	99.92%	225691 (in >= 59/88 females)	87.53%	154755 (in >= 59/88 females)	60.02%	33279 (in >= 59/88 females)	12.91%	3698 (in 88/88 females)	1.43%
Autosomal	3606033	3602413	99.90%	2926297	81.15%	2079587	57.67%	679645	18.85%	137371	3.81%

*too few high coverage genomes (9 males, 3 females) to accurately estimate

Table S13. List of regions showing mirror effects, in which copy number variation occurring at a distant but homologous locus is detected by CGH or total read depth mapping but not supported by paralog-specific markers.

chr1:499-77296	chr14:105144802-105204335	chr19:8702687-8764008	chr6:170773423-170896961
chr1:511276-713934	chr14:105490591-105556154	chr19:38161285-38223951	chr7:34035-137131
chr1:1557744-1673566	chr14:105641496-105701838	chr19:45054470-45098440	chr7:32965738-33019569
chr1:6399215-6437393	chr15:18260050-18809093	chr19:48308360-48340279	chr7:38251784-38277712
chr1:12764158-12975525	chr15:19578259-19663437	chr19:59931522-60053003	chr7:43965543-44053608
chr1:13025610-13071008	chr15:19806304-19928954	chr2:13417182-13484676	chr7:56838073-56864221
chr1:13084311-13122423	chr15:20197602-20373877	chr2:87420977-87476309	chr7:57672481-57704192
chr1:13218798-13328379	chr15:20769611-20821158	chr2:87905414-88068613	chr7:57737500-57780156
chr1:16927478-16998216	chr15:20851648-21066254	chr2:88989402-89325711	chr7:62844017-62869981
chr1:17048309-17078719	chr15:21881157-21930208	chr2:89613729-89913809	chr7:64533527-64579210
chr1:21608059-21687749	chr15:26121887-26158746	chr2:91225894-91333583	chr7:64612545-64679379
chr1:25457726-25537715	chr15:26201755-26321807	chr2:94689944-94777006	chr7:64731171-64785020
chr1:25561153-25626740	chr15:26321862-26378706	chr2:94819012-94932793	chr7:64873354-64944010
chr1:83370836-83727807	chr15:26478772-26742609	chr2:95955189-96011581	chr7:66373788-66405265
chr1:103937591-104122772	chr15:26748304-26774250	chr2:97162966-97221482	chr7:71941119-71987586
chr1:108580035-108656868	chr15:28228784-28467743	chr2:106451384-106498087	chr7:73939568-73973416
chr1:108714654-108802432	chr15:29695733-29806849	chr2:110044996-110090723	chr7:74953470-74982353
chr1:116931478-117008345	chr15:30233196-30468651	chr2:111724832-112091455	chr7:75898835-75982664
chr1:120328470-120498601	chr15:30485989-30532114	chr2:112191129-112298835	chr7:75983944-76011844
chr1:120556168-120738188	chr15:45148821-45183847	chr2:112309964-112357768	chr7:76473646-76538796
chr1:120788253-120844675	chr15:70719459-70750478	chr2:113884126-113983779	chr7:99734143-99775575
chr1:142004106-142090112	chr15:72145166-72169711	chr2:114038741-114078338	chr7:100420686-100434591
chr1:142186144-142261160	chr15:73329425-73378675	chr2:165536378-165565682	chr7:101758653-101785006
chr1:142436913-142562145	chr15:75990028-76014384	chr20:25679048-25714947	chr7:101895096-102119471
chr1:142612557-142700497	chr15:80420460-80468829	chr20:25821968-25860787	chr7:126301909-126340192
chr1:143020644-143113064	chr15:80676755-80851330	chr20:26007188-26064813	chr7:142926646-143202833
chr1:143522106-143594098	chr15:80906475-81016695	chr20:28141672-28267541	chr7:143508477-143539578
chr1:143900869-144094047	chr15:82739692-82775415	chr21:13852857-14025073	chr7:143541360-143665697
chr1:146349000-146460960	chr15:83523026-83609542	chr21:14205091-14274260	chr7:143667406-143705959
chr1:146542717-146712807	chr15:100209456-100338508	chr22:15241227-15430274	chr7:151535690-151621339
chr1:146840958-146919371	chr16:2528458-2575015	chr22:17038085-17257818	chr7:157813890-157828171
chr1:147004580-147090322	chr16:5069063-5132802	chr22:18706922-18750743	chr8:73-160197
chr1:147776324-148071632	chr16:16320057-16404187	chr22:18939490-19056898	chr8:6813531-6866789

chr1:194976941-195091905	chr16:18343176-18418104	chr22:19369489-19424484	chr8:7047741-7461991
chr1:195146908-195191991	chr16:18775476-18844533	chr22:19791592-20008506	chr8:7606131-7904302
chr1:204239359-204384868	chr16:21634210-21717977	chr22:20008734-20127283	chr8:7907662-7927204
chr1:204620809-204650800	chr16:21775793-21853183	chr22:20142750-20247196	chr8:8024623-8146538
chr1:205762356-205821816	chr16:29290525-29560625	chr22:21979375-22002856	chr8:11937321-11977411
chr1:220705662-220760196	chr16:30107412-30212745	chr22:22133474-22158812	chr8:12023029-12099266
chr1:222163349-222246574	chr16:31867401-32004307	chr22:22963691-23005073	chr8:12199406-12597502
chr1:232984323-233024889	chr16:32151318-32204882	chr22:23334960-23387034	chr8:86737456-86763666
chr1:241220014-241331740	chr16:32755320-32946521	chr3:587318-627139	chr8:86851058-86913407
chr1:246670883-246701528	chr16:33004830-33206817	chr3:75464498-75899889	chr9:485-199354
chr10:50073-125923	chr16:33508063-33725429	chr3:131246442-131405726	chr9:38754947-39054160
chr10:37480058-37535779	chr16:68528337-68601385	chr3:197157403-197201638	chr9:39134780-39270784
chr10:38733715-38812330	chr16:68791706-68842482	chr3:199317200-199446787	chr9:39703727-39800094
chr10:38928865-39116929	chr16:72917919-73021767	chr4:8933780-9021628	chr9:40014824-40222966
chr10:47442802-47486994	chr16:86067717-86121360	chr4:9117494-9354801	chr9:40465886-40617095
chr10:47725753-47774013	chr16:86446982-86517872	chr4:48975057-49025982	chr9:40752326-40823000
chr10:48365756-48489335	chr16:88689291-88822221	chr4:70060340-70266462	chr9:41405823-41469268
chr10:48935328-49060727	chr17:4965165-4989740	chr4:119557627-119586189	chr9:41801881-41910251
chr10:51536286-51588070	chr17:18869145-19081192	chr4:132778147-133121338	chr9:41961482-42332096
chr10:81128690-81224323	chr17:20284345-20432628	chr4:191032481-191137850	chr9:43150072-43203636
chr10:81495386-81567953	chr17:21473391-21507142	chr4:191182042-191222546	chr9:44011107-44055066
chr10:89179686-89265924	chr17:22981956-23027009	chr5:17561678-17668989	chr9:44084665-44118479
chr11:50059-206243	chr17:23091385-23119190	chr5:21296394-21429236	chr9:44191033-44375254
chr11:3224806-3315713	chr17:31430268-31469339	chr5:21487748-21556948	chr9:45267857-45616738
chr11:3382657-3631565	chr17:31505984-31699929	chr5:21952601-22061452	chr9:46261321-46351007
chr11:4195474-4343961	chr17:33324909-33429029	chr5:34204972-34303863	chr9:65207526-65658150
chr11:60723617-60777607	chr17:33529260-33666666	chr5:49921606-50025172	chr9:65708247-65927481
chr11:71176703-71281616	chr17:33869857-33942552	chr5:61546963-61591513	chr9:67007950-67056088
chr11:88254398-88307369	chr17:40929257-41067483	chr5:68865176-69131061	chr9:67106152-67379756
chr12:17370-64628	chr17:41762682-42142846	chr5:69319374-69433460	chr9:67627850-68003967
chr12:8205041-8481089	chr17:42446144-42532220	chr5:69569367-69817074	chr9:68054080-68142779
chr12:9524006-9609667	chr17:42966030-43026172	chr5:69865639-70533749	chr9:68278801-68379303
chr12:31116113-31245092	chr17:55434740-55460046	chr5:70567964-70697786	chr9:69797264-69975213
chr12:52754366-52806269	chr17:55527349-55559551	chr5:85602854-85633701	chr9:83717389-83756427
chr13:18065871-18110282	chr17:57650372-57720112	chr5:180643835-180763081	chr9:114860491-114896671
chr13:34562609-34609933	chr17:60263711-60377923	chr6:26831072-26883319	chr9:140102069-140273191
chr13:56609653-56650416	chr17:63378002-63419729	chr6:26950195-27097376	chrX:49061057-49129879
chr13:92080803-92145168	chr18:657-95963	chr6:32066995-32093119	chrX:49180859-49256931
chr14:18070219-18431517	chr18:14421322-14589321	chr6:32099567-32124504	chrX:114866273-114919727
chr14:19264340-19377293	chr18:42796008-42816438	chr6:46903506-46939935	chrX:119890554-119948231
chr14:27280845-27401766	chr19:11070-79285	chr6:58352136-58382297	
chr14:73064493-73121425	chr19:143071-196320	chr6:160940570-160990675	

Table S14. GenBank accessions for 383 fosmid clone insert sequences corresponding to deletions used to validate SUN-based paralog-specific genotyping.

AC208148	AC216799	AC209543	AC225597	AC210423	AC214076	AC214160	AC213263
AC210994	AC225381	AC231957	AC209553	AC210988	AC229602	AC217516	AC217324
AC217141	AC225383	AC207996	AC203586	AC217060	AC205938	AC216810	AC216970
AC210971	AC225622	AC205871	AC210769	AC206486	AC216064	AC216091	AC211957
AC211947	AC207974	AC214989	AC215277	AC225702	AC215704	AC226063	AC218915
AC210431	AC212911	AC216897	AC212839	AC207974	AC211406	AC230047	AC206436
AC225590	AC210894	AC215924	AC208949	AC213535	AC215523	AC210755	AC226450
AC216084	AC226182	AC221040	AC212900	AC210436	AC225767	AC214162	AC207966
AC211782	AC215706	AC213282	AC215999	AC209204	AC229864	AC213243	AC210773
AC213505	AC229892	AC208872	AC210963	AC207593	AC209285	AC203589	AC209340
AC208591	AC216804	AC225589	AC208006	AC216693	AC225386	AC211949	AC217952
AC210708	AC225578	AC223410	AC208062	AC215336	AC210702	AC209297	AC216119
AC209280	AC210709	AC207435	AC215330	AC226126	AC207606	AC206358	AC231522
AC208065	AC213232	AC217015	AC213467	AC226739	AC216967	AC225779	AC203603
AC213246	AC208867	AC226617	AC225633	AC206489	AC226158	AC216896	AC220967
AC209539	AC208868	AC216230	AC226631	AC210996	AC217409	AC208299	AC225385
AC226589	AC216746	AC215932	AC216284	AC216026	AC212750	AC225573	AC226451
AC212295	AC206480	AC213290	AC226066	AC213259	AC225553	AC217521	AC203620
AC209547	AC226386	AC231961	AC210900	AC219165	AC225769	AC215698	AC210879
AC231268	AC213294	AC214991	AC207579	AC213059	AC226160	AC216995	AC211963
AC225832	AC210432	AC203652	AC209198	AC226743	AC212494	AC206361	AC208508
AC225576	AC213727	AC226178	AC209561	AC213061	AC216748	AC217544	AC223425
AC208179	AC213729	AC206483	AC217055	AC225708	AC216080	AC203590	AC225914
AC213260	AC213752	AC217013	AC203664	AC215337	AC213215	AC216959	AC225585
AC207977	AC215707	AC208052	AC206485	AC210440	AC216116	AC207169	AC216124
AC203608	AC213273	AC212591	AC213469	AC225581	AC206020	AC214170	AC216973
AC209239	AC216132	AC209550	AC207580	AC207595	AC209287	AC207780	AC214811
AC225033	AC216796	AC215990	AC212908	AC217139	AC207054	AC210761	AC206438
AC209237	AC203645	AC213291	AC225642	AC209277	AC214158	AC231120	AC216131
AC214014	AC214985	AC209317	AC226103	AC214988	AC216816	AC211952	AC225387
AC208167	AC208598	AC206478	AC216007	AC205937	AC229890	AC216993	AC231952
AC216809	AC212262	AC207447	AC210408	AC208151	AC217411	AC209303	AC208584
AC220942	AC207302	AC226624	AC207431	AC216741	AC215696	AC213255	AC216976

AC214174	AC225998	AC210913	AC208070	AC225327	AC207711	AC225574	AC206469
AC207984	AC210882	AC226004	AC210965	AC214075	AC210710	AC226186	AC215795
AC209048	AC215924	AC208877	AC209200	AC226134	AC226172	AC217189	AC226526
AC213220	AC206475	AC208059	AC204959	AC208159	AC216238	AC207170	AC207179
AC213737	AC209541	AC217514	AC207367	AC211321	AC213239	AC231265	AC212261
AC213744	AC226595	AC223412	AC215331	AC226156	AC225571	AC206416	AC210881
AC213749	AC221032	AC214993	AC225695	AC225708	AC208185	AC213261	AC213272
AC207778	AC216989	AC209545	AC225038	AC211237	AC216133	AC217956	AC209540
AC225916	AC216991	AC209552	AC216022	AC215328	AC216082	AC203598	AC214980
AC216921	AC203647	AC216232	AC216115	AC210539	AC225098	AC210764	AC225586
AC210542	AC206476	AC226630	AC209202	AC213116	AC225778	AC225580	AC207985
AC214809	AC212490	AC226061	AC204970	AC215344	AC230043	AC208389	AC203622
AC206736	AC225588	AC207578	AC213226	AC216800	AC209289	AC226371	AC207301
AC214182	AC207429	AC215992	AC208107	AC207604	AC203588	AC225582	AC221030
AC212258	AC216179	AC210916	AC226695	AC217405	AC206356	AC207221	

Table S15. Clones resequenced overlapping predicted duplication-embedded deletions

Supplementary Note Figure (if shown)	Band and approximate location	Genes	Genome-wide psCN map	Fosmid end sequence mapping	Individual	Clone ID	Bar-code	Reads on-target: Run A	Reads on-target: Run B	Notes
Fig. S61	4p16, 9.05-9.10 Mbp	*DEFB131*	Hemizygous deletion	Deletion	NA18517	ABC7_43084800_J22	C2	Yes	Yes	Supports predicted deletion
			Homozygous deletion	Deletion	NA18956	ABC9_43883700_M19	C3	Yes	Yes	Supports predicted deletion
			Homozygous deletion	Deletion	NA18555	ABC11_48240400_H11	C4	Yes	Yes	Supports predicted deletion
			Homozygous deletion	Deletion	NA12878	ABC12_49048900_K21	C5	Yes	Yes	Supports predicted deletion
			Homozygous deletion	Deletion	NA12156	ABC14_50205200_K7	C6	Yes	Yes	Supports predicted deletion
Fig. S62	19q13.31, 47.9-48.3 Mb	*PSG3, PSG8, PSG1, PSG7, PSG11*	Hemizygous deletion	Deletion	NA18956	ABC9_43881200_L12	G10	Yes	Yes	Supports predicted deletion
			Hemizygous deletion	Deletion	NA18956	ABC9_43862300_O17	G11	Yes	Yes	Supports predicted deletion
Fig. S63	2q11.2, ~97.61 Mbp		Homozygous deletion (~8 kb)	Normal (deletion < detection threshold)	NA12156	ABC14_1038314_G16	B1	Yes	Yes	Supports predicted deletion but does not confirm distal end
			Homozygous deletion (~8 kb)	Normal (deletion < detection threshold)	NA12878	ABC12_47948200_G24	B2	Yes	Yes	Supports deletion
Not shown	2q11.2, 97.25-97.54 Mbp	*ANKRD36B*	Deletion, or interlocus conversion	Inversion	NA12878	ABC12_46994000_B3	B3	Yes	No	Supports deletion or interlocus conversion (not shown)
			Deletion, or interlocus conversion	Inversion	NA12878	ABC12_46984600_N4	B4	Yes	No	No coverage
			Deletion, or interlocus conversion	Inversion	NA18517	ABC7_42377600_B5	B5	Yes	No	Supports deletion or interlocus conversion (not shown)
			Deletion, or interlocus conversion	Normal	NA18517	ABC7_42407100_N7	B6	Yes	Yes	Supports deletion or interlocus conversion (not shown)
Fig. S64	2q12.2, ~106.25 Mb		Homozygous deleted	deletion	NA18555	ABC11_48223100_B23	D1	yes	no	Supports deletion
			Homozygous deleted	deletion	NA19240	ABC10_44489600_C1	D2	yes	yes	Supports deletion
			Hemizygous deleted	deletion	NA12156	ABC14_50976300_M13	D3	yes	yes	Supports deletion

			Hemizygous deleted	concordant (non-deleted allele)	NA12156	ABC14_50144300_G23	D4	yes	yes	Supports non-deletion allele
			Hemizygous deleted	concordant (non-deleted allele)	NA12156	ABC14_50151700_G24	D5	yes	yes	Supports non-deletion allele
Fig. S65	4q13, ~68.81-70.38 Mbp	*UGT2B17, UGT2B15, TMPRSS1E*	Hemizygous deletion in vicinity of UGT2B17	deletion	NA19240	ABC10_1585670_B20	B1	Yes	Yes	Misassembly in reference causes apparent deletion
			Hemizygous deletion in vicinity of UGT2B17	deletion	NA12878	ABC12_47965400_B12	B2	Yes	Yes	Misassembly in reference causes apparent deletion
			Homozygous deletion in vicinity of UGT2B28	deletion	NA19129	ABC13_48744100_J5	B3	No	Yes	Supports deletion
Fig. S66	8p23.1	*FAM86B1, DEFB130*	Homozygous deletion within amplified region	Deletion	NA12878	ABC12_7942349_K16	D8	Yes	Yes	Supports predicted deletion; difference in breakpoints between clones plus genome-wide psCN signal suggests two distinct nested deletions within an amplified region with psCN = 4 flanking these deletions
			Homozygous deletion within amplified region	Deletion	NA12878	ABC12_49206400_O2	D9	Yes	Yes	Supports predicted deletion
Fig. S67	17q12, 31.40-32.02 Mbp	*CCL3L1/4L/4L2/3L3, TBC1D3C/F/G*	Homozygous deleted, except for small patches	deletion	NA12156	ABC14_1185722_G2	C9	No	No	No coverage
			Homozygous deleted, except for small patches	deletion	NA12156	ABC14_50418700_K9	C10	Yes	Yes	Supports deletion
			Homozygous deleted, except for small patches	deletion	NA12878	ABC12_46881000_B6	C11	No	Yes	Supports deletion
			Homozygous deleted, except for small patches	deletion	NA12878	ABC12_49212300_D22	C12	Yes	Yes	Supports deletion - different breakpoints
Fig. S68	15q11.2, 19.33-19.92 Mbp		Deletion(s) within amplified region	deletion	NA18956	ABC9_43887900_I3	C7	Yes	Yes	Supports deletion
			Deletion(s) within amplified region	deletion	NA18956	ABC9_43885500_N14	C8	Yes	Yes	Supports deletion
Fig. S69	17q21.31-32, 41.71-41.97 Mbp	*LRRC37A/A2, ARL14/17/17B*	Deletions within amplified region	Deletion	NA12878	ABC12_47837300_D11	E1	Yes	No	Supports deletion; does not confirm distal end. Patch of hits to proximal duplication at proximal duplication (~41.000 Mbp) supports alternative architecture bringing that patch in proximity to this deletion.
			Deletions within amplified region	Deletion	NA12878	ABC12_49206300_L15	E2	Yes	Yes	Same
			Homozygous deletion	Deletion	NA18507	ABC8_43243500_A3	E3	No	No	No coverage
			Homozygous deletion	Deletion	NA18507	ABC8_2149840_M3	E4	Yes	Yes	Supports deletion
			Partially	Inversion	NA18507	ABC8_42582000_E3	E5	Yes	Yes	Supports deletion

			homozygous (i.e. Homozygous deletion	Deletion	NA18517	ABC7_43063500_H12	E6	No	No	No coverage
			Homozygous deletion	Deletion	NA18517	ABC7_42485000_O5	E7	No	Yes	Supports deletion
Fig. S70	15q11.2, 20.254-21.009 Mbp	*TUBGCP5, CYFIP1, NIPA2, NIPA1, WHAMML, HERC2P2, GOLGA9P/ 8D/8E/6L1 NIPA1*	Hemizygous deletion	Deletion	NA18507	ABC8_41095200_D24	D11	No*	No*	*No coverage among SUNKs, but non-unique positions at locus are covered. Deletion haplotype may be contained on unanchored contigs (chr15_random) which were excluded from SUN mapping. Deletion confirms by array CGH, FISH, and Mendelian consistency.
	15q11.2, 20.632-20.690 Mbp		Single-copy deletion (to diploid) nested psCN = 3 amplification	Deletion	NA18555	ABC11_49597500_C7	D12	Yes	Yes	Supports predicted deletion
Not shown	9p12, ~39.16-39.74 Mbp		Hemizygous deletion	Deletion	NA19240	ABC10_44588000_L12	A1	Yes	Yes	Supports predicted deletion
			Hemizygous deletion	Normal (non-deleted allele)	NA19240	ABC10_43623500_O17	A2	Yes	Yes	Supports presence of non-deleted allele
			Hemizygous deletion	Normal (non-deleted allele)	NA19240	ABC10_44477800_E4	A3	Yes	Yes	Supports presence of non-deleted allele
			Homozygous deletion	deletion	NA12878	ABC12_49028400_G17	A4	No	No	No coverage
Not shown	9p11, ~44.16-46.87 Mbp		Deletion, partially homozygous	deletion	NA18507	ABC8_41061800_A22	H10	Yes	Yes	Supports predicted deletion
			Deletion, partially homozygous	inversion	NA19240	ABC10_44472100_I19	H11	Yes	Yes	Supports deletion and/or inversion
			Deletion, partially homozygous	inversion	NA18555	ABC11_48043400_I14	H12	Yes	Yes	Ambiguous - other end may be in gap
Not shown	22q11.21		Hemizygous deletion with other nested events	deletion	NA12156	ABC14_50417400_I12	D10	yes	yes	Supports deletion
Not shown	12p13.31	*KLRC2 KLRC3*	Hemizygous deletion, ~10.449-10.475 Mbp (~26 kbp)	Deletion	NA19129	ABC13_48611600_I24	G5	Yes	Yes	Ambiguous
			Hemizygous deletion, ~10.449-10.475 Mbp (~26 kbp)	Deletion	NA19129	ABC13_48888000_M12	G6	Yes	Yes	Unclear

	12p13.2		Hemizygous deletion ~11.111-11.138 Mbp (~37 kbp)	Deletion	NA19129	ABC13_48729400_D23	G9	Yes	Yes	Supports deletion
Not shown		*PRB1* *PRB2*	Hemizygous deletion, ~11.397-11.433 Mbp (~46 kbp)	Deletion	NA12878	ABC12_49283600_H13	G7	Yes	Yes	Supports deletion
			Hemizygous deletion, ~11.397-11.433 Mbp (~46 kbp)	Deletion	NA12878	ABC12_46656200_N8	G8	Yes	Yes	Supports deletion
Not shown	15q11.2		Hemizygous deletion, ~22.204-22.276 Mbp	Deletion	NA18517	ABC7_43086600_F18	F10	Yes	Yes	Supports deletion
			Hemizygous deletion, ~22.204-22.276 Mbp	Deletion	NA18507	ABC8_40979300_O18	F11	Yes	Yes	Supports deletion
			Hemizygous deletion, ~22.204-22.276 Mbp	Deletion	NA19129	ABC13_926622_P22	F12	Yes	Yes	Supports deletion
Not shown	1q23.3		Hemizygous deletion	Deletion	NA18517	ABC7_42430700_E21	F5	Yes	Yes	Supports deletion
Not shown	7p22.1, ~68.54-68.84 Mbp		Hemizygous deletion	Deletion	NA18507	ABC8_40884700_F19	F8	Yes	Yes	Supports deletion
						ABC8_40870800_M21	F9	Yes	Yes	Supports deletion

Table S16. Selected high-scoring loci for interlocus conversion signatures

Duplication coordinates		Percent identity	Conversion signature score	Gene(s)	Re
chr8 12027750-12099344	chr8 12276250-12348580	98.55%	19.39	*DUB3, FAM86B1*	
chr8 6815345-6834499	chr8 6834500-6853602	98.93%	16.02	*DEFA10P, DEFA1B, DEFA1, DEFA3*	(*S5*
chr5 783526-820318	chr5 874953-904001	97.52%	14.72	*ZDHHC11*	
chr10 17817718-18014681	chr10 18064682-18261586	99.79	14.69	*MRC, MRC1L1*	
chr1 246689711-246749190	chr1 246758720-246818330	98.61%	12.64	*OR2T34, OR2T29, OR2T3, OR2T5*	
chr2 130923990-131193978	chr2 130875596-131143174	99.24%	10.83	*CFC1, CFC1B*	
chr8 7121221-7431099	chr8 7629812-7929091	99.10%	10.11	*DEFB103B, SPAG11B, DEFB104A, DEFB106A, DEFB105A, DEFB107A, HE2, DEFB4, DEFB109*	(*S5*
chr11 89115225-89293439	chr11 89296665-89470331	99.60%	9.88	*TRIM49, TRIM53, TRIM64B, TRIM64*	
chr10 45896971-46325469	chr10 48284311-48715542	99.78%	9.06	*PTPN20A, PTPN20B*	
chr5 175282971-175491278	chr5 177066105-177280072	99.14%	7.45	*THOC3, NY-REN-7*	(*S6*
chr16 32717851-32943799	chr16 33539405-33771971	98.06%	6.68	*SLC6A10P, IGHV*	
chr7 5899795-5995202	chr7 6741801-6839043	99.07%	13.22	*RSPH10B, PMS2, PMS2CL*	(*S6*

SUPPORTING REFERENCES

S1. F. Hach *et al.*, *Nat Methods* **7**, 576 (Aug).
S2. M. A. Quail *et al.*, *Nat Methods* **5**, 1005 (Dec, 2008).
S3. C. Alkan *et al.*, *Nat Genet* **41**, 1061 (Oct, 2009).
S4. P. Lichter *et al.*, *Science* **247**, 64 (Jan 5, 1990).
S5. D. F. Conrad *et al.*, *Nature*, (Oct 7, 2009).
S6. N. Craddock *et al.*, *Nature* **464**, 713 (Apr 1, 2010).
S7. A. Itsara *et al.*, *Am J Hum Genet* **84**, 148 (Feb, 2009).
S8. S. A. McCarroll, D. M. Altshuler, *Nat Genet* **39**, S37 (Jul, 2007).
S9. J. M. Kidd *et al.*, *Nature* **453**, 56 (May 1, 2008).
S10. H. Park *et al.*, *Nat Genet* **42**, 400 (May, 2010).
S11. J. I. Kim *et al.*, *Nature* **460**, 1011 (Aug 20, 2009).
S12. S. M. Ahn *et al.*, *Genome Res* **19**, 1622 (Sep, 2009).
S13. S. C. Schuster *et al.*, *Nature* **463**, 943 (Feb 18, 2010).
S14. J. Wang *et al.*, *Nature* **456**, 60 (Nov 6, 2008).
S15. R. E. Green *et al.*, *Nature* **444**, 330 (Nov 16, 2006).
S16. G. Benson, *Nucleic Acids Res* **27**, 573 (Jan 15, 1999).
S17. A. J. Iafrate *et al.*, *Nat Genet* **36**, 949 (Sep, 2004).
S18. J. A. Bailey *et al.*, *Science* **297**, 1003 (Aug 9, 2002).
S19. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res* **11**, 1005 (Jun, 2001).
S20. W. J. Kent *et al.*, *Genome Res* **12**, 996 (Jun, 2002).
S21. S. A. McCarroll *et al.*, *Nat Genet* **40**, 1166 (Oct, 2008).
S22. J. P. Noonan *et al.*, *Am J Hum Genet* **72**, 621 (Mar, 2003).
S23. T. Bhattacharya *et al.*, *Nat Med* **15**, 1112 (Oct, 2009).
S24. S. F. Field *et al.*, *Nat Med* **15**, 1115 (Oct, 2009).
S25. E. Gonzalez *et al.*, *Science* **307**, 1434 (Mar 4, 2005).
S26. W. He *et al.*, *Nat Med* **15**, 1117 (Oct, 2009).
S27. T. J. Urban *et al.*, *Nat Med* **15**, 1110 (Oct, 2009).
S28. R. Redon *et al.*, *Nature* **444**, 444 (Nov 23, 2006).
S29. D. Brown, J. Trowsdale, R. Allen, *Tissue Antigens* **64**, 215 (Sep, 2004).
S30. K. Hirayasu *et al.*, *Am J Hum Genet* **82**, 1075 (May, 2008).
S31. Y. Xue *et al.*, *Am J Hum Genet* **83**, 337 (Sep, 2008).
S32. D. M. Altshuler *et al.*, *Nature* **467**, 52 (Sep 2, 2010).
S33. M. E. Johnson *et al.*, *Nature* **413**, 514 (Oct 4, 2001).
S34. G. H. Perry *et al.*, *Nat Genet* **39**, 1256 (Oct, 2007).
S35. J. M. Kidd, T. L. Newman, E. Tuzun, R. Kaul, E. E. Eichler, *PLoS Genet* **3**, e63 (Apr 20, 2007).
S36. H. Stefansson *et al.*, *Nature* **455**, 232 (Sep 11, 2008).
S37. M. Doornbos *et al.*, *Eur J Med Genet* **52**, 108 (Mar-Jun, 2009).
S38. S. K. Murthy *et al.*, *Cytogenet Genome Res* **116**, 135 (2007).
S39. K. Vandepoele, V. Andries, F. van Roy, *Mol Biol Evol* **26**, 1321 (Jun, 2009).
S40. I. Moore *et al.*, *Blood* **115**, 379 (Jan 14).
S41. M. C. Popesco *et al.*, *Science* **313**, 1304 (Sep 1, 2006).

S42. B. B. Rosenblum, A. D. Merritt, *Am J Hum Genet* **30**, 434 (Jul, 1978).
S43. E. Tuzun *et al.*, *Nat Genet* **37**, 727 (Jul, 2005).
S44. G. Q. Zhou *et al.*, *Placenta* **18**, 491 (Sep, 1997).
S45. J. Sambrook, D. W. Russell, *Molecular cloning : a laboratory manual*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., ed. 3rd, 2001).
S46. J. M. Chen, D. N. Cooper, N. Chuzhanova, C. Ferec, G. P. Patrinos, *Nat Rev Genet* **8**, 762 (Oct, 2007).
S47. N. D. Avent *et al.*, *Blood* **89**, 1779 (Mar 1, 1997).
S48. H. Innan, *Proc Natl Acad Sci U S A* **100**, 8793 (Jul 22, 2003).
S49. T. Kitano, N. Saitou, *J Mol Evol* **49**, 615 (Nov, 1999).
S50. F. F. Wagner, W. A. Flegel, *Blood* **95**, 3662 (Jun 15, 2000).
S51. G. P. Patrinos, P. Kollia, A. Loutradi-Anagnostou, D. Loukopoulos, M. N. Papadakis, *Hum Genet* **102**, 629 (Jun, 1998).
S52. K. Ezawa, O. O. S, N. Saitou, *Mol Biol Evol* **23**, 927 (May, 2006).
S53. D. R. Bentley *et al.*, *Nature* **456**, 53 (Nov 6, 2008).
S54. R. E. Green *et al.*, *Science* **328**, 710 (May 7, 2010).
S55. L. W. H. Tarjei S. Mikkelsen, Evan E. Eichler, Michael C. Zody,David B. Jaffe, Shiaw-Pyng Yang, Wolfgang Enard, Ines Hellmann, Kerstin Lindblad-Toh, Tasha K. Altheide, Nicoletta Archidiacono,Peer Bork, Jonathan Butler, Jean L. Chang, Ze Cheng, Asif T. Chinwalla, Pieter deJong, Kimberley D. Delehaunty, Catrina C., L. L. F. Fronick, Yoav Gilad, Gustavo Glusman, Sante Gnerre, Tina A. Graves, Toshiyuki Hayakawa, Karen E. Hayden,, H. J. Xiaoqiu Huang, W. James Kent, Mary-Claire King, Edward J. KulbokasIII, Ming K. Lee, Ge Liu, Carlos Lopez-Otin,, O. M. Kateryna D. Makova, Elaine R. Mardis, Evan Mauceli, Tracie L. Miner, William E. Nash, Joanne O. Nelson, Svante Pa¨a¨bo, Nick J. Patterson, Craig S. Pohl, Katherine S. Pollard, Kay Pru¨fer, Xose S. Puente, David Reich, Mariano Rocchi, Kate Rosenbloom, Maryellen Ruvolo, Daniel J. Richter, Stephen F. Schaffner, Arian F. A. Smit, Scott M. Smith, Mikita Suyama, James Taylor, David Torrents, Eray Tuzun, Ajit Varki, Gloria Velasco, Mario Ventura, John W. Wallis, Michael C. Wendl, Richard K. Wilson, Eric S. Lander & Robert H. Waterston, *Nature* **437**, 69 (Sep 1, 2005).
S56. T. Marques-Bonet *et al.*, *Nature* **457**, 877 (Feb 12, 2009).
S57. E. J. Hollox, J. C. Barber, A. J. Brookes, J. A. Armour, *Genome Res* **18**, 1686 (Nov, 2008).
S58. C. Nusbaum *et al.*, *Nature* **439**, 331 (Jan 19, 2006).
S59. S. Taudien *et al.*, *BMC Genomics* **5**, 92 (2004).
S60. N. Kurotaki, P. Stankiewicz, K. Wakui, N. Niikawa, J. R. Lupski, *Hum Mol Genet* **14**, 535 (Feb 15, 2005).
S61. J. Auclair *et al.*, *Hum Mutat* **28**, 1084 (Nov, 2007).

Appendix B - Supplementary Material for Chapter 3

Great ape genetic diversity and population history

Supplementary Material

Section 1: Sample collection and sequencing 199
Section 2: Mapping and SNP calling 200
2.1. Human reference mappings 200
2.2. Species reference mapping 201
2.3. Ancestral allele calls and variant orientation 204
Section 3: Validation 207
3.1. Reference fixed sites 207
3.2. Validation of segregating variants by Sanger 208
3.3. HGDP SNP arrays 209
3.4. Heterozygous variants in Clint 210
3.5. Indel validation 211
Section 4: Kinship and contamination 212
4.1. Kinship among samples 212
4.2. Contamination assessment 213
Section 5: Divergence 216
Section 6: Heterozygosity 218
Section 7: Mitochondrial reconstruction 221
Section 8: Population structure 230
8.1. Genetic structure analysis 230
8.2. Statistical support of population structure 230
8.3. Admixture deconvolution 242
8.4. AIMs 244
8.5. Phylogeny 249
8.6. Cross River gorilla 251
Section 9: Inbreeding 252
Section 10: Loss-of-function variants 255
Section 11: X versus autosomes 264
Section 12: Recent demography 269
12.1. D-statistic 272

12.2 TreeMix 273
12.3 Approximate Bayesian Computation analysis of chimpanzee demography 276
12.4 PSMC 292
Section 13: Analysis of demography and incomplete lineage sorting (ILS) 302
13.1 Incomplete lineage sorting (ILS) 302
13.2 Isolation model CoalHMM 308
References 312

Section 1: Sample collection and sequencing

A total of 88 individuals (79 non-human and 9 human great apes) were collected and sequenced in this study encompassing all the species and subspecies of great apes except for mountain gorillas (*Gorilla beringei beringei*). These populations are summarized as follows: 9 humans (*Homo sapiens*) sampled from the HGDP panel[1]; 13 bonobos without known geographical origin (*Pan paniscus*); 25 chimpanzees covering from west to east Africa (10 *Pan troglodytes ellioti*, 6 *Pan troglodytes schweinfurthii*, 4 *Pan troglodytes troglodytes*, 4 *Pan troglodytes verus*, and 1 chimpanzee hybrid); 31 gorillas from Rwanda, Cameroon and Congo (3 *Gorilla beringei graueri*, 1 *Gorilla gorilla diehli,* and 27 *Gorilla gorilla gorilla*); and 10 Sumatran and Bornean orangutans (5 *Pongo abelii* and 5 *Pongo pygmaeus*). As the aim of this study was to assess the genomic variation among natural populations of great apes, our sampling criteria maximized wild-born individuals (77%) or the first generation of captive individuals (23%). Moreover, the samples were mostly obtained from blood with the exception of at least three samples coming from low passage cell lines (**Table S1**). All samples were sequenced on an Illumina sequencing platform (HiSeq 2000) with data production at four different sequencing centers; samples were collected under the supervision of ethical committees and CITES permissions were obtained when necessary.

Section 2: Mapping and SNP calling

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Joanna L. Kelley, Dorina Twigg, Carlos D. Bustamante, Evan E. Eichler, Tomas Marques-Bonet

The goal of this study was to explore the wide spectrum of diversity in great ape populations and, thus, many of the analyses must be provided against a single reference genome. As the human reference is the best annotated primate genome, most analyses were performed with mappings to the human reference NCBI Build 36. We were aware that this could introduce biases, so we also mapped and called variants against the available nonhuman primate reference genomes. Analyses particularly sensitive to mapping were provided against these references.

2.1. Human reference mappings

Genomes were mapped to the human reference assembly NCBI Build 36 (UCSC hg18) using the BWA mapping software[2]. Read qualities were first converted/scaled to Sanger format then BWA paired-end mapping was performed using the BWA *aln* and *sampe* tools. All reads were mapped using the *aln* trim parameter *–q 15* and nonhuman genomes were additionally mapped with the increased edit distance parameter of *–n 0.01.* Pairing was performed using the *sampe* tool and limiting the maximum occurrences of a read for pairing to 1000 using the *–o 1000* option.

SNP calling was performed using the Genome Analysis Toolkit (GATK) software (version 1.4)[3]. First, samples combined by species were realigned around putative indels. SNP calling was then performed on the combined individuals for each species. SNPs were finally filtered if they met any of the following criteria:

DP < (mean_read_depth/8.0) || DP > (mean_read_depth*3)
QUAL < 33
FS > 26.0
-sites within 5 bp of a reported indel-
MQ < 25
MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)

To ensure contamination (**Suppl. Section 4**) would not contribute to any reported heterozygous calls, we applied an allele balance (AB) filter that theoretically removed the 10% of heterozygous calls with the most skewed allele balance assuming a binomial distribution of reads with *p=0.5* for the A or B alleles. Genome-wide assessments of heterozygous counts were corrected by a factor of *1/0.9*. Finally, sites overlapping predicted segmental duplications, as predicted by mrsFAST-based read-depth counts, were filtered.

We next identified all those base pairs called as a *reference* or *variant* versus those that were *uncallable* as a result of filtering thresholds (as defined by the above noted filtering parameters) to delineate the fraction of the genome that was *callable* in any particular species. We thus

generated VCFs with all bases annotated and applied the filters described above. These so-called *uncallable* bases and segmentally duplicated base pairs were then combined into a mask that defined the *callable* versus the *uncallable* fraction of the genome. Ancestry informative markers, or AIMs, (**Table 1**) were defined as specific, fixed variants at the subspecies level.

2.2. Species reference mapping

Mapping to species reference assemblies

We assessed the limitations of calling variants based on mapping primate sequences to the human genome reference by undertaking an independent alignment and variant discovery process that utilized the chimpanzee[4], gorilla[5], and orangutan[6] reference genomes. The specifics of the analysis are described below, but each species was processed using the same basic pipeline.

For each species, Illumina reads were mapped to the corresponding species reference assembly using BWA[2] version 0.5.9 with dynamic quality trimming (-q 15) and default read alignment identity thresholds (-n 0.04). Paired-end placements were identified using BWA *sampe* (with -o 1000). We performed empirical base quality score recalibration for each sequencing lane using the GATK[3,7] version 1.2-65 and identified duplicate read pairs from each library using Picard version 1.62 (http://picard.sourceforge.net/). For each species, we performed indel realignment using GATK jointly across all samples and produced a preliminary SNP callset using the GATK Unified Genotyper.

From the resulting set of candidate SNPs, we identified a high-quality set of variants using the variant quality score recalibration (VQSR) procedure implemented in GATK[3]. This procedure utilizes a training set of known variant positions to define the characteristics of high-quality calls (based on a joint analysis of criteria such as total depth, mapping quality, strand balance, etc.) and identifies a subset of the candidate SNP set that meets the resulting criteria. Using the VQSR methodology, we selected a set of SNP positions such that 99% of the SNP positions in the training set were retained. For humans, the training set for the VQSR procedure is typically derived from sets of positions known to be variant based on SNP genotyping arrays, the HapMap Project, or the 1000 Genomes Project. Such a resource is not available for the nonhuman primates considered in this paper. Instead, we created set of training SNP positions that are independent of Illumina short-read data based on capillary sequence traces obtained for each species from the NCBI trace archive based on the whole-genome shotgun (WGS) reads produced for the species reference assemblies. We mapped capillary reads using ssaha2[8] and called SNPs using the neighborhood quality score criteria implemented in ssahaSNP.

To limit the impact of segmental duplications and copy number variants in the nonhuman primate species, we further constrained the training set to those SNPs with a unique mapping to

the human genome (hg18, based on the UCSC liftOver program) and that did not overlap with duplications identified in humans, chimpanzees, gorillas, or orangutans.

All analyses were limited to the autosomes. The VQSR step utilized the following parameters:
-resource:capillary,known=false,training=true,truth=true,prior=12.0
-an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an MQ -an FS -an DP

After identifying sites such that 99% of the training set positions were retained, we created phased and imputed individual genotypes based on the Unified Genotyper output using BEAGLE[3,9] version 3.3.2. For some analyses, we masked out individual heterozygous genotypes that failed an AB filter with a two-tailed binomial p-value less than 0.05.

Gorilla
For gorilla, we used the gorGor3 gorilla genome assembly available as part of Ensembl release 62. We created a VQSR training set using 8,308,425 gorilla capillary WGS reads obtained from the NCBI trace archive:

SPECIES_CODE = 'GORILLA GORILLA' and TRACE_TYPE_CODE = 'WGS'

and limited the training set to 1,539,968 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Chimpanzee and Bonobo
For chimpanzee and bonobo, we used the panTro-2.1.4 assembly obtained from Ensembl release 65. We created a VQSR training set using 20,596,701 chimpanzee capillary WGS reads obtained from the NCBI trace archive:

CENTER_NAME = 'BI' and SPECIES_CODE = 'PAN TROGLODYTES' and CENTER_PROJECT = 'G591' and
SPECIES_CODE = 'PAN TROGLODYTES' and CENTER_NAME = 'WUGSC' and strategy = 'WGS'

and limited the training set to 1,287,455 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Orangutan
For orangutan, we used the ponAbe2 assembly obtained from the UCSC genome browser. We created a VQSR training set using 26,569,515 orangutan capillary WGS reads obtained from the NCBI trace archive:

SPECIES_CODE = 'PONGO ABELII' and TRACE_TYPE_CODE = 'WGS'

and limited the training set to 2,706,869 SNP positions with a unique liftOver to the human genome (hg18, NCBI Build 36) that did not overlap with any segmental duplications.

Comparison to hg18 SNP positions

We used the species reference SNP callset to estimate the number of SNPs missed due to mapping to the human reference, which is diverged from each nonhuman primate species. In this analysis, we did not consider SNPs as missing because of lineage-specific deletions, which result in sequences that are absent from the human genome reference.

To avoid confounding calls due to fixed differences, we limited analysis to autosomal variants that were identified as polymorphic among the analyzed samples. Only the positions of segregating sites, not individual genotypes or allele frequencies, were considered. First, we identified all segregating sites in the species reference mappings that liftOver to hg18 (**Suppl. Table 2.2.1**). We find that 10.8%–21% of the segregating sites identified from the species reference mappings with the VQSR procedure are not called as variable based on mapping to hg18. However, the hg18 callset includes hard filters as well as an explicit mask of regions of the genome that are not callable. When we limit the analysis only to the regions of hg18 where reliable calls are reported for each species, we find that 6.3%–8.7% of the segregating sites identified in the species reference mappings are not found to be polymorphic based on the hg18 mappings. We note that orangutan, which has the highest divergence from human, shows the highest rate of missing SNPs. The SNPs identified from the species reference mappings contain a mixture of true and false positive sites, and we thus take the values in **Suppl. Table 2.2.1** as an upper bound on the rate of missing variation. Overall, however, we estimate that less than 9% of variable positions are missed when consideration is limited to the portion of the hg18 genome reference where reliable calls can be made. In the opposite direction, around 15% of the sites are called in the human mappings and not in the species-specific mappings.

Species	Autosomal SNPs (species reference, only segregating sites)	Has liftOver to hg18	Not called as segregating site in hg18	Percent not identified in hg18	Has liftOver to hg18 and pass callability mask	Not called as segregating site in hg18 and pass callability mask	Percent not identified in hg18, pass mask
Pan paniscus	8,924,485	8,775,058	1,276,443	14.5%	7,993,823	507,110	6.3%
Pan troglodytes ellioti	12,977,210	12,754,863	1,373,932	10.8%	11,721,326	801,748	6.8%
Western gorilla (*Gorilla gorilla*)	17,071,573	16,412,409	2,646,214	16.1%	14,425,674	964,084	6.7%
Pongo pygmaeus	10,160,078	9,640,321	1,939,178	20.1%	8,370,456	693,949	8.3%
Pongo abelii	15,029,715	14,316,146	3,057,831	21.4%	12,287,173	1,065,615	8.7%

Suppl. Table 2.2.1 – *Identification of segregating sites called based on mapping to the species references that are not called based on mapping to the human genome reference.*

Species	Autosomal SNPs (hg18, only segregating sites)	Has liftOver to species reference (autosomes)	Not called as segregating site in species reference	Percent not identified in species reference
Pan paniscus	8,950,002	8,615,793	1,128,532	13.1%
Pan troglodytes ellioti	13,715,319	13,162,422	1,795,182	13.6%
Western gorilla (*Gorilla gorilla*)	17,217,951	16,279,942	2,548,820	15.7%
Pongo pygmaeus	10,321,213	9,540,296	1,893,039	19.8%
Pongo abelii	14,543,573	13,475,311	2,284,908	17.0%

Suppl. Table 2.2.2 – *Identification of segregating sites called based on mapping to hg18 that are not called based on mapping to the species references.*

2.3. Ancestral allele calls and variant orientation

Asger Hobolth, Marta Mele, Anders E. Halager, Thomas Mailund

To call ancestral alleles and orient variants present in the extant groups into ancestral and derived alleles, we employ Felsenstein's pruning algorithm to compute probability distributions for alleles at the inner nodes in the phylogeny and weigh these with population frequencies.

For each possible allele at the tips of the phylogeny, we built a table of the posterior probabilities at the inner nodes. Given allele counts for a site, we then compute a weighted average of posterior probabilities from this table and use this as the basis for the ancestral allele call algorithm. To avoid that varying number of calls have a large effect on the weighted posteriors, we use a pseudo count for the weighting, and we group the two *Gorilla gorilla* subspecies together as the Cross River gorilla would weigh too much otherwise.

Allele calling algorithm

The first step after computing the weighted posteriors is to classify inner nodes as either polymorphic or monomorphic.

For this, we first assume that an inner node can have at most two different alleles and then consider the two alleles with the highest and second highest posterior probability. We use the second highest posterior probability to determine if we consider the node monomorphic or polymorphic; if the second highest probability falls below a threshold, we consider the node monomorphic for the highest probability allele, and otherwise we consider it polymorphic for the two most probable alleles.

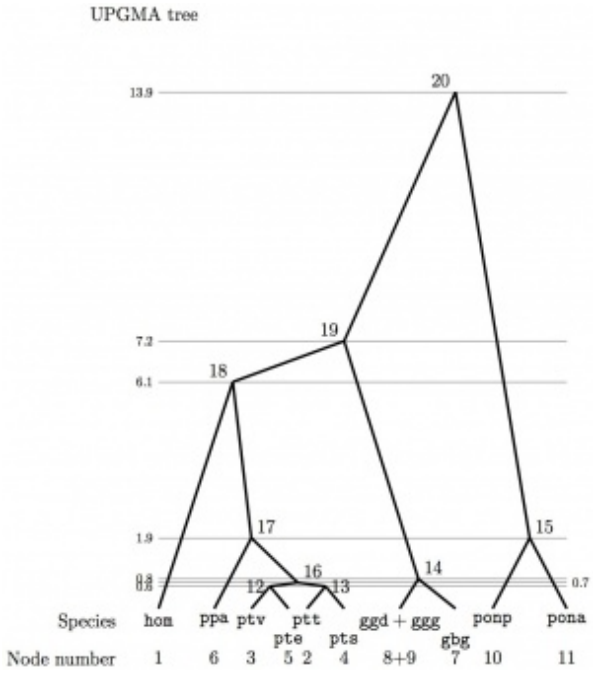
The threshold used in the monomorphic/polymorphic classification depends on the depth of the node in the tree and was determined by careful inspection of outputs from the algorithm with different choices of thresholds. The nodes within common chimpanzees (inner nodes 12, 13, and 16) have a threshold of 1%, the within-genera nodes (inner nodes 14, 15, and 17) have a threshold of 5%, and the inter-genera nodes (inner nodes 18, 19, and 20) have a threshold of 10% (**Suppl. Figure 2.3.1**).

One exception to using the posterior probabilities to determine if a node is polymorphic is when both children of a node are called as monomorphic with the same allele. In this case, we always call the parent as monomorphic.

This, generally, calls the alleles at inner nodes, but the thresholds are chosen so they are likely biased to call monomorphic nodes as polymorphic rather than the other way around, since this ensures that the allele orientation algorithm (described below) will be conservative.

Orienting polymorphisms

To orient polymorphisms, we search up the tree from each polymorphic leaf. Ancestral nodes with the same polymorphism are just stepped over. If we reach a monomorphism, the allele in this node can either be one of the alleles in the polymorphism, in which case we call that allele as ancestral, or it can be a third allele, in which case we again cannot orient the polymorphism. If the search up the tree reaches an ancestral node with a different polymorphism, we will not be able to orient the polymorphism and we give up. If we reach the root of the tree without seeing a monomorphism, we also cannot orient the polymorphism and we give up. For the root of the tree, if African and Asian apes have different alleles, we orient the root using the macaque genome.



Suppl. Figure 2.3.1 – *UPGMA phylogenetic tree. To provide the proper orientation and classification of the internal nodes, distances and different weights were used.*

Section 3: Validation

Peter H. Sudmant, Javier Prado-Martinez, Carl Baker, Maika Malig, Jessica Hernandez-Rodriguez, James C. Mullikin, Tomas Marques-Bonet, Evan E. Eichler

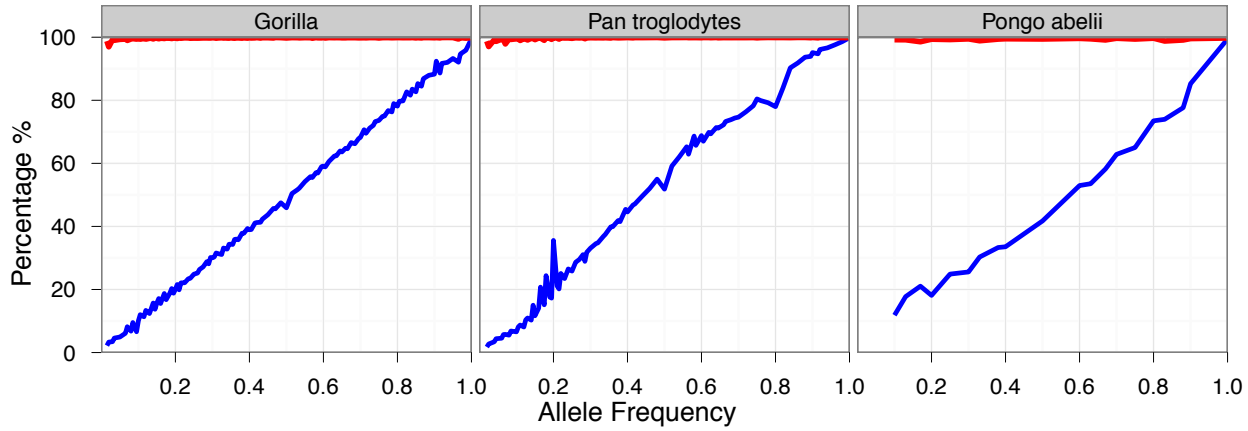
3.1. Reference fixed sites

The first evaluation of our variation was made comparing to the available reference genome assemblies[4–6]. We downloaded the genome alignments to the human NCBI Build 36 reference and the nonhuman primate genomes from the UCSC database and retrieved all annotated variants. As these alignments include many short and low-quality read alignments that lead to an excess of variation between the assemblies, we only considered those alignments longer than 1 Kbp (**Suppl. Table 3.1**). It is important to note that these variants do not only correspond to fixed events between the lineages and a significant proportion of these variants are segregating within the populations of great apes.

Reference Genome	Callable (1 Kbp alignments)	Reference SNV (compared to hg18)	Variants per bp	Overlap in Fixed Variants	Concordance in Fixed Variants
Pan troglodytes (Pantro2)	2,559,497,801 bp	34,178,305	0.0134	99.86%	99.89%
Gorilla gorilla (Gorgor3)	2,530,410,350 bp	42,055,936	0.0166	99.5%	99.87%
Pongo abelii (PonAbe2)	2,471,794,228 bp	84,759,693	0.0343	99.44%	99.76%

Suppl. Table 3.1 – *Summary table of the variants found in the reference alignments compared to the human reference genome (hg18). The overlap with the fixed variants corresponds to the percentage of our variants that are also found in the reference alignments. The concordance is the rate of variants that overlap with the reference and have the same variant allele.*

As expected, given that the references mostly consist of a single individual, the intersection of variants called from our dataset and those in the reference genomes increases as a function of the allele frequency (**Suppl. Figure 3.1**), reaching the highest overlap at fixed variants. The percentage of variants concordant with our callset varies between 99.44%–99.86% (orangutan-chimpanzee), commensurate with what we would expect given the divergence with respect to the human reference genome. We note that despite the fact that the percentage of sites in our dataset intersecting variants in the reference genomes varies widely depending on the allele frequency, the allele concordance in the variants that are found from both sources is over 99% irrespectively of the allele frequency.



Suppl. Figure 3.1 – *Comparison of the allele frequency to the percentage of overlap (blue) and the allele concordance (red) with the reference alignments. The overlap between the variants present in the reference genomes (Gorgor3, Pantro2, and Ponabe2) with respect to the variation in the different populations increases as a function of the frequency of the variants in the species. In contrast, the variants that overlap in both sources have a high concordance of the alleles that are found, meaning that when a variant is present in both the reference and our sequencing data they agree in the derived allele found. The increase in the percent of overlap in chimpanzees around 20% corresponds to the Pan troglodytes verus samples, corresponding to the subspecies used in the chimpanzee assembly.*

3.2. Validation of segregating variants by Sanger

We also performed Sanger capillary sequencing on a subset of ~480 random variants (**Suppl. Table 3.2**). Two individuals from each species were selected to be tested for a random subset of heterozygous and homozygous sites (80% heterozygotes and 20% homozygotes). Both forward and reverse strands were then sequenced. Though we obtained low validations rates for the two individuals Abe and Tzambo, likely due to poor DNA quality, our overall genotype concordance was >96%.

			Correct Calls			Incorrect Calls				
Species	**Sample**	**Sites**	**Het**	**Hom**	**Hom Ref**	**Het**	**Hom**	**Hom Ref**	**%correct**	**%correct variant only**
Chimp	Koto	61	33	17	8	1	1	1	95.1	96.0
Chimp	Vincent	68	40	9	17	1	1	0	97.1	95.9
Gorilla	Abe	54	23	16	6	8	0	1	83.3	79.5
Gorilla	Tzambo	59	34	13	6	4	2	0	89.8	87.2
Gorilla	Kokamo	77	34	13	28	1	0	1	97.4	97.9
Gorilla	Azizi	34	28	6	0	1	0	0	97.1	97.1
Orangutan	Dunja	27	12	7	6	1	1	0	92.6	89.5
Orangutan	Napoleon	23	10	7	6	0	0	0	100.0	100.0
Bonobo	Dzeeta	38	18	10	10	0	0	0	100.0	100.0
Bonobo	Hermien	38	19	10	7	2	0	0	94.7	93.1
Total/Median		**479**	**251**	**108**	**94**	**19**	**5**	**3**	**96.1**	**96.0**

Suppl. Table 3.2 – *We performed 479 Sanger sequencing experiments to validate our Illumina sequencing-based SNP calls. Overall, our median per individual concordance at variant sites was 96% confirming our SNP calls to be of high quality.*

3.3. HGDP SNP arrays

The human samples used in this study are part of the HGDP panel and were previously analyzed using Illumina 650Y SNP arrays (http://hagsc.org/hgdp/files.html). This resource allows us to compare the quality of our variant calls genome-wide. This is especially useful in the characterization of loss of heterozygotes as a consequence of the AB filters (**Suppl. Section 2.2**) applied in the contamination correction. This filter should theoretically remove 10% of the heterozygous calls, but with the SNP arrays we can assess the true overall effect of this filter. We find a general reduction of ~12%, slightly higher than the theoretical estimation, but this may also account for false heterozygous calls with skewed allele balance and additionally the HGDP samples assessed in this study were of lower coverage than the nonhuman primates, making them more susceptible to AB filtering. We also estimated the total sensitivity of both homozygous and heterozygous positions varying from the reference. Before AB filtering, we obtained a sensitivity of 99.5% compared to the SNP array while this number dropped to ~93% after the filter. Finally, we determined there to be a >99.5% overall genotype concordance in our callset. (**Suppl. Table 3.3**)

		Before Allele Balance Filter			After Allele Balance Filter		
Sample	Coverage	Heterozygous Concordance	Non-reference Sensitivity	Genotype Concordance	Heterozygous Concordance	Non-reference Sensitivity	Genotype Concordance
San HGDP01029	28.56	99.48	99.58	99.88	85.16	92.24	99.92
Han HGDP00778	21.91	99.52	99.59	99.89	87.78	93.01	99.94
French HGDP00521	21.07	99.50	99.57	99.86	88.03	92.61	99.91
Mandenka HGDP01284	19.85	99.40	99.54	99.70	87.34	92.34	99.85
Sardinian HGDP00665	19.68	99.43	99.55	99.83	87.93	92.61	99.88
Dai HGDP01307	16.82	99.23	99.49	99.77	87.48	92.86	99.83
Karitiana HGDP00998	15.56	98.91	99.44	99.62	86.88	93.68	99.78
Papuan HGDP00542	15.03	98.84	99.42	99.64	87.27	93.83	99.75
Mbuti HGDP00456	14.30	98.72	99.35	99.50	87.70	93.43	99.63
All	19.20	99.23	99.50	99.74	87.29	92.96	99.83

Suppl. Table 3.3 – *Effect of allele balance in the human samples based on SNP array comparison.*

3.4. Heterozygous variants in Clint

To provide further validation in the heterozygous calls genome-wide in a nonhuman primate, we made use of the chimpanzee reference WGS data (Clint)[4]. We downloaded all the data produced from WGS sequencing (ftp://ftp.ncbi.nih.gov/pub/TraceDB/pan_troglodytes/) and mapped to it to Pantro2 with the following parameters:

ssahaSNP des_qual 23 maxSNPs/1 Kbp 45 maxDepth 10 Qne 15 Nne 6 maxNdiff 2

Among the data used for this assembly, the primary donor was Clint; however, other chimpanzees contributed, most notably Donald. We filtered out the reads from Donald and infer the variants using the coverage and allele frequency criteria; a variant was only considered if the region was covered between 6 and 10 NQS (neighborhood quality standard) with an allele frequency ranging between 0.3 and 0.7 in order to call heterozygous variants. Variants called against the chimpanzee genome were lifted over to the human reference genome were discarded if they did not fall within the callable fraction determined for chimpanzee Illumina sequence mapping. In total, 428,022 heterozygous variants were identified from this procedure.

We compared these heterozygous positions to the Illumina calls before and after the AB filter and obtained similar performances to those with the HGDP SNP arrays. Before this filter the rate of validation was 94.3%, a number that appears low, but two factors may influence this rate: first, the liftOver process (which may produce some misalignments across species) and second, some of the variants may be derived from related sequencing projects and not necessarily from the individual Clint. After the AB filtering, the concordance in heterozygotes dropped to 86.0%, a change similar to that observed with the HGDP SNP arrays.

3.5. Indel validation

The quality of the indels (**Suppl. Section 12**) was first assessed by comparing indel variants to reference genome alignments, as previously with the fixed SNVs. Analyses were limited to 1 Kbp alignments between the references. As the alignment algorithms can place the indels with some variation around a small region, we screened for an indel of the same length and within 20 bp of the prediction. With this method, we were able to validate computationally ~97% of the variants that are predicted to damage human gene models. We further validated a random subset of 111 indels with Sanger capillary sequencing and confirmed 110 of 111 events (99.1%), representing a very low FDR (<1%) (**Suppl. Table 3.4**).

Method	Sample	Sites	Incorrect Calls	Correct Calls	%Correct
Sanger Validations	Chimpanzee	16	0	16	100.00
	Bonobo	20	0	20	100.00
	Gorilla	23	1	22	95.65
	Orangutan	52	0	52	100.00
	All	111	1	110	99.10
Reference Alignments	Chimpanzee	732	14	718	98.09
	Gorilla	813	41	772	94.96
	Orangutan	1584	45	1539	97.16
	All	3129	100	3029	96.80

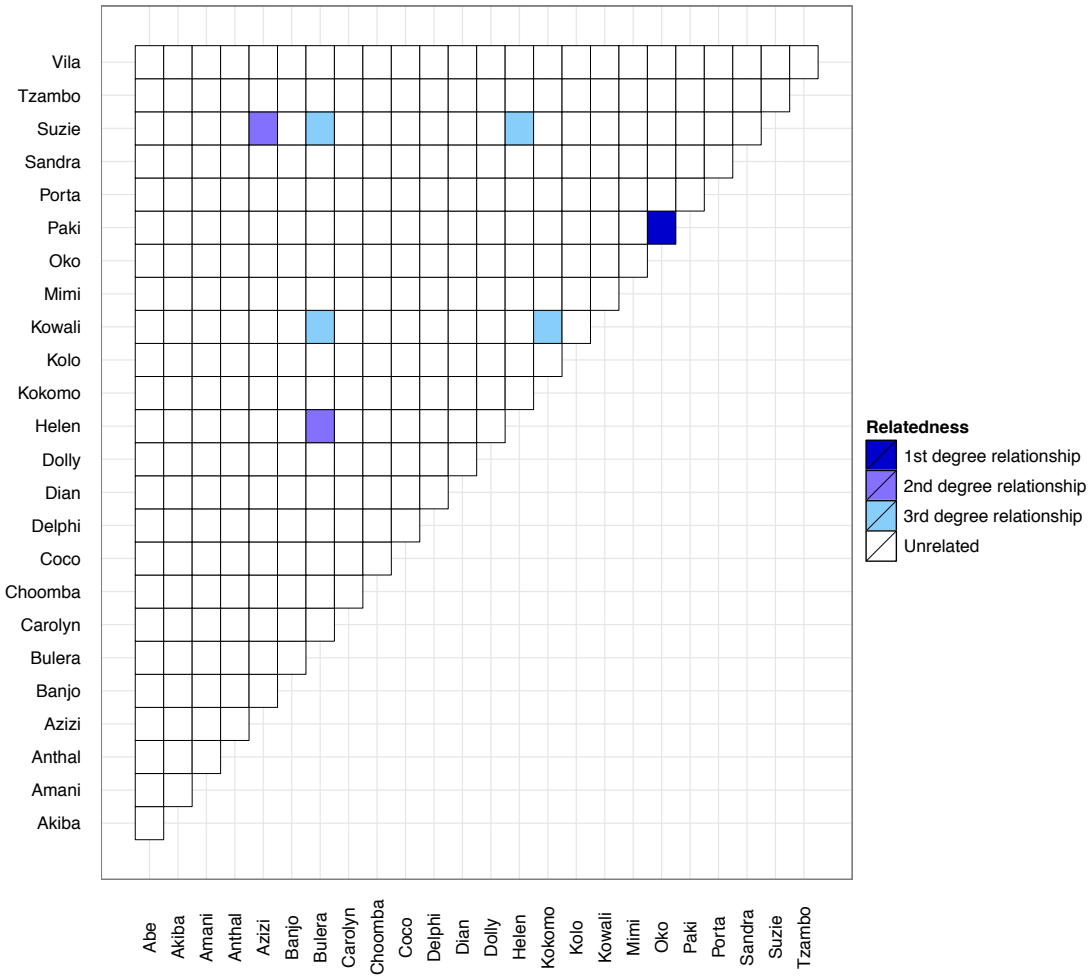
Suppl. Table 3.4 – *Indel validation rates with Sanger sequencing validation and comparison with the reference alignments.*

Section 4: Kinship and contamination

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Joanna L. Kelley, Evan E. Eichler, Tomas Marques-Bonet

4.1. Kinship among samples

Some individuals sequenced in this study are offspring of wild-caught individuals that were bred in zoos (**Table S1**). The software KING was used to estimate kinship coefficients between all the samples[10]. As expected, no hidden relatedness was identified within any of the chimpanzees, orangutans, humans, or bonobos sequenced because of the criteria of selection. Among the gorillas, the relationship between Helen and Bulera was correctly identified (**Suppl. Figure 4.1.1**) and additionally we identified a 2nd degree relationship between Azizi and Suzie, a 1st degree relationship between Paki and Oko, and 3rd degree relationships between Kowali, Bulera and Kokamo. We revised the gorilla studbook relationships with our findings and discarded the related individuals Bulera, Kowali, Suzie and Oko from all population genetic analyses.



Suppl. Figure 4.1.1 – *The degree of relatedness is plotted between all pairs of Western lowland gorillas as estimated from the coefficient of kinship.*

4.2. Contamination assessment

During the SNP quality checks we detected a significant difference in the amount of singletons in some samples when comparing the human mappings and the species-specific mappings. The larger divergence between mitochondrial genomes (4% between bonobo and chimpanzee and 14% between bonobo and orangutan mitochondria) combined with the higher mtDNA coverage allowed us to study the extent contamination. We applied two different methods to determine inter-species contamination and intra-species contamination.

For interspecific contamination we mapped all WGS data to all available mitochondrion sequences of all great apes (human, bonobo, Western chimpanzee, Western gorilla, and Bornean and Sumatran orangutans) and we recorded the unique best-quality mappings to each mitochondrion. Then, we computed the ratio of coverage between the endogenous mtDNA and the contaminant sample that could be translated into the percentage of contamination (**Suppl. Table 4.2.1**).

			Inter-species Contamination				Intra-species Contamination		
Sample	Sample Autosomic Coverage	Ratio mtDNA/ Autosomic	Sample mtDNA Coverage	Contami-nation From	Contaminated Coverage	% of mtDNA Contamination	Contamination from	Fraction of Overlap	% of mtDNA Contamination
Gbg-A929_Kaisi*	41.3	0.56	23	Ppa	7	23.333			
Ggg-A932_Mimi	38.7	27.80	1076	Ppa	1	0.093	Gbg-A929_Kaisi	0.833	1.52
Ggg-A934_Delphi	40.1	38.25	1534	Ppa	1	0.065			
Ppa-A927_Salonga	28.7	55.64	1597	None	0	0.000	Ppa-Catherine	0.918	2.21
Ppa-A928Kumbuka	43.2	35.60	1538	Ggg	10	0.646	Ppa-Catherine	0.846	2.36
Pts-100040_Andromeda	23.2	10.26	238	Ppa	2	0.833			
Pts-A911_Kidongo	49.8	2.27	113	Ppa	1	0.877			
Pts-A912_Nakuu	46.4	7.67	356	Ppa	1	0.280			
Ptt-A957_Vaillant	35	3.57	125	Ppa Pab	1 1	0.787 0.787	Ptv-A956_Jimmie	0.818	2.51
Ptt-A958_Doris	39.4	2.46	97	Pab	1	1.020			
Ptv-9730_Donald	21.731	38.01	826	None	0	0.000	Ptv-9668_Bosco	0.800	1.78
Ptv-A956_Jimmie	31.7	27.51	872	Pab	8	0.909			
Ptv-X00100_Koby	39.3	118.51	4657.5	Ggg	6	0.129			
Pab-A947_Elsi	39.8	56.51	2249	Ppa	2	0.089	Pab-A949_Dunja	0.918	2.07
Pab-A948_Kiki	34.1	69.65	2375	Ppy	1	0.042			
Pab-A949_Dunja	41.1	96.23	3955	Ppa	10	0.252			
Ppy-A940_Temmy	29.2	76.99	2248	Pab	1	0.044			
Ppy-A941_Sari	32.3	35.54	1148	Pab	1	0.087			
Ppy-A943_Tilda	37.7	75.41	2843	Pab	38	1.319			
Ppy-A944Napoleon	36.8	65.52	2411	Ppa Pab	1 29	0.040 1.188			

Suppl. Table 4.2.1 – *Summary of low contaminated samples used in the study.*

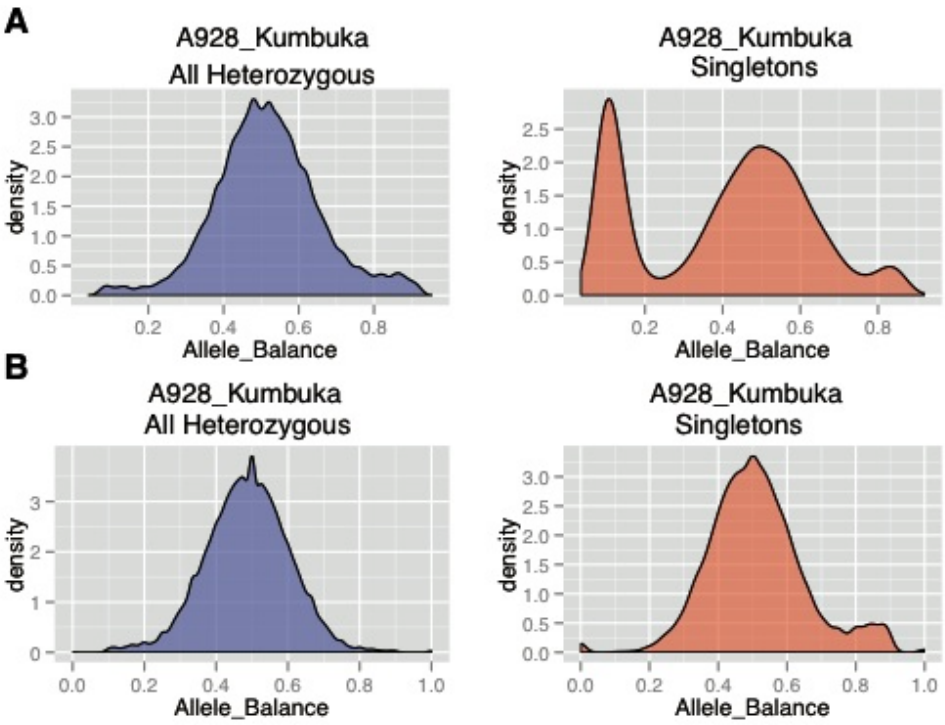
**Gbe-Kaisi: This sample show only 23X of mitochondrial coverage because it was mapped against the Western lowland gorilla mtDNA. As a result, the estimation of the contamination seems very high because only 7X of mitochondrial coverage is found.*

Assuming that the Eastern gorillas have a similar proportion of mitochondrial coverage as the Western lowland gorillas (~1000X), this estimation would be 0.7%.

To assess intraspecific contamination, we first called the mitochondrion variation to the specific references supported by the majority of reads (>90%) and the variants that had low frequency (between <50% of t2he reads). We considered the former variants as real mutations and the latter could be considered as contaminants, heteroplasmy, as well as sequencing errors. We then intersected variants with low support in one species to variants with high support in other species, which may be indicative of cross-contamination (**Suppl. Table 4.2.1**). This analysis is more ambiguous and presents biases, such as a reference effect (that can mask contamination from samples that do not have variants with respect to the reference), and possible heteroplasmy in those samples is not considered, so this number should be considered an overestimate. After this analysis, only a few samples showed traces of cross-contamination. We removed samples with higher levels of contamination (>2%), although there was a high heterogeneity in mitochondrial coverage due to different sample sources (cell lines had a higher mitochondrial coverage than blood samples) and due to the reference effect that lowered the coverage as a result of stringent mapping parameters.

We also used the autosomal portion of the genome to study the proportion of alleles present from each copy in heterozygous calls. We observed that samples with traces of contamination appear to have skewed allele balance distributions compared to the rest of the samples, where we expect a normal distribution centered on 50%. Indeed, the samples with inter-species contamination had a skewed distribution mostly in singletons (**Suppl. Figure 4.2.1**).

We corrected this problem by applying filters to the allele distribution of heterozygous variants. We applied a 0.1 two-tailed probability filter to the expected binomial distribution as a function of coverage. This conservative filter reduced the peaks in the tails of the distributions as a result of the elimination of variants with skewed allele balance, a large proportion of which indicative of contamination. However, a proportion of real heterozygous variants have been affected by this strict filtering (~10%) (**Suppl. Figure 4.2.1**).



Suppl. Figure 4.2.1 – *Effect of the filters on allele balance distribution in the heterozygous calls. This sample (Kumbuka - Bonobo), with traces of gorilla in the sequencing, shows an increase of singletons with skewed allele balance (A). These peaks are significantly reduced with the application of the allele balance (AB) filter (B).*

Section 5: Divergence

Peter H. Sudmant, Javier Prado-Martinez, Tomas Marques-Bonet, Evan E. Eichler

We initially explored the genetic relationships between individuals assessed in our study by constructing phylogenies based on the autosomal variant calls and *de novo* assembled mitochondrial genomes (**Suppl. Section 7 – Mitochondrial reconstruction**). The autosomal phylogenies were constructed from the consensus of genetic distance based neighbor-joining trees from all 10 Mbp subsegments of the genome. Consensus trees were also calculated for 1, 5, and 20 Mbp windows with little change in the overall topology of the tree. Divergence was estimated between all pairs of individuals using only sites callable among all species, defined as: (2**homs*+*hets*)/(2*callable fraction of the genome), where *homs* refers to the number of homozygous bases differing between the two individuals and *hets* the number of heterozygous sites. Divergence between two populations was thus computed as the mean pairwise divergence between all pairs of individuals between the two populations.

Homo sapiens non-African	*Homo sapiens African*	*Pan troglodytes ellioti*	*Pan troglodytes schweinfurthii*	*Pan troglodytes troglodytes*	*Pan troglodytes verus*
0.0009	0.0011	0.0121	0.0121	0.0120	0.0121
Pan paniscus	*Gorilla gorilla gorilla*	*Gorilla gorilla diehli*	*Gorilla beringei graueri*	*Pongo abelii*	*Pongo pygmaeus*
0.0122	0.0157	0.0156	0.0158	0.0303	0.0305

Suppl. Table 5.1 – *Mean divergence estimates between populations sequenced in this study and the human reference genome*.

	Western Gorilla	Bonobo	Eastern Chimpanzee	Bornean Orangutan	Sumatran Orangutan	Western Chimpanzee	Non-African Human	Nigeria-Cameroon Chimpanzee	Central Chimpanzee	African Human	Eastern Gorilla
Bonobo	0.0138										
Eastern Chimpanzee	0.0138	0.0035									
Bornean Orangutan	0.0287	0.0288	0.0288								
Sumatran Orangutan	0.0286	0.0287	0.0287	0.0032							
Western Chimpanzee	0.0138	0.0036	0.0019	0.0288	0.0287						
Non-African Human	0.0141	0.0117	0.0117	0.0291	0.0290	0.0116					
Nigeria-Cameroon Chimpanzee	0.0138	0.0036	0.0018	0.0288	0.0287	0.0017	0.0117				
Central Chimpanzee	0.0137	0.0035	0.0017	0.0287	0.0286	0.0019	0.0116	0.0018			
African Human	0.0141	0.0117	0.0116	0.0291	0.0290	0.0116	0.00095	0.0117	0.0116		
Eastern Gorilla	0.0020	0.0138	0.0138	0.0287	0.0286	0.0138	0.0141	0.0138	0.0137	0.0140	
Cross River Gorilla	0.0016	0.0137	0.0137	0.0286	0.0286	0.0137	0.0140	0.0138	0.0136	0.0140	0.00199

Suppl. Table 5.2 – *Pairwise species and subpopulation genetic distances.*

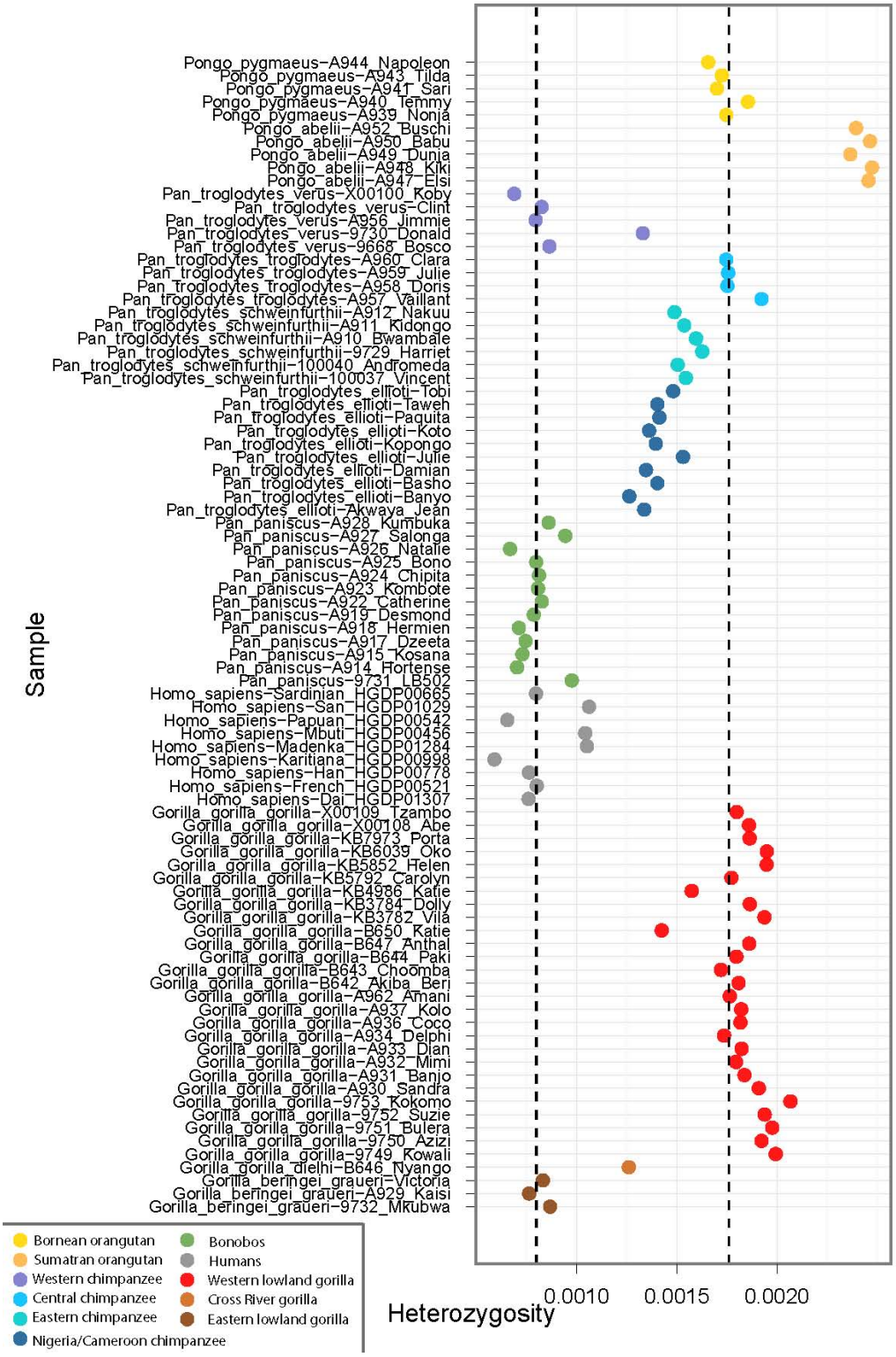
As expected, both mitochondrial and autosomal cladograms showed 100% bootstrap support for all known species relationships; however, subspecies relationships and additional population substructures varied between the two trees. Among chimpanzees, Nigeria-Cameroon and the Western individuals strikingly formed two distinct, high-confidence clades that cluster together. This result is supported by both the mitochondrial and autosomal cladograms, which contract previous reports that may have been biased due to the lack of informative markers targeted[11]. Central and Eastern chimpanzee populations each form clades that cluster together separately from the Nigeria-Cameroon and Western chimpanzees. This split for each of these populations is supported by 72% of autosomal trees, however, and 100% of mitochondrial trees place Eastern chimpanzees into a subclade of Central chimpanzees supporting these populations as being closely related.

Section 6: Heterozygosity

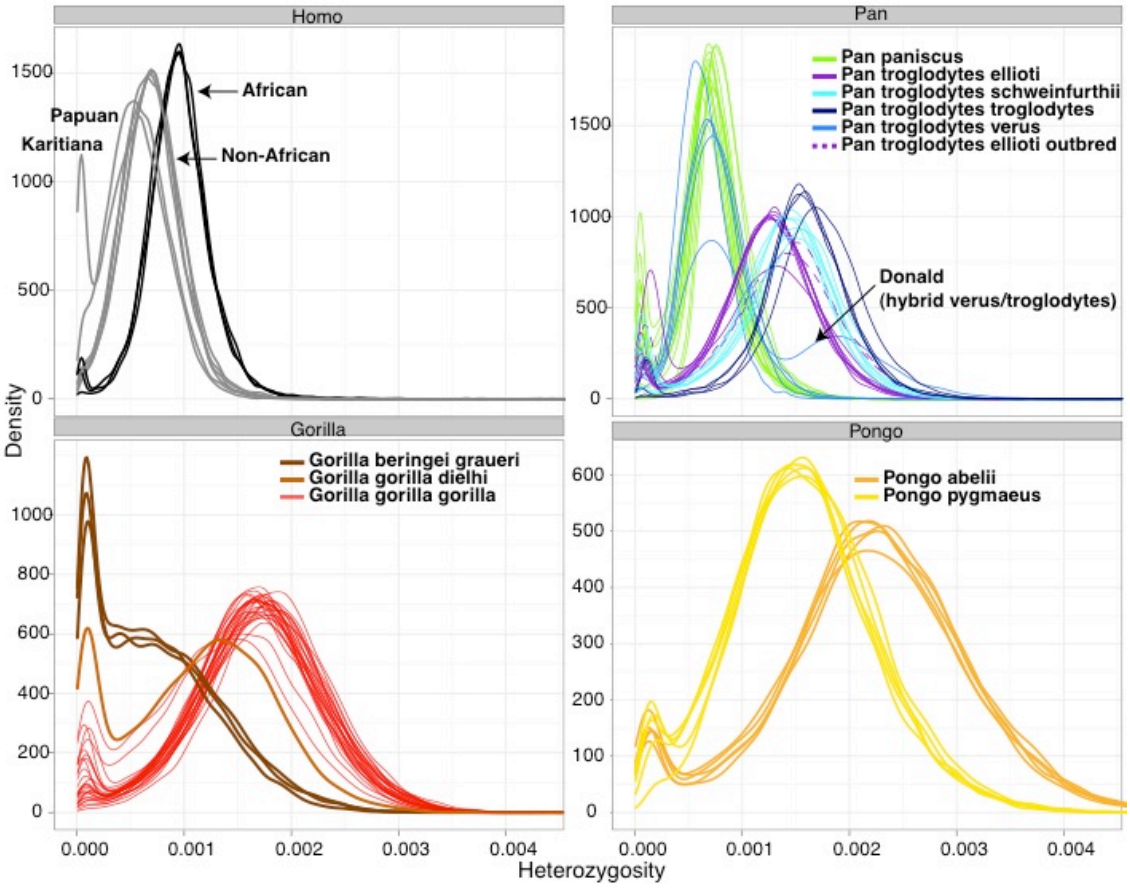
Peter H. Sudmant, Javier Prado-Martinez, Tomas Marques-Bonet, Evan E. Eichler

To assess the genetic diversity within and between great apes, we analyzed the distribution of heterozygosities in each of the species and subpopulations targeted in our study (**Figure 1b**). Relationships between the population groups were inferred by constructing a neighbor-joining tree based on inter-population divergence estimates between species and subpopulations (**Suppl. Table 5.2**). Heterozygosity estimates were computed for each individual (**Suppl. Figure 6.1**) and then combined into species/population distributions. The four related Western lowland gorilla individuals (Bulera, Kowali, Suzie and Oko) were discarded from this analysis in addition to the admixed individual Donald. We find a fourfold range in genome-wide nucleotide diversities for different hominid species and subpopulations. They range from ~6e-4 in highly bottlenecked human Karitiana and Papuan individuals to ~25e-4 in Sumatran orangutans. Non-African humans, Eastern lowland gorillas, bonobos, and Western chimpanzees all demonstrate very similar and comparatively low heterozygosities (~8e-4). Each of the chimpanzee, gorilla and orangutan genera contain high-diversity subpopulations with heterozygosities approximately twice that of these low diversity populations. Populations of intermediate diversity are also present in the gorilla and chimpanzee populations. The orangutan species demonstrates the highest overall heterozygosity of all the great apes.

A more in-depth analysis of the distribution of heterozygosity in 1 Mbp sliding windows (200 Kbp overlap) across the genome in each individual (**Suppl. Figure S6.2**) revealed a striking homogeneity in the distribution of heterozygosity in all the populations assessed, with the exception of humans. The complex demography of the human population has resulted in distinct genome-wide distributions of diversity in different populations. For example, African populations, Asian and European populations, and inbred individuals each show distinct distributions.



Suppl. Figure 6.1 – *Individual heterozygosity estimates corrected for the AB filter. The dotted lines represent heterozygosity estimated by the Chimpanzee Sequencing Consortium for Western and Central chimpanzees, respectively.*



Suppl. Figure 6.2 – *Distributions of heterozygosity of 1 Mbp windows with 200 Kbp overlap for each individual and grouped by the four genera targeted in this study. There is an enrichment of windows with low or no heterozygosity that point to events of recent inbreeding.*

Section 7: Mitochondrial reconstruction

Belen Lorente-Galdos, Gabriel Santpere, Marc Dabad, Tomas Marques-Bonet

We assembled the mitochondrial genome (mtDNA) of the samples with paired-end reads with length between 94 and 114 bp: 9 *Homo sapiens*, 5 *Pan troglodytes ellioti,* 6 *Pan troglodytes schweinfurthii,* 4 *Pan troglodytes troglodytes,* 5 *Pan troglodytes verus,* 13 *Pan paniscus,* 27 *Gorilla gorilla gorilla,* 1 *Gorilla gorilla diehli,* 2 *Gorilla beringei graueri,* 5 *Pongo abelii* and 5 *Pongo pygmaeus*. All the mtDNA were reconstructed from WGS data only (**Suppl. Table 7.1**).

	#	Read length	Median coverage per genome	Minimum coverage	Maximum coverage
Homo sapiens	9	94	22.65	16.44	35.52
Pan troglodytes ellioti	5	100	12.11	11.14	13.21
Pan troglodytes schweinfurthii	6	100	34.57	13.22	48.26
Pan troglodytes troglodytes	4	100	31.32	24.83	38.17
Pan troglodytes verus	5	100	21.05	17.21	38.09
Pan paniscus	13	100,101	32.15	11.61	46.65
Gorilla gorilla gorilla	27	100,101	21.53	12.31	38.88
Gorilla gorilla diehli	1	100	23.05	23.05	23.05
Gorilla beringei graueri	2	100,101	25.92	18.34	33.50
Pongo abelii	5	100	38.53	33.04	39.85
Pongo pygmaeus	5	100	31.60	28.24	36.55

Suppl. Table 7.1 – *Summary per species of the initial WGS reads used to reconstruct the mtDNA.*
**Coverage is computed relative to the 3 Gbp of the human assembly.*

For each sample, we captured reads from the mitochondrial genome by mapping the raw data to previously published mitochondrial assemblies of the corresponding species (**Suppl. Table 7.2**). In a second round of mapping, in order to increase the number of captured reads at the extremes of the assemblies and take advantage of the circularity of mtDNA, we aligned the reads to a modified sequence assembly, changing the origin of the reference assembly at the middle of the mtDNA in the databases (8 Kbp from the start). The mapping was carried out using mrFAST[12] with paired-end mode and 6% of divergence. We removed low-quality reads when at least one of both paired-ends had a median Phred quality score lower than 32 (**Suppl. Table 7.3**).

Species	Accession Code	Length	Length without D-loop
Homo sapiens	NCBI36.1	16,571	15446
Pan troglodytes verus	PanTro2	16,554	15441
Pan paniscus	NC_001644.1	16,563	15442
Gorilla gorilla gorilla	NC_011120.1	16,412	15448
Pongo abelii	X97707.1	16,499	15483
Pongo pygmaeus	NC_001646.1	16,389	15472

Suppl. Table 7.2 – *Mitochondrial assemblies used to capture mitochondrial reads. Notice that some control regions are not totally represented for some species.*

	Map to reference			Map to reference with modified origin		
	Median coverage per genome	Minimum coverage	Maximum coverage	Median coverage per genome	Minimum coverage	Maximum coverage
Homo sapiens	9,261.48	5,634.85	28,714.16	9,240.06	5,637.32	28,675.46
Pan troglodytes ellioti	937.89	873.75	1,252.12	933.10	865.04	1,248.50
Pan troglodytes schweinfurthii	2,076.92	546.73	2,969.60	2,059.71	542.49	2,944.84
Pan troglodytes troglodytes	980.72	503.21	1,726.18	968.19	496.28	1,712.90
Pan troglodytes verus	1,170.62	834.48	8,449.28	1,163.05	826.79	8,442.76
Pan paniscus	1,293.91	779.67	4,146.29	1,274.47	769.53	4,072.05
Gorilla gorilla gorilla	1,663.86	534.28	18,350.96	1,643.94	528.09	18,086.83
Gorilla gorilla diehli	1,069.95	1,069.95	1,069.95	1,057.71	1,057.71	1,057.71
Gorilla beringei graueri	2,277.75	953.12	3,602.38	2,247.87	938.80	3,556.95
Pongo abelii	1,325.77	1,186.36	8,877.87	1,307.68	1,170.10	8,777.60
Pongo pygmaeus	1,823.21	925.90	2,344.22	1,831.11	930.65	2,344.79
All	**1,620.45**	**503.21**	**28,714.16**	**1,597.90**	**496.28**	**28,675.46**

Suppl. Table 7.3 – *Mitochondrial coverage relative to the length of the mitochondrial reference of the corresponding species.*

We then constructed contigs for mtDNA using Hapsembler[13] (-p Illumina -t 4 -d no --PHRED_OFFSET 33 --MIN_CONTIG_SIZE 1000 –EPSILON 0.05). These contigs were finally oriented via local alignments to the corresponding reference assembly (using BLAST[14]) and then joined (using mafft[15]) incorporating N's in the existing gaps.

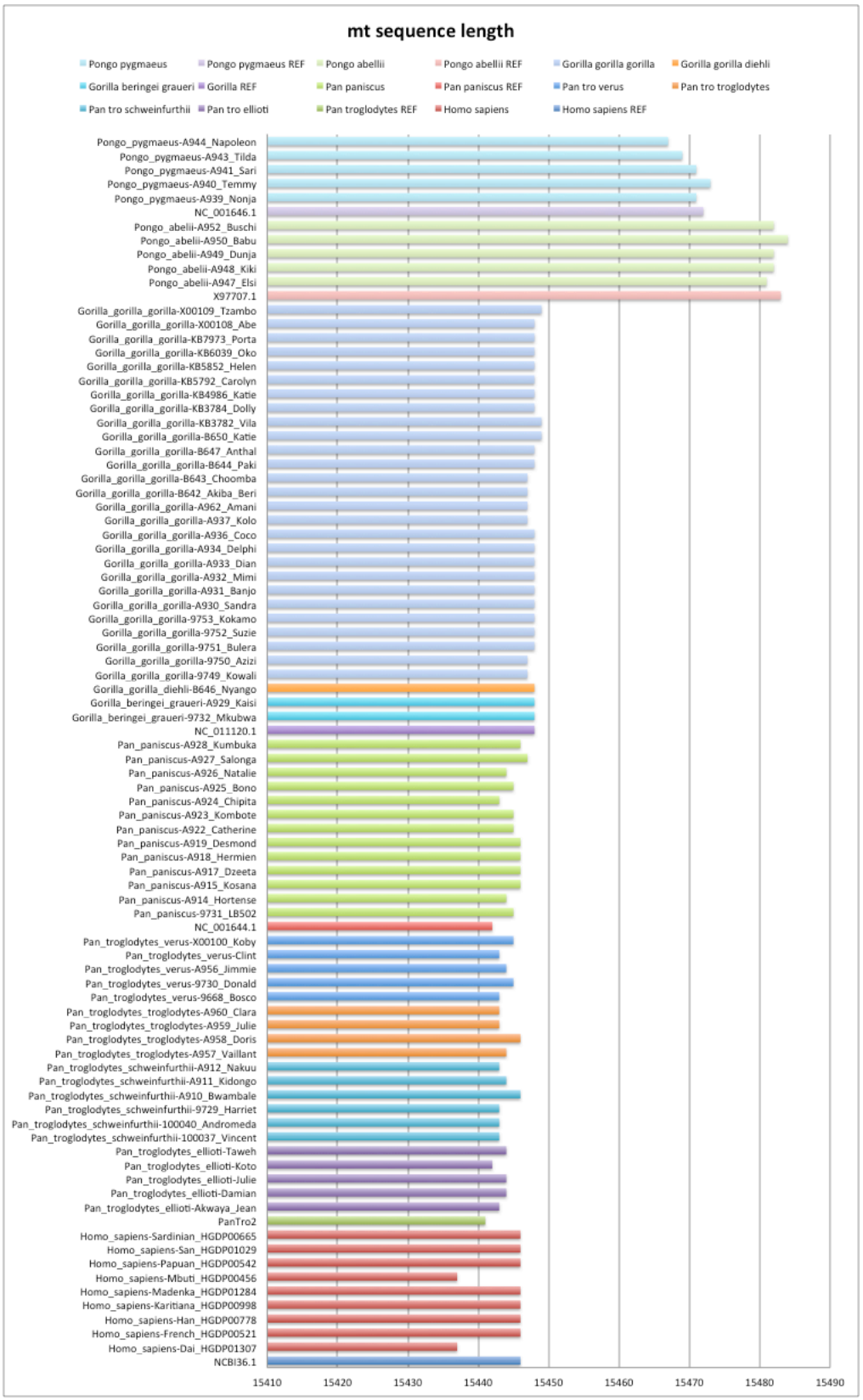
We reduced the final coverage to 350X so the assembler improved its efficiency; the highly variable D-loop region of the mitochondria was also eliminated. To compensate the random representation caused by this reduction and also to reduce the potential problem of *numts* into our reconstructions, we recreated a random reduction of coverage to 350X replicated five times. We did this step twice, once per each reference assembly (standard and changing the origin of the assembly). Thus, we created 10 mitochondrial assemblies per individual. The consensus sequence resulting from the 10 assemblies is the final mitochondrial assembly per sample.

A phylogenetic tree from the final assemblies was created using RAxML (parameters -m GTRGAMMA -# 1000 -n T1 -T 8, for deducing the best tree and, -T 8 -n result -# 1000 -x 12345 -p 12345 -m GTRGAMMA, for calculating bootstrap values).

We were able to reconstruct the mitochondrial sequence of the 82 samples studied, with no gaps outside the D-loop. Except for one individual, the resampling was carried out by reducing coverage to 350X. For the orangutan A949_Dunja, we considered the sequence obtained with the resampling done to 300X. The length of the sequences we obtained is shown in **Suppl. Figure 7.1**.

The mitochondrial sequences of 14 of our samples (4 *Pan troglodytes troglodytes*, 7 *Pan paniscus*, and 3 *Gorilla gorilla gorilla*) were independently obtained via long-range PCR and Illumina sequencing[16] (Hvilsom et al. in prep). We compared the sequences obtained through both methods (**Suppl. Table 7.4**).

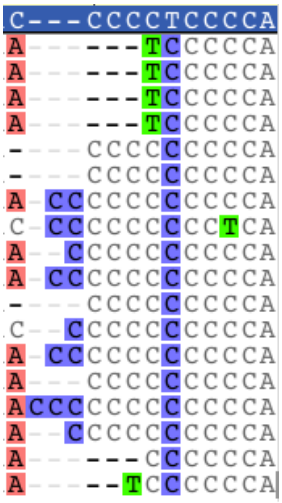
All sequences analyzed show a high level of identity. The differences of these samples correspond only to three regions (~379-389, ~7963-7965, ~8450-8471, starting from the end of the D-loop). The first region is shown in **Suppl. Figure 7.2**, and it is a complex region with homopolymers that complicates its correct identification by any of the two methods.


mt sequence length
Pongo pygmaeus
Pongo pygmaeus REF
Pongo abellii
Pongo abellii REF
Gorilla gorilla gorilla
Gorilla gorilla diehli
Gorilla beringei graueri
Gorilla REF
Pan paniscus
Pan paniscus REF
Pan tro verus
Pan tro troglodytes
Pan tro schweinfurthii
Pan tro ellioti
Pan troglodytes REF
Homo sapiens
Homo sapiens REF
Pongo_pygmaeus-A944_Napoleon
Pongo_pygmaeus-A943_Tilda
Pongo_pygmaeus-A941_Sari
Pongo_pygmaeus-A940_Temmy
Pongo_pygmaeus-A939_Nonja
NC_001646.1
Pongo_abelii-A952_Buschi
Pongo_abelii-A950_Babu
Pongo_abelii-A949_Dunja
Pongo_abelii-A948_Kiki
Pongo_abelii-A947_Elsi
X97707.1
Gorilla_gorilla_gorilla-X00109_Tzambo
Gorilla_gorilla_gorilla-X00108_Abe
Gorilla_gorilla_gorilla-KB7973_Porta
Gorilla_gorilla_gorilla-KB6039_Oko
Gorilla_gorilla_gorilla-KB5852_Helen
Gorilla_gorilla_gorilla-KB5792_Carolyn
Gorilla_gorilla_gorilla-KB4986_Katie
Gorilla_gorilla_gorilla-KB3784_Dolly
Gorilla_gorilla_gorilla-KB3782_Vila
Gorilla_gorilla_gorilla-B650_Katie
Gorilla_gorilla_gorilla-B647_Anthal
Gorilla_gorilla_gorilla-B644_Paki
Gorilla_gorilla_gorilla-B643_Choomba
Gorilla_gorilla_gorilla-B642_Akiba_Beri
Gorilla_gorilla_gorilla-A962_Amani
Gorilla_gorilla_gorilla-A937_Kolo
Gorilla_gorilla_gorilla-A936_Coco
Gorilla_gorilla_gorilla-A934_Delphi
Gorilla_gorilla_gorilla-A933_Dian
Gorilla_gorilla_gorilla-A932_Mimi
Gorilla_gorilla_gorilla-A931_Banjo
Gorilla_gorilla_gorilla-A930_Sandra
Gorilla_gorilla_gorilla-9753_Kokamo
Gorilla_gorilla_gorilla-9752_Suzie
Gorilla_gorilla_gorilla-9751_Bulera
Gorilla_gorilla_gorilla-9750_Azizi
Gorilla_gorilla_gorilla-9749_Kowali
Gorilla_gorilla_diehli-B646_Nyango
Gorilla_beringei_graueri-A929_Kaisi
Gorilla_beringei_graueri-9732_Mkubwa
NC_011120.1
Pan_paniscus-A928_Kumbuka
Pan_paniscus-A927_Salonga
Pan_paniscus-A926_Natalie
Pan_paniscus-A925_Bono
Pan_paniscus-A924_Chipita
Pan_paniscus-A923_Kombote
Pan_paniscus-A922_Catherine
Pan_paniscus-A919_Desmond
Pan_paniscus-A918_Hermien
Pan_paniscus-A917_Dzeeta
Pan_paniscus-A915_Kosana
Pan_paniscus-A914_Hortense
Pan_paniscus-9731_LB502
NC_001644.1
Pan_troglodytes_verus-X00100_Koby
Pan_troglodytes_verus-Clint
Pan_troglodytes_verus-A956_Jimmie
Pan_troglodytes_verus-9730_Donald
Pan_troglodytes_verus-9668_Bosco
Pan_troglodytes_troglodytes-A960_Clara
Pan_troglodytes_troglodytes-A959_Julie
Pan_troglodytes_troglodytes-A958_Doris
Pan_troglodytes_troglodytes-A957_Vaillant
Pan_troglodytes_schweinfurthii-A912_Nakuu
Pan_troglodytes_schweinfurthii-A911_Kidongo
Pan_troglodytes_schweinfurthii-A910_Bwambale
Pan_troglodytes_schweinfurthii-9729_Harriet
Pan_troglodytes_schweinfurthii-100040_Andromeda
Pan_troglodytes_schweinfurthii-100037_Vincent
Pan_troglodytes_ellioti-Taweh
Pan_troglodytes_ellioti-Koto
Pan_troglodytes_ellioti-Julie
Pan_troglodytes_ellioti-Damian
Pan_troglodytes_ellioti-Akwaya_Jean
PanTro2
Homo_sapiens-Sardinian_HGDP00665
Homo_sapiens-San_HGDP01029
Homo_sapiens-Papuan_HGDP00542
Homo_sapiens-Mbuti_HGDP00456
Homo_sapiens-Madenka_HGDP01284
Homo_sapiens-Karitiana_HGDP00998
Homo_sapiens-Han_HGDP00778
Homo_sapiens-French_HGDP00521
Homo_sapiens-Dai_HGDP01307
NCBI36.1
15410
15420
15430
15440
15450
15460
15470
15480
15490

Suppl. Figure 7.1 – *Lengths of the sequences without D-loop. The lengths of the references used for mapping are also shown.*

Sample-ID *	Sample-ID **	Length *	Length **	Matches	Mismatches	Indels	# bp in indels
Ptt_Vaillant	A957_Vaillant	15444	15444	15444	0	0	0
Ptt_Doris	A958_Doris	15447	15446	15446	0	1	1
Ptt_Julie	A959_Julie	15443	15443	15443	0	0	0
Ptt_Clara	A960_Clara	15449	15443	15443	0	1	6
Ppa_Hortense	A914_Hortense	15444	15444	15443	1	0	0
Ppa_Kosana	A915_Kosana	15448	15446	15445	1	1	2
Ppa_Dzeeta	A917_Dzeeta	15447	15446	15446	0	1	1
Ppa_Hermien	A918_Hermien	15448	15446	15445	1	1	2
Ppa_Desmond	A919_Desmond	15448	15446	15445	1	1	2
Ppa_Natalie	A926_Natalie	15444	15444	15444	1	0	0
ppa_Kumbuka	A928_Kumbuka	15446	15446	15446	1	0	0
Ggg_Mimi	A932_Mimi	15448	15448	15443	5	0	0
Ggg_Dian	A933_Dian	15448	15448	15443	5	0	0
Ggg_Amani	A962_Amani	15447	15447	15442	5	0	0

Suppl. Table 7.4 – *Comparison of the sequences obtained via two independent methods. *Long-range PCR plus Illumina sequencing. ** Assemblies in this study from Illumina data.*



Suppl. Figure 7.2 – *Alignment of the sequences obtained through long-PCR and Illumina sequencing. In this region, approximately at 380 bp from the D-loop, many of these sequences differ from the ones obtained via our method.*

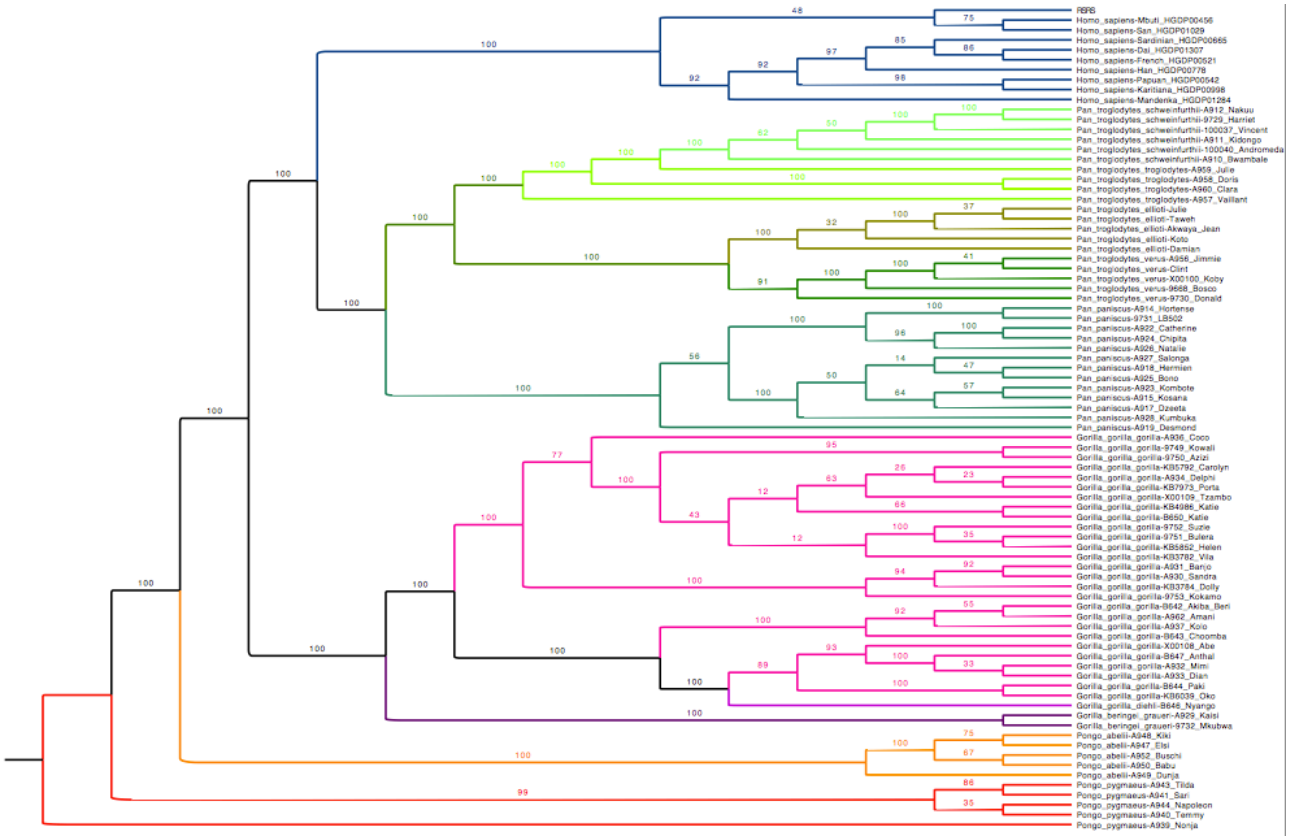
Human haplogroups

For the human samples, we identified the variants relative to a high-quality sequence[17] (**Suppl. Table 7.5**). The haplogroups for these sequences match the expected population they belong to. All common variants that define each haplogroup are found in our sequences.

Population	Sample ID	Haplogroup	Variants
Dai	HGDP01307	B4a1c4	709A, 769G, 825T, 1018G, 2758G, 2885T, 3594C, 4104A, 4312C, 5465C, 7146A, 7256C, 7521G, 8281-8289d, 8468C, 8655C, 8701A, 9123A, 9540T, 10238C, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11914G, 12705C, 12904G, 13105A, 13276A, 13506C, 13650C
French	HGDP00521	T1a	709A, 769G, 825T, 1018G, 1888A, 2758G, 2885T, 3394C, 3594C, 4104A, 4216C, 4312C, 4639C, 4917G, 7146A, 7256C, 7521G, 8468C, 8655C, 8697A, 8701A, 9540T, 10398A, 10463C, 10664C, 10688G, 10810T, 10873T, 10915T, 11251G, 11914G, 12633A, 12705C, 13105A, 13276A, 13368A, 13506C, 13650C, 14905A, 15452A, 15607G, 15928A
Han	HGDP00778	A5b	663G, 769G, 825T, 961C, 1018G, 1709A, 1736G, 2758G, 2885T, 3594C, 4104A, 4248C, 4312C, 4316G, 4824G, 7146A, 7256C, 7521G, 8468C, 8563G, 8655C, 8701A, 8794T, 9540T, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11536T, 11914G, 13105A, 13276A, 13506C, 13650C, 13999T
Karitiana	HGDP00998	D1	769G, 825T, 1018G, 1821G, 2092T, 2758G, 2885T, 3010A, 3591A, 3594C, 4104A, 4312C, 4883T, 5178A, 7146A, 7256C, 7521G, 8414T, 8468C, 8655C, 10118C, 10400T, 10664C, 10688G, 10810T, 10915T, 11914G, 11928G, 12732C, 13105A, 13276A, 13506C, 13650C, 14256C, 14668T, 14783C, 15043A, 15301A
Mandenka	HGDP01284	L2c3a	680C, 709A, 825T, 1442A, 2332T, 2416C, 2589G, 2758G, 2885T, 3200A, 4312C, 5255T, 6521T, 7146A, 7624A, 8206A, 8468C, 8655C, 8733C, 9221G, 10115C, 10664C, 10688G, 10810T, 10915T, 11914G, 11944C, 12236A, 13105A, 13276A, 13506C, 13590A, 13928C, 13958C, 15077A, 15110A, 15217A, 15301A, 15849T
Mbuti	HGDP00456	L0a2b	1048T, 2245G, 3372C, 3516A, 4586C, 5147A, 5231A, 5237A, 5442C, 5460A, 5603T, 5711G, 6185C, 6257A, 8281-8289d, 8428T, 8460G, 8566G, 9042T, 9347G, 9755A, 9818T, 10589A, 11172G, 11176A, 11269T, 11641G, 12007A, 12172G, 12720G, 13281C, 14308C, 15136T, 15431A
Papuan	HGDP00542	Q3a	769G, 825T, 1018G, 2758G, 2768G, 2885T, 3594C, 4104A, 4117C, 4312C, 4335T, 5843G, 7146A, 7256C, 7521G, 8468C, 8578T, 8655C, 8790A, 10400T, 10664C, 10688G, 10810T, 10915T, 11260C, 11914G, 12940A, 13105A, 13276A, 13500C, 13506C, 13650C, 14783C, 15043A, 15172A, 15301A
San	HGDP01029	L0d1b1	719A, 1048T, 1438A, 2706A, 3438A, 3516A, 3618C, 3756G, 4232C, 5029C, 5442C, 6185C, 6266G, 6815C, 7283C, 8113A, 8152A, 8251A, 8383C, 8937C, 9042T, 9347G, 9755A, 10589A, 12007A, 12121C, 12720G, 13759A, 14315T, 14659T, 15466A, 15692G, 15930A, 15941C
Sardinian	HGDP00665	H3u	769G, 825T, 1018G, 2706A, 2758G, 2885T, 3594C, 4104A, 4312C, 6776C, 7028C, 7146A, 7256C, 7521G, 8468C, 8655C, 8701A, 9540T, 9966A, 10398A, 10664C, 10688G, 10810T, 10873T, 10915T, 11719G, 11914G, 12705C, 13105A, 13276A, 13506C, 13650C, 14766C, 15315T

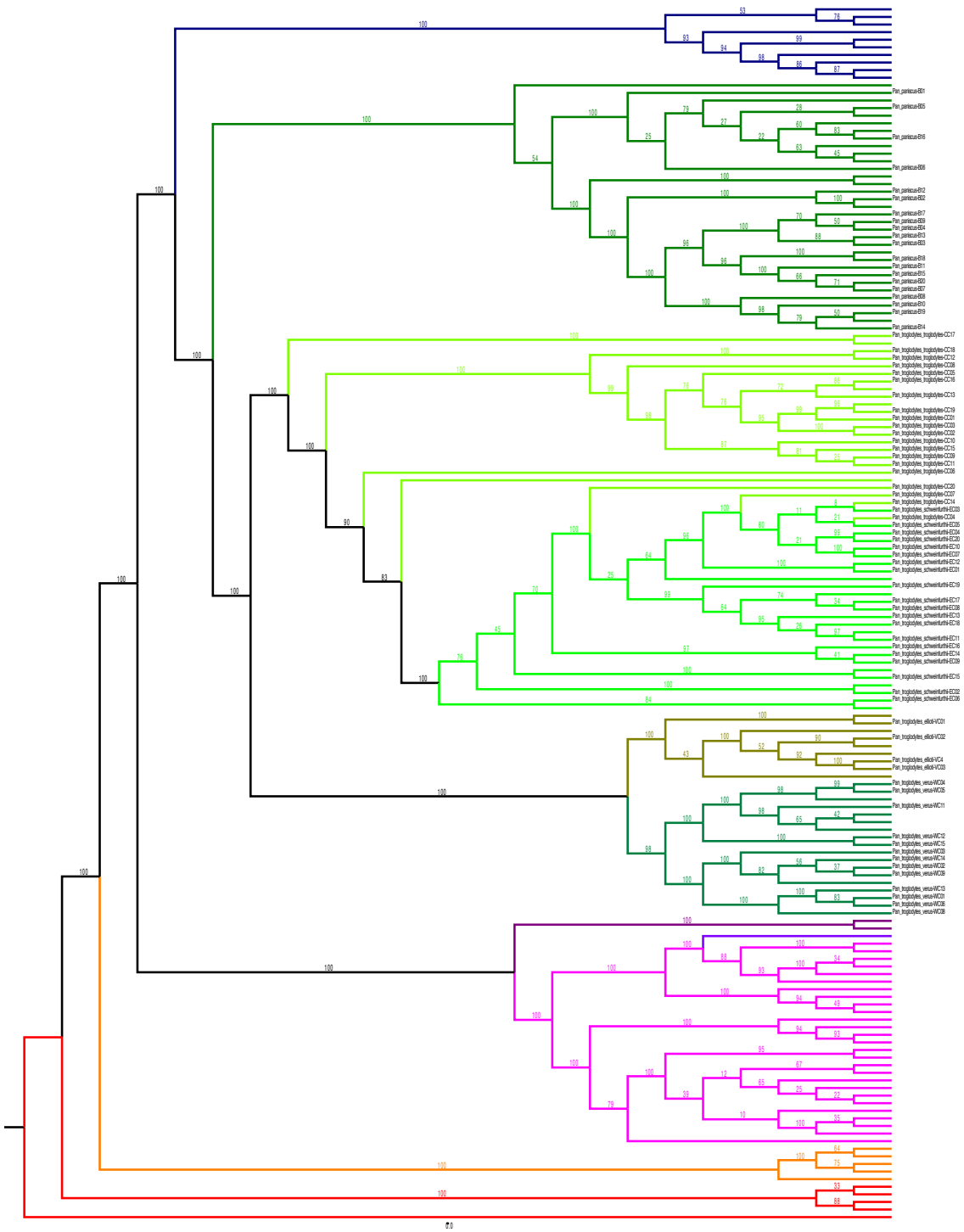
Suppl. Table 7.5 – *Human haplogroups. The variants are given in coordinates of a high-quality human reference*[17]*.*

The phylogenetic tree of our sequences is shown in **Suppl. Figure 7.3**. All sequences were ultimately aligned to the human reference[17], which is also included in the tree.



Suppl. Figure 7.3 – *Phylogenetic tree of 82 samples and the human reference.*

Another 77 individuals (20 *Pan troglodytes troglodytes*, 20 *Pan troglodytes schweinfurthii*, 4 *Pan troglodytes ellioti*, 13 *Pan troglodytes verus*, and 20 *Pan paniscus*) from a previous publication[16] have been included in the global phylogenetic tree to ensure the results of our classification (**Suppl. Figure 7.4**).



Suppl. Figure 7.4 – *Phylogenetic tree of our 82 samples, 77 from a previous study*[16] *and the human reference. Only the 77 added individuals are labeled in this tree.*

Section 8: Population structure

8.1. Genetic structure analysis

Timothy D. O'Connor, Joshua Akey

For all groups we used the following settings from PLINK[18]:
PLINK –geno 0.1 –maf MAF –indep-pairwise 50 5 0.1

In the case of *Homo sapiens*, *Pongo,* and *Pan paniscus*, we used MAF = 0.06 to remove singletons, otherwise we used 0.05 as the MAF threshold. Additionally, we removed all SNVs from the X chromosome to make the resulting data autosomal only. This resulted in SNV counts of 96,473 (*Homo sapiens*); 330,941 (*Gorilla* combined; with Katies 361049); 139,547 (*Pan paniscus*); 342,781 (*Pan troglodytes*); and 271,889 (*Pongo*).

Using this filtered data, we performed a Frappe analysis[19]. Frappe is a maximum likelihood method that finds a preset number of clusters of structure (K) in genetic data. For the expectation-maximization algorithm utilized by Frappe we used a maximum of 50,000 iterations and a likelihood increase termination threshold of 0.001 per step, the latter being the primary criterion for termination with these data. For each value of K (2 to 8) we ran the analysis 10 times and selected the highest log likelihood. All visualization was done in the R software environment (**Figure 1**).

8.2. Statistical support of population structure

Jeffrey M. Kidd, Joanna L. Kelley

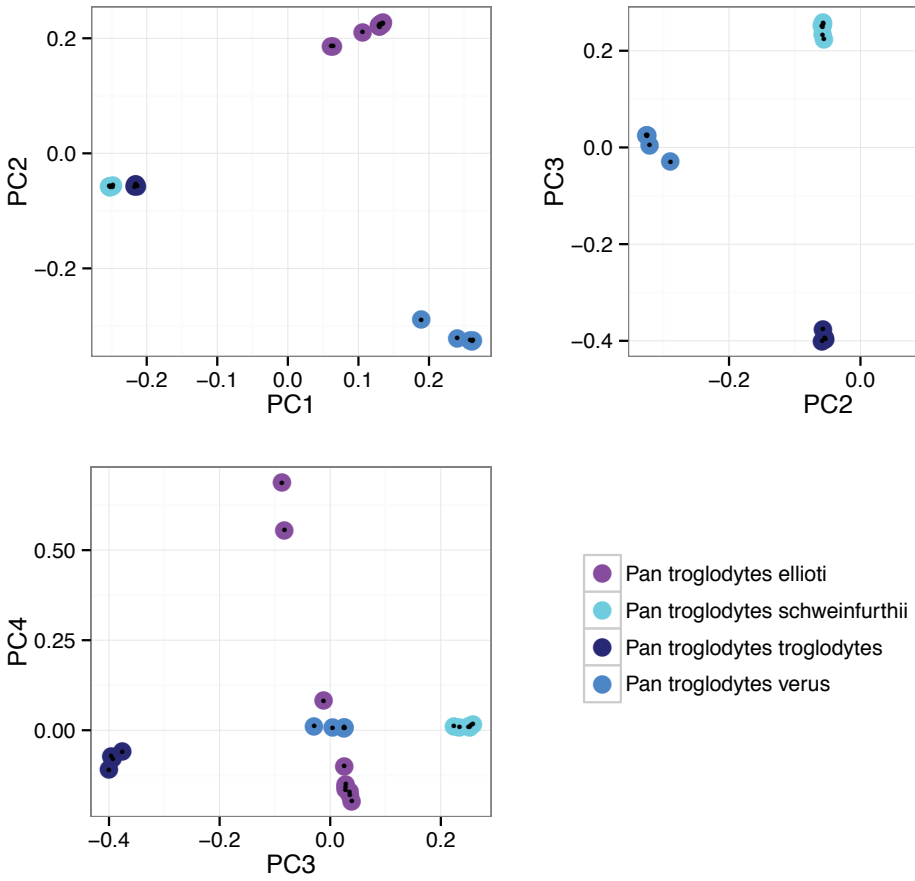
We based our analysis on an "LD-thinned" set of polymorphic sites identified in each group of samples. We first limited analysis to the indicated samples, then pruned for minor allele frequency, rate of missing genotypes, and LD using PLINK[18] (--geno 0.1 --maf 0.05 and --indep-pairwise 50 5 0.1). We only considered autosomal SNPs and removed variants on random or unassigned chromosomes. The dataset we used had an AB filter applied to heterozygous sites. The AB filter introduces a bias against high frequency events in the final dataset, since higher frequency alleles have more heterozygous genotypes, which can potentially be filtered out. Using this set of variants (**Suppl. Table 8.2.1**), we assessed the evidence of population structure based on PCA (using the smartpca program in EIGENSOFT[20] and associated Tracy–Widom statistics) and ADMIXTURE[21] with 10-fold cross-validation[22]. ADMIXTURE analysis was performed five times using independent random seeds for each interrogated value of K. We conducted analysis on SNP sets with and without the AB filter and obtained concordant results. We therefore focus on the results obtained with the AB filter.

Group	Samples	SNPs (with AB filter)	SNPs (no AB filter)
Chimpanzee	25	342,781	471,650
Bonobo	13	139,547	213,022
Gorilla	27	361,050	559,330
Orangutan	10	430,014	548,199
Chimpanzee + Bonobo	38	392,940	505,507

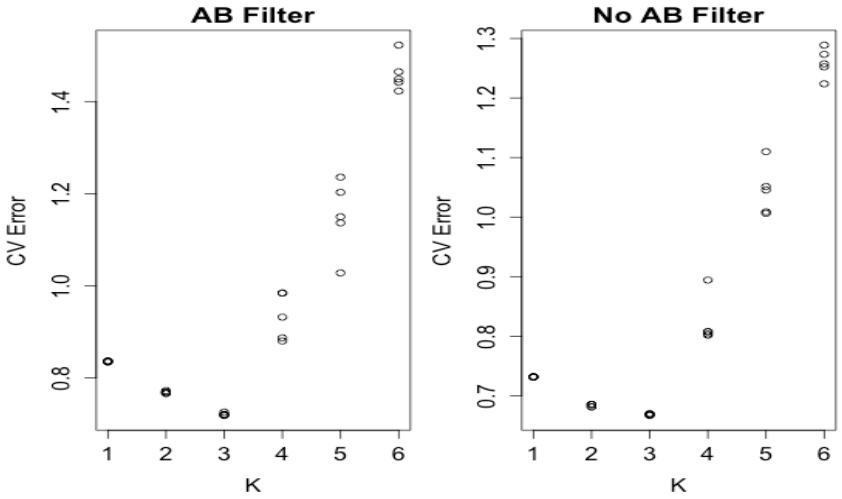
Suppl. Table 8.2.1 – *Summary of SNPs used for analysis. The combined chimpanzee and bonobo callset was constructed based on the individual species calls using the merged mask of the callable genome.*

Chimpanzee

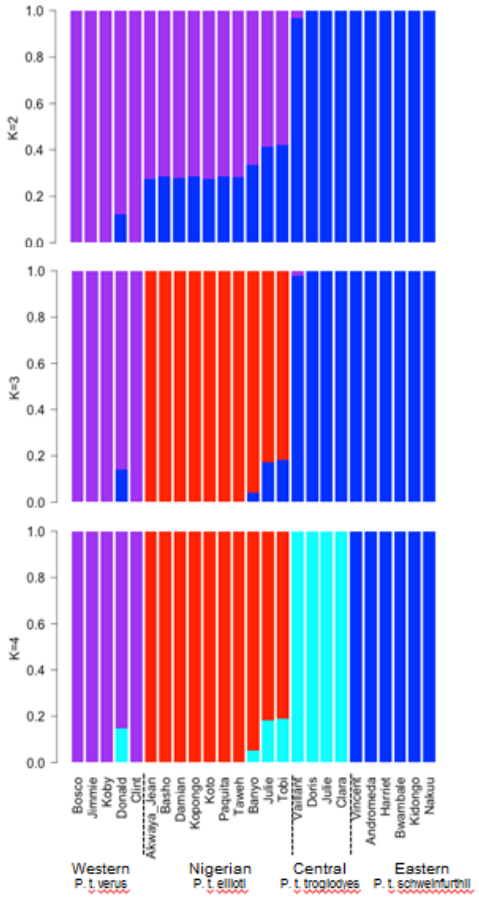
PCA of SNP genotypes identified three significant principal components. The first PC (20.8% of the variance, p = 0.0023) separates the Central and Eastern chimpanzees from the Nigeria-Cameroon and Western chimpanzees. The second PC (16.6% of the variance, p = 1.9e-07) separates the Nigeria-Cameroon and Western chimpanzees while the third PC (5.9% of the variance, p = 3.405e-06) separates the Central and Eastern chimpanzees (**Suppl. Figure 8.2.1**). Cross-validation analysis using ADMIXTURE offers support for three clusters (**Suppl. Figure 8.2.2**). Central and Eastern chimpanzees are separated at K = 4, a value not supported by the cross-validation metric. (**Suppl. Figure 8.2.3**)



Suppl. Figure 8.2.1 – *PC analysis of chimpanzees.*



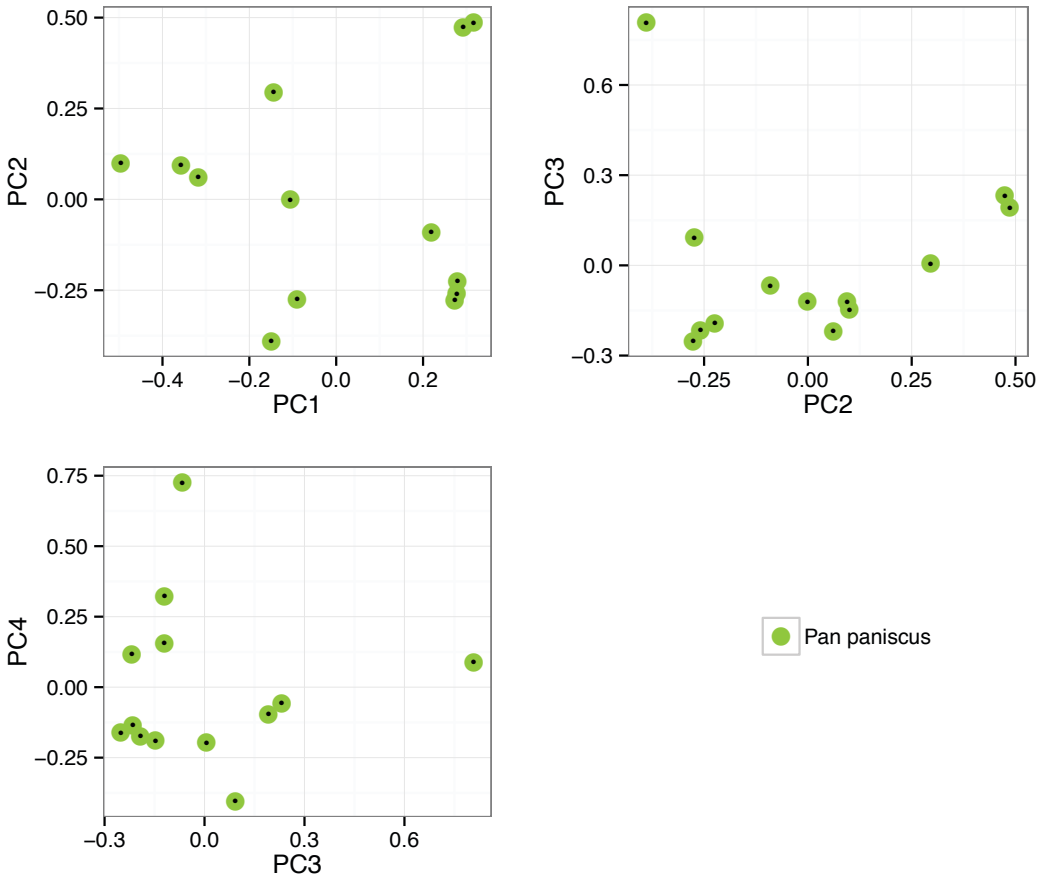
Suppl. Figure 8.2.2 – *ADMIXTURE cross-validation errors for chimpanzees.*



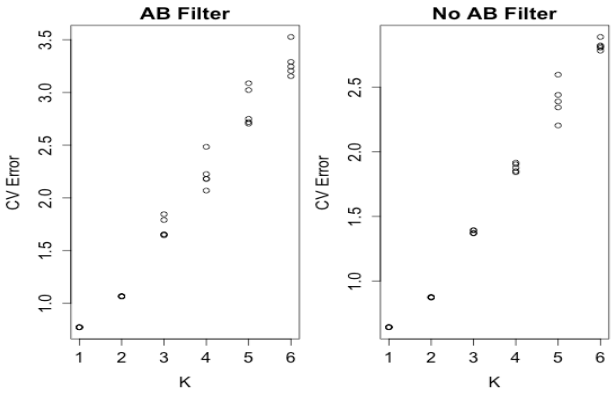
Suppl. Figure 8.2.3 – *ADMIXTURE cluster-membership for chimpanzees. The assignments at K = 4 are not supported by the cross-validation metric.*

Bonobo

PCA of SNP genotypes for the bonobos identifies a single significant component (14.3% of the variance p = 0.000393) along which the 13 bonobos are arrayed (**Suppl. Figure 8.2.4**). The second PC is not significant (p = 0.09). ADMIXTURE cross-validation indicates that K = 1 has the best support (**Suppl. Figure 8.2.5**).



Suppl. Figure 8.2.4 – *PCA of bonobo genotypes.*

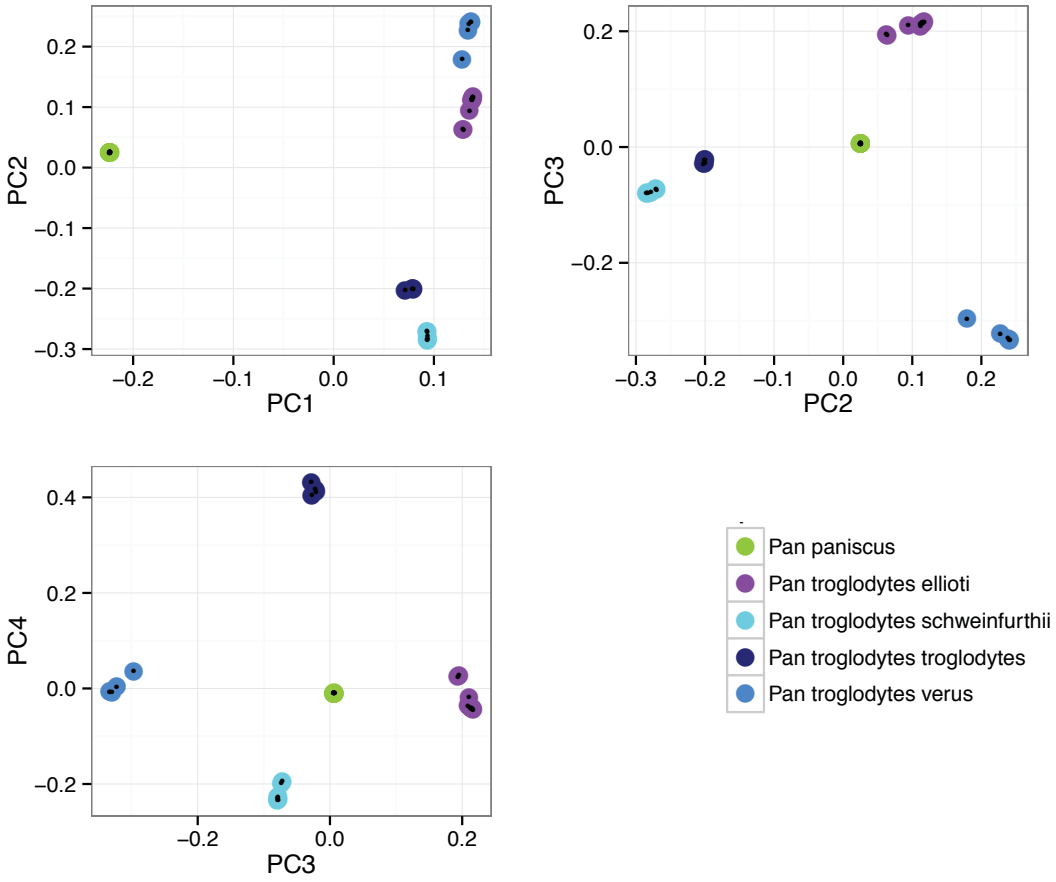


Suppl. Figure 8.2.5 – *ADMIXTURE cross-validation errors for bonobos.*

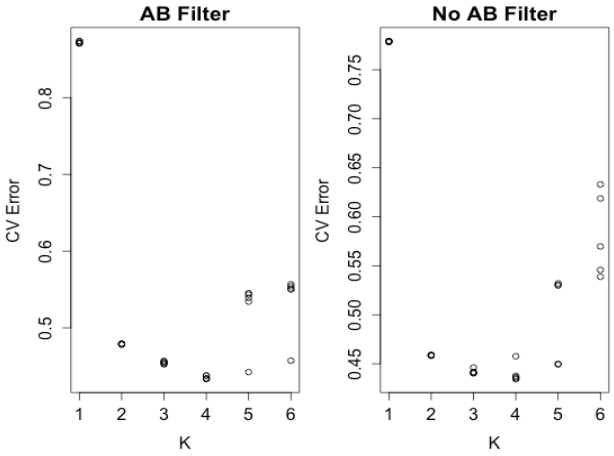
Chimpanzee and Bonobo

We also performed a joint analysis of the combined chimpanzee and bonobo data. In this sample set, we find that the first four PCs are significant (p = 0.00142, p = 7.93e-06, p = 1.67e-5,

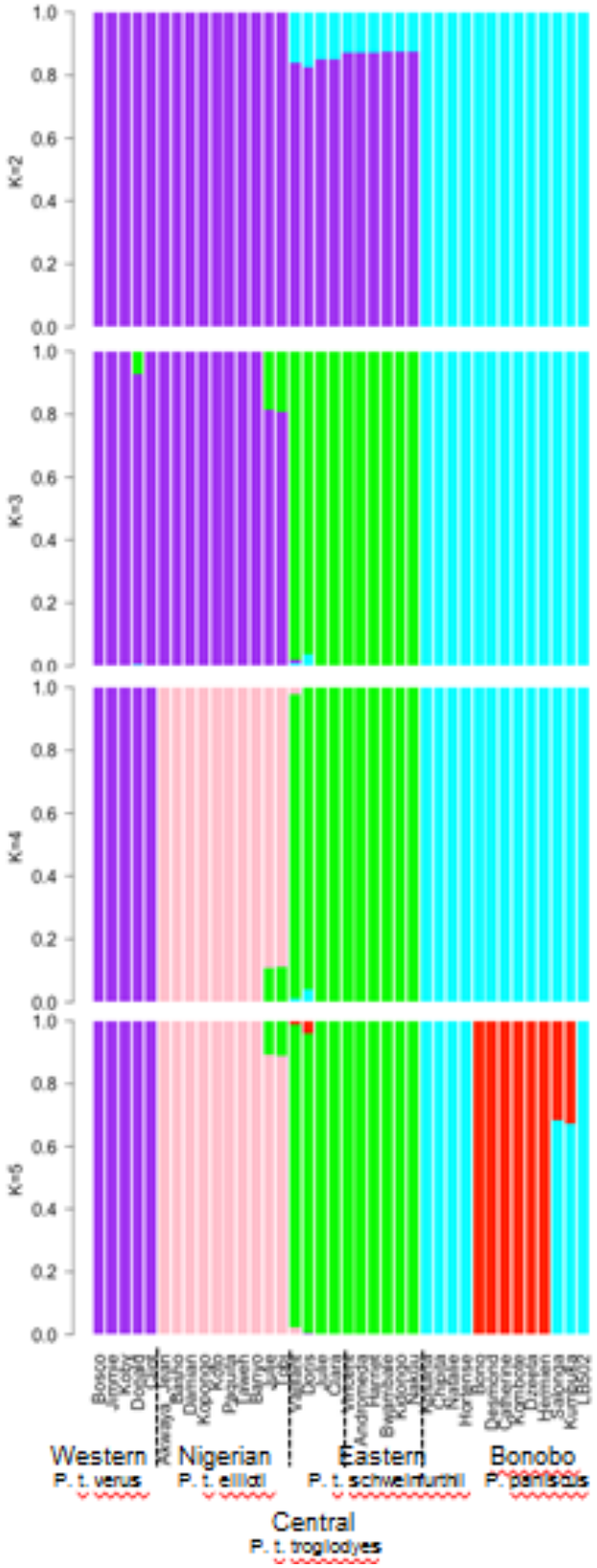
and p = 0.000169). PC1 accounts for 43.2% of the total variance and separates bonobos from chimpanzees while the higher PCs separate out the chimpanzees (**Suppl. Figure 8.2.6**). ADMIXTURE cross-validation shows the lowest assignment error at K = 4 (**Suppl. Figure 8.2.7**). At K = 2, chimpanzees and bonobos are separated, with the Central and Eastern chimpanzees showing ~12.5% component of shared ancestry with bonobos. At higher K, the Central and Eastern chimpanzees pull out as their own component. Interestingly, K = 5, which has a worse cross-validation error than K = 4, separates bonobos into two groups rather than dividing the Central and Eastern chimpanzee populations. (**Suppl. Figure 8.2.8**)



Suppl. Figure 8.2.6 – *PC analysis of combined chimpanzee and bonobo data.*



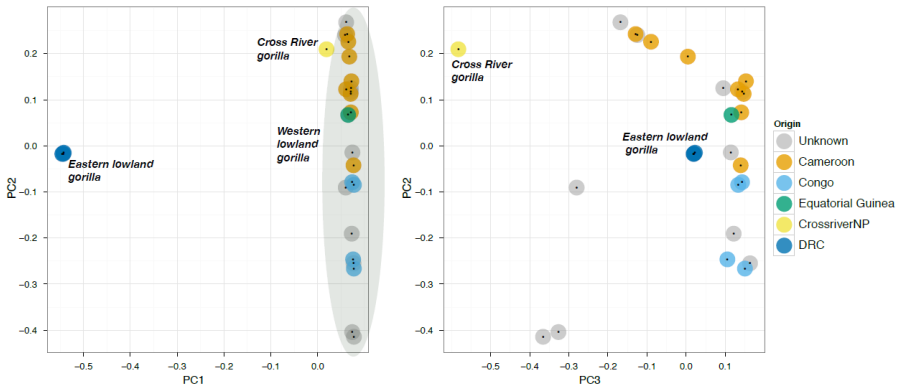
Suppl. Figure 8.2.7 – *Admixture cross-validation errors for combined chimpanzee and bonobo.*

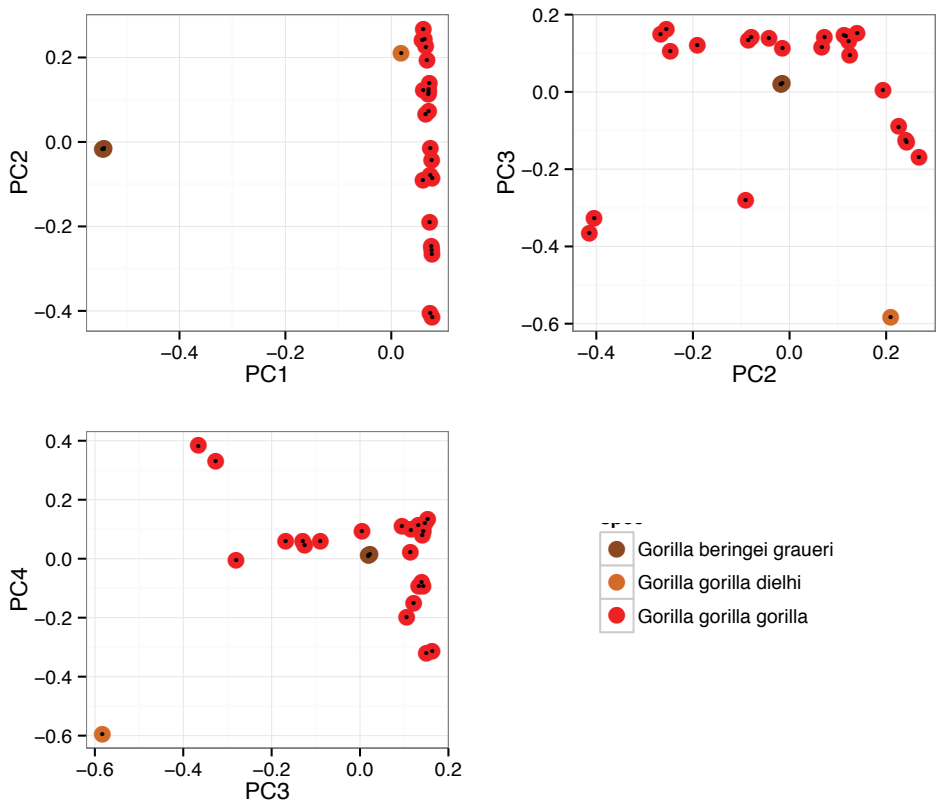

K=2
K=3
K=4
K=5
1.0
0.8
0.6
0.4
0.2
0.0
Bosco
Jimmie
Koby
Donald
Jean
Western
P. t. verus
Nigerian
P. t. ellioti
Central
P. t. troglodytes
Eastern
P. t. schweinfurthii
Bonobo

Suppl. Figure 8.2.8 – *Admixture cluster-membership for combined chimpanzees and bonobos. The assignments at K = 5 are not supported by the cross-validation metric.*

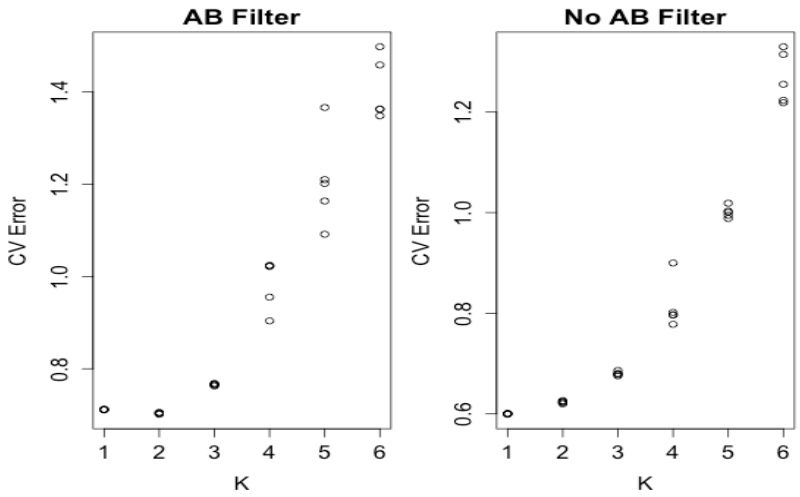
Gorilla

Analysis of the gorilla genotypes identifies two PCs that are significant. The first component (13.5% of the variance, p = 8.45e-12) separates the Eastern lowland from the Western lowland gorillas (**Suppl. Figure 8.2.9**). The Eastern lowland and the Cross River gorilla are arrayed out along the second component (6.6% of the variance, p = 2.64e-07). Cross-validation indicates that K = 2 has the lowest error rate, although we note that without the AB filter, cross-validation suggests that K = 1 is the best fit. At K = 2, Eastern and Western lowland gorillas are separated, with the single Cross River sample showing some evidence of a minor component shared with the Eastern gorillas (**Suppl. Figure 8.2.10**), which can also be inferred by the position of this sample along PC1 (**Suppl. Figure 8.2.9**).

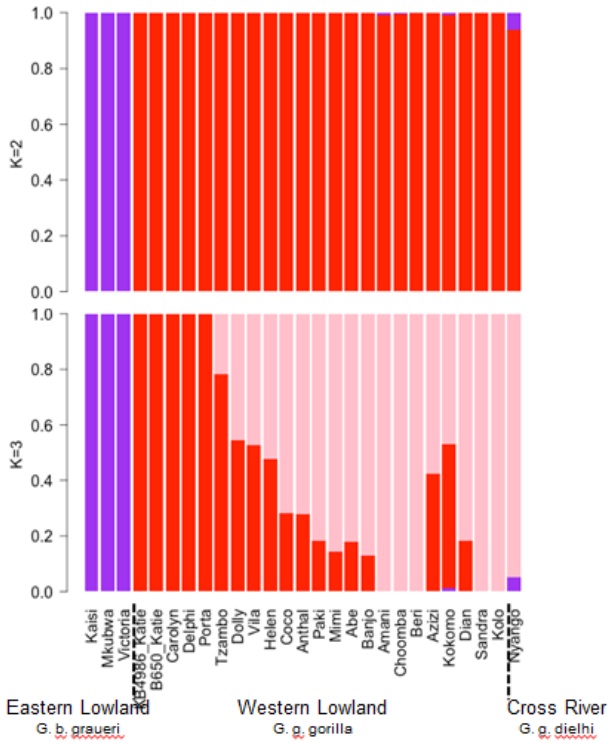




Suppl. Figure 8.2.9 – *PCA of gorillas showing the origin of each sample. Notice that PC2 tends to separate samples according to the origin of gorillas.*



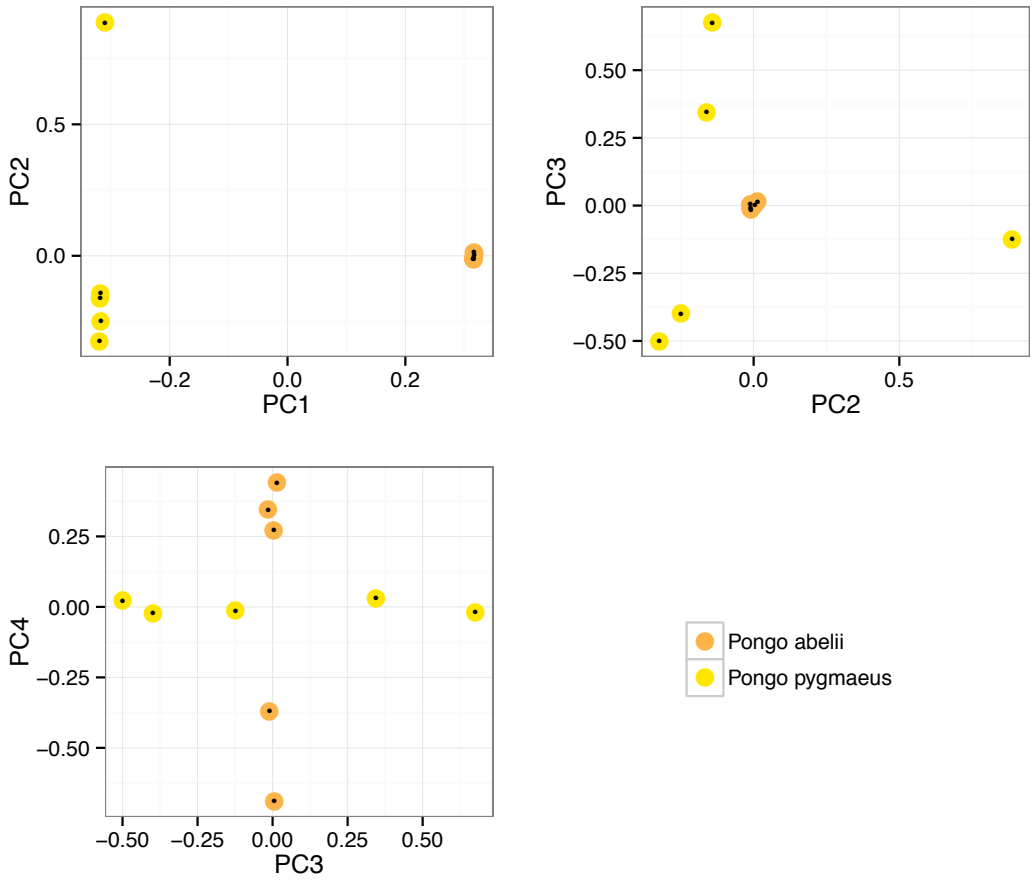
Suppl. Figure 8.2.10 – *Cross-validation results for gorillas.*



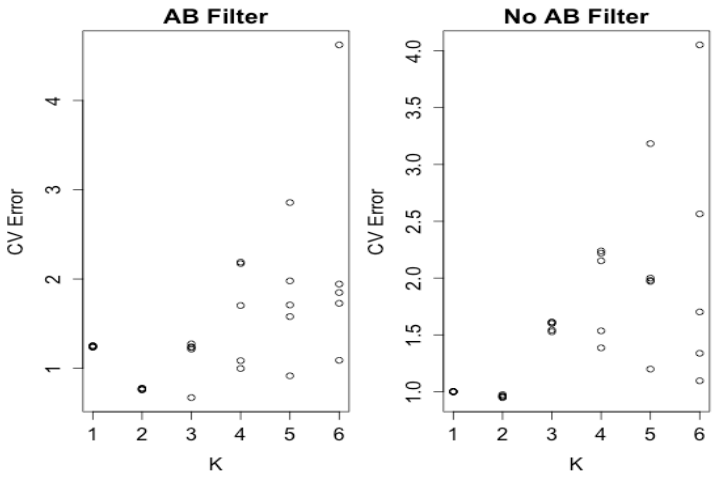
Suppl. Figure 8.2.11 – *ADMIXTURE cluster-membership for gorillas. The cross-validation metric does not support the selection of K = 3.*

Orangutan

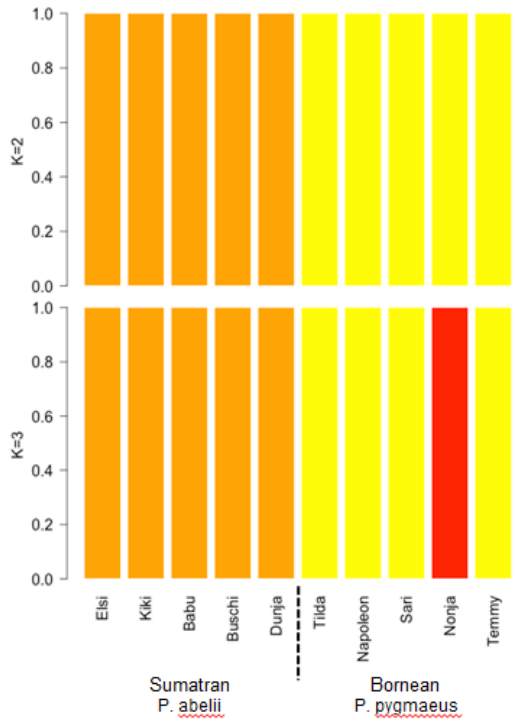
For the orangutans, only the first PC is significant. It accounts for 72% of the total variance and separates the Sumatran and Bornean samples. Most ADMIXTURE runs show K = 2 as having the lowest cross-validation error (**Suppl. Figure 8.2.12**), but a single run supports K = 3 with a single Bornean individual pulling out as a separate component (**Suppl. Figures 8.2.13 and 8.2.14**).



Suppl. Figure 8.2.12 – *PC analysis of orangutans.*



Suppl. Figure 8.2.13 – *Cross-validation results for orangutans.*



Suppl. Figure 8.2.14 – *ADMIXTURE cluster-memberships for orangutans. The cross-validation metric supports K = 3 in only one of five runs.*

8.3. Admixture deconvolution

Jeffrey M. Kidd, Joanna L. Kelley, Heng Li

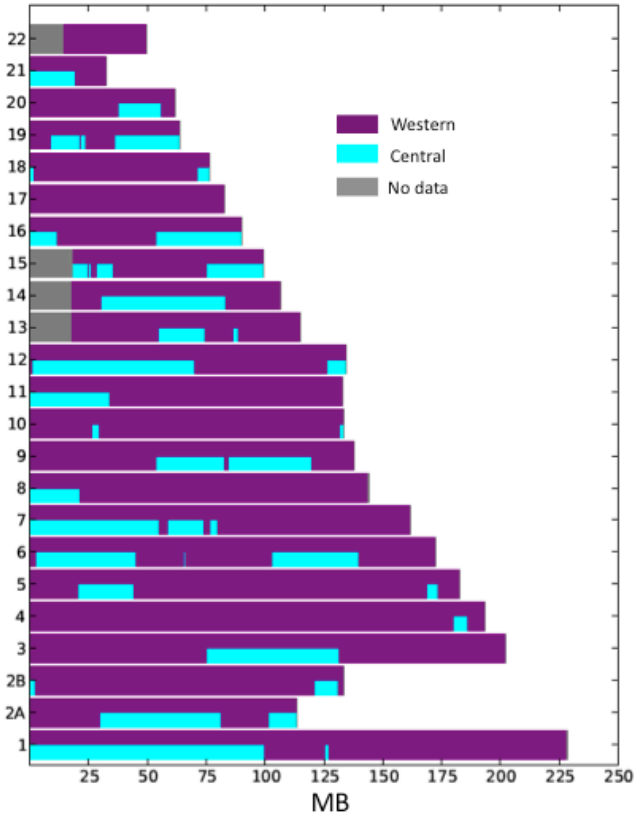
Analysis with frappe and ADMIXTURE indicated that Donald is a hybrid with ancestry from Western and Central chimpanzees and identified three Nigeria-Cameroon chimpanzees (Banyo, Julie, and Tobi) as having an unusual ancestry component. To investigate this further, we performed admixture deconvolution to infer the ancestral origin of each position along the genome of these samples. We based this analysis on the autosomal SNP calls identified based on mapping to the chimpanzee reference genome and used HAPMIX v2[23] along with the genetic map previously inferred for Western chimpanzees[24]. We ran HAPMIX in diploid mode, with the mutation parameter increased by a factor of 10 from the default value, as recommended for use on resequencing data instead of genotypes from SNP arrays.

For Donald, as source populations we used haplotypes from the four Western and four Central chimpanzees. We filtered the output from HAPMIX to require that tracts be at least 1 cM long and contain 1,000 SNPs in order for an ancestry switch to occur. Overall, we assign 15.5% of the genome to Central ancestry, and note that one complete set of chromosome homologs is inferred to be entirely Western in origin.

Min 1,000 SNPs and 1 cM			
Ancestry	**Tracts**	**Bp**	**FractionBp**
Central (P. t. troglodytes)	41	863,907,777	15.5%
Western (P.t. verus)	71	4,694,834,609	84.5%

Suppl. Table 8.3.1 – *Local ancestry assignments for Donald based on HAPMIX.*

We attempted the same analysis for Banyo, Julie, and Tobi using as reference panels the seven remaining Nigeria-Cameroon chimpanzees and either the four Central chimpanzees alone or a combined set of 10 Central and Eastern chimpanzees. We were unable to identify ancestry tracks, suggesting that these three samples are not the result of a recent admixture with Central chimpanzees.



Suppl. Figure 8.3.1 – *Ancestry painting for Donald.*

Sample U	Sample X	Sample Y	F3(U;X,Y)
ptv-Donald	ptt-Doris	ptv-Clint	-38.69
pte-Julie	ptt-Doris	pte-Koto	5.14
pte-Tobi	ptt-Doris	pte-Koto	21.89
pte-Banyo	ptt-Doris	pte-Koto	58.77

Note: If sample U is from a recent admixture of X and Y, F3(U;X,Y) significantly below zero.

Suppl. Figure 8.3.2 – *Results from the F3 statistic (Patterson et al. 2012).*

8.4. AIMs

Javier Prado-Martinez, Gabriel Santpere, Peter H. Sudmant, Jessica Hernandez-Rodriguez, Belen Lorente, Irene Hernado-Herraez, Arcadi Navarro, Evan E. Eichler, Tomas Marques-Bonet

Ancestry informative markers (AIMs) are variants that exhibit large allele-frequency differences between populations. These variants can accurately predict the population ancestry even when using a small subset of markers and can be extremely useful in conservation genetics and breeding programs of great apes. This resource may be a starting point towards the study of population genetics at a genome-wide level and from it we have retrieved the variants with a major degree of differentiation between the populations within each genus. Given the small sample sizes we could obtain from specific populations, we performed further analysis on whether these predicted AIMs are informative.

Expectation of AIMs FDR

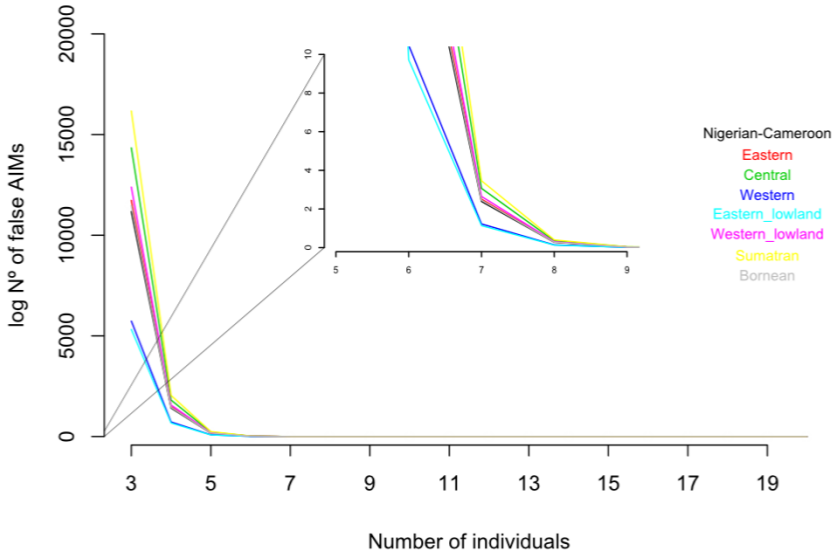
To predict the fixation of our AIMs, they were defined as fixed alleles in a particular species, while all other species being compared carry the other allele. That accounts for divergence between species but, given a low sample size, a number of segregating sites may mimic fixed differences. We estimated how many SNPs could be expected to resemble fixed differences with different sample sizes and diversity values. Given a segregating site (derived alleles), we calculated the probability of extracting one homozygous individual, considering Hardy-Weinberg equilibrium, from a neutral unfolded site frequency spectrum of 46 alleles, the maximum sample size used in this study. We use this probability to calculate the binomial probability of finding all sampled individuals of a given species homozygous for the derived allele in this site. Finally, the number of expected false AIMs in sampled individuals from one species is the product of all segregating sites identified in these individuals by the calculated probability of finding one false AIM.

Species	Sample size	Prob(All-HZ)	SS	AIMs Observed*	AIMs Expected	% false AIMs
Nigeria-Cameroon	10	2.9E-10	12,605,585	1,941	0	0
Eastern	6	1.9E-06	11,264,879	1,117	21	2
Central	**4**	**1.5E-04**	**11,820,858**	427	1,814	425
Western	4	1.5E-04	4,729,933	136,061	726	1
Eastern lowland	3	1.4E-03	3,866,117	278,190	5,330	2
Western lowland	23	1.2E-22	17,314,403	3,009	0	0
Sumatran	5	1.7E-05	14,543,573	446,133	248	0
Bornean	5	1.7E-05	10,321,213	733,535	176	0

Suppl. Table 8.4.1 – *Expected proportion of false AIMs as the product of the probability of extracting all individuals homozygous for one allele (Prob(All-HZ)) (for a given species sample size) by the number of segregating sites. *Derived alleles oriented with hg18.*

We found that for species with a very low sample size but higher diversity, such as Central chimpanzees, the number of expected AIMs, under neutrality assumptions, surpasses the number AIMs reported (**Suppl. Table 8.4.1.**).

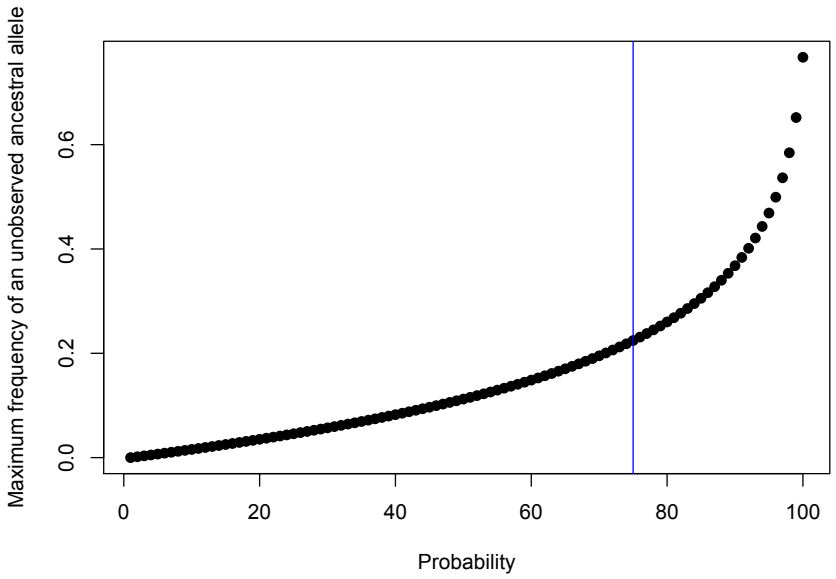
We have also projected the number of expected false AIMs sampling a different number of individuals for each species, scaling the number of segregating sites with the harmonic proportion. For all species, by sampling more than seven individuals, again considering a neutral site frequency spectra (SFS), no false AIMs are expected (**Suppl. Figure 8.4.1.**).



Suppl. Figure 8.4.1 – *Expected proportion of false AIMs as the product of the probability of extracting all individuals homozygous for one allele (Prob(All-HZ)) (for a given species sample size) by the number of segregating sites. *Derived alleles oriented with hg18.*

Given a number of sampled alleles, all bearing the derived allele, it is also possible to estimate which is the maximum frequency in the population that could achieve the ancestral allele to be unobserved in our sampling. Considering a 95% probability, the highest possible allele frequency (F) for an unobserved ancestral allele could be estimated as $1\text{-}e^{\text{-}Fn} = 0.95$, according to the Poisson distribution. With our minimum sample size (n) of three individuals (6 chromosomes), then F≈50%. If we consider a probability of 99%, F grows to ≈77%. That means a derived allele could be in a frequency in the population as low as 23% and still create false AIMs if we only sample three individuals. However, it also means that 95% of the time, the allele frequency of the derived allele in the population for these false AIMs is higher than 50%.

In fact, in the 75% percentile, the derived allele frequency increases to 77.4% (**Suppl. Figure 8.4.2**), which would not invalidate these segregating sites completely to be used as AIMs, especially if we considered combinations of them.



Suppl. Figure 8.4.2 – *Quantile distribution of the maximum frequency that can be achieved by a possible unobserved ancestral allele at site with an apparent AIM. The blue line represents the 75% quantile.*

AIM validation

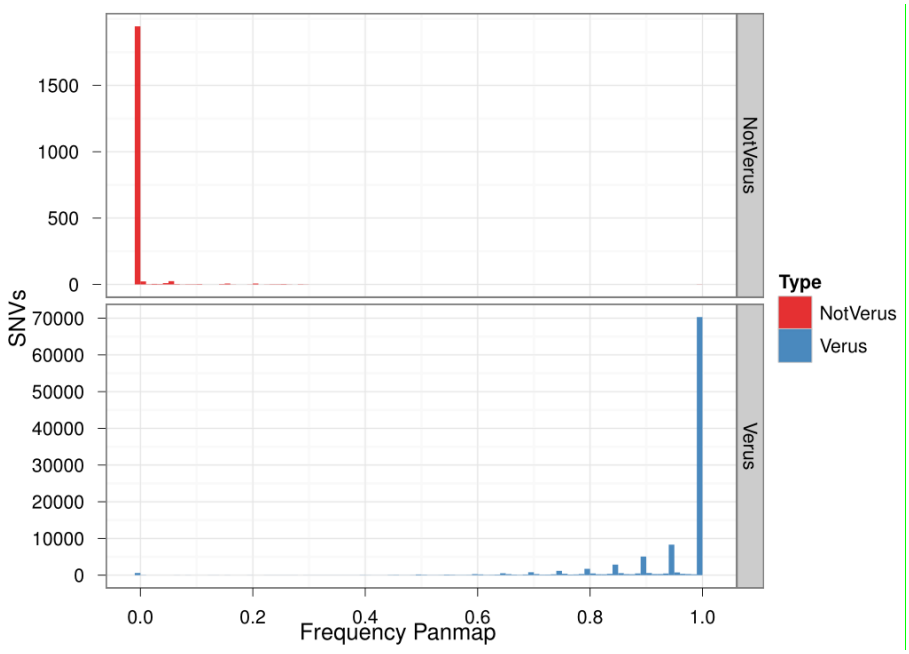
To complement this approach, we have also genotyped a set of 22 AIMs (~6 from each subspecies) in seven new individuals of known ancestry (1 *P.t.troglodytes*, 3 *P.t.schweinfurthii* and 3 *P.t.verus*) and a wild-born individual with unknown origin, using Sanger capillary sequencing. We genotyped 171 positions (5 sequences failed), 44 of which were tested for AIMs specific to the group and 127 were used as a negative controls (AIMs not specific). 96.9% of the negative controls have been correctly assigned (123/127). All but four sequences confirmed the presence of these alleles among these populations but they are mostly polymorphic in *P.t.schweinfurthii* and *P.t.troglodytes* due to the larger variation in these subspecies. Most of the specific AIMs are validated in *P.t.verus* and were determined as fixed variants in this group. As expected, the power of the AIM approach increased with the combination of them. Just as with 22 AIMs, all the individuals were correctly classified to their subspecies and we were able to determine the sample with unknown origin as a *P.t.schweinfurthii* (**Suppl. Table 8.4.2**).

			Allele		TROGLODYTES	SCHWEINFURTHII			VERUS			Unknown
			REF	AIM	**Noemie**	**Jac**	**Kina**	**Washu**	**Alice**	**Benita**	**Mike**	**Bihati**
AIMs	VERUS	chr1:156246124	G	A	-	-	-	+/-	+	+	+	+/-
		chr2:83764348	T	A	-	-	-	-	+	+	+	-
		chr12:78540784	G	A	-	-	-	-	+	+	+	-
		chr14:41791060	T	A		-	-	-	+	+	+	-
		chr13:35136820	T	A	-	-	-	-	+	+	+	-
		chr3:183884588	C	T	-	-	-	-	+	+	+	-
	SCHWEINFURTHII	chr6:118685644	A	G	-	+	+	+	-	-	-	+
		chr10:93452416	T	C	-	+	+/-	+/-	-	-	-	+
		chr1:183152149	T	A		+	+/-	-	-	-	-	+
		chr10:93382776	G	A	-	+	+	+/-	-	-	-	+
		chr6:73412868	A	G	-	+/-	+/-	-	-	-	-	+/-
	ELLIOTI	chr12:122387054	A	G	-	-	-	-	-	-	-	-
		chr8:111236243	C	G	-	-	-	-	-	-	-	-
		chr12:54999690	G	A	-	-	+/-	-	-	-	-	-
		chr1:179913228	G	C	-	-	-	-	-	-	-	-
		chr2:108857399	T	C	-		-	-	-	-	-	
	TROGLODYTES	chr8:139513779	G	A	-	-	-	-	-	-	-	-
		chr11:114695783	A	T	+	-	-		-	-	-	-
		chr9:271906	A	G	+	-	-	-	-	-	-	-
		chr3:34545483	G	T	-	-	-	-	-	+/-	-	-
		chr22:35007755	T	C	+/-	-	-	-	-	-	-	-
		chr14:94709629	A	T	+/-	-	-	-	-	-	-	-
Distances	Distance P.t.verus 22 AIMS				16	21	20	15	0	1	0	20
	Distance P.t.schweinfurthii 22 AIMS				14	1	4	7	22	23	22	2
	Distance P.t.ellioti 22 AIMS				16	17	16	15	22	23	22	18
	Distance P.t.troglodytes 22 AIMS				6	21	20	15	24	23	24	22

Suppl. Table 8.4.2 – *Sanger capillary genotyping of 22 predicted AIMs in eight new individuals. The global distance to the combination of AIMs for each subspecies (bottom) allows us to unambiguously classify each individual to the proper population. The allele REF corresponds to the allele for which we are not predicting to be at high frequency in the population we are testing, i.e., a "+" in the table always correspond to the AIM column. Blank spaces denote that the sequencing did not work.*

Whole-genome validation

Finally, we have also computationally genotyped for our set of AIMs the 10 chimpanzee genomes analyzed in a previous work (Auton et al. Science 2012). All the individuals but one are wild born and they belong to *Pan troglodytes verus* subspecies. Most of the non-verus AIMs (~95%) are not present in these individu and the remaining 5% are present at very low frequencies. The *verus*-specific AIM are mostly fixed (~70% of them) and the others are found at high frequencies, i.e. ~90% of these AIMs are found with an allele frequency (AF) higher than 0.8 in th samples. (**Suppl. Table 8.4.3**).



Suppl. Table 8.4.3 – *Frequency of P.t.verus AIMs and non-P.t.verus AIMs in the chimpanzee individuals from (Auton et al Science 2012). Almost all the non-verus AI are absent or at low frequency (top panel) whereas the verus-specific AIMs (bottom panel) are fixed (70%) or at high frequency.*

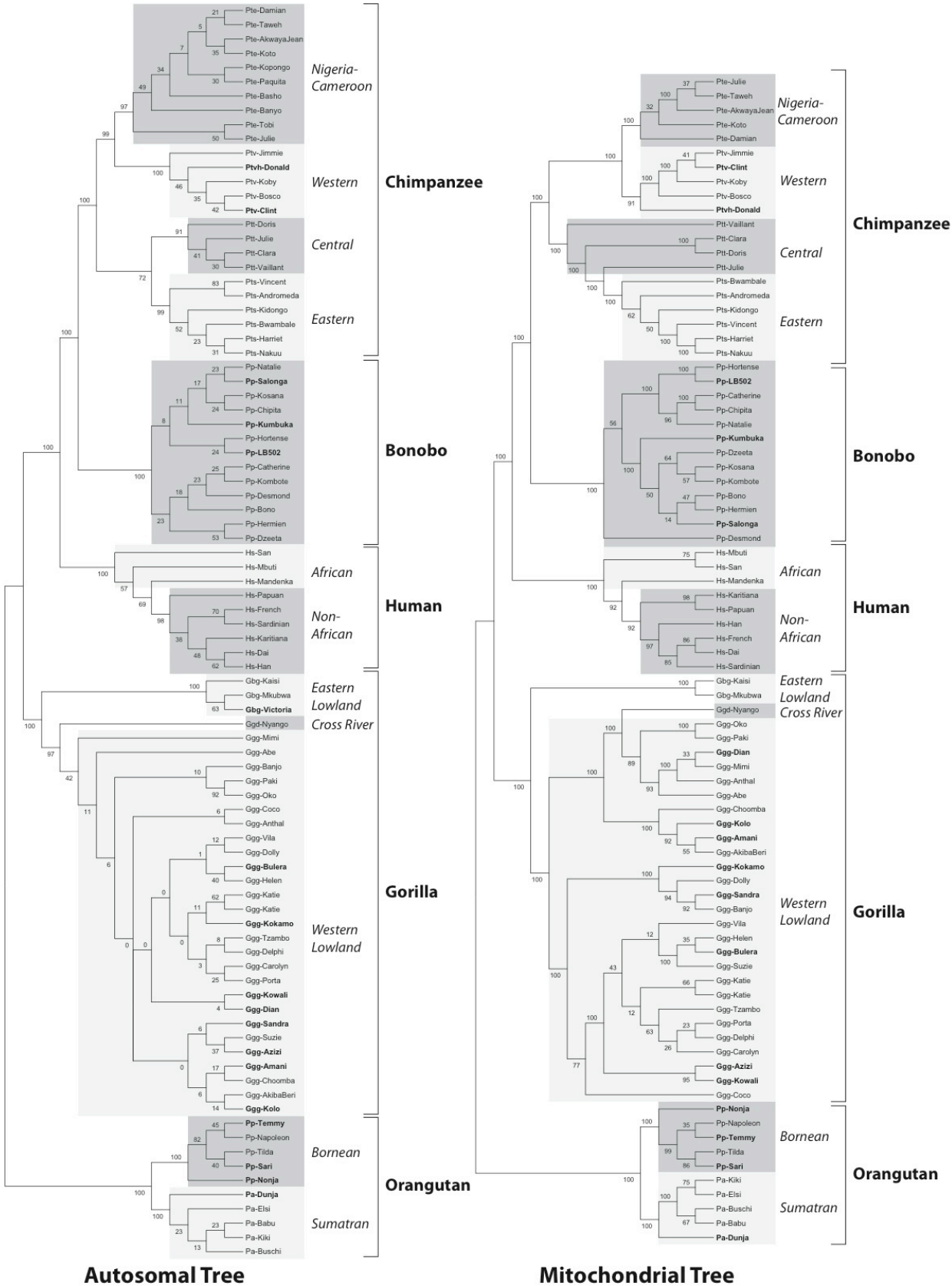
Nevertheless, overfitting is definitely an important consideration in light of our small sample size and this effect is most pronounced for subspecies with the greatest diversity. While our AIMs provide a starting point towards further characterization of unknown samples (especially for Nigeria-Cameroon and Western chimpanzees), we have minimized discussion of conservation and the us of these markers until more experimental validation emerges.

8.5. Phylogeny

Peter H. Sudmant, Belen Lorente-Galdos, Evan E. Eichler

We created an autosomal phylogeny based on a consensus of neighbor-joining trees (n=560) constructed from non-overlapping 5 Mbp blocks across the genome, whereas mitochondrial phylogeny was assessed using a maximum likelihood method from *de novo* assembled mitochondrial genomes (Section 7). The autosomal tree shows separate monophyletic groupings for each species/subspecies designation (**Suppl. Figure 8.5.1**) and support a split of extant chimpanzees into two groups. Nigeria-Cameroon and Western chimpanzees form a monophyletic clade (>97% of all autosomal trees). Central and Eastern chimpanzees form a second group (72% of the genomic trees separate these subspecies). The topology of mitochondrial phylogeny supports this bipartite division with the exception that Eastern chimpanzee and Central chimpanzees are grouped into a single clade. This difference may be consistent with the evolutionary stochasticity of a single locus and or sex-specific processes associated with the maternal transmission of the mitochondria. The single Cross River gorilla genome also shows some evidence of genetic differentiation from the Western lowland gorillas—although only 42% of the autosomal trees suggest it as a potential outgroup of all Western lowland gorillas and the mtDNA tree groups it with other Western gorillas. There is evidence of additional Western lowland gorilla substructure based on the tree topology and additional analyses (see PCA and PSMC). As expected, orangutan species cluster into two distinct monophyletic clades with complete support.

Suppl. Figure 8.5.1 – Phylogeny. *(Below) Cladograms constructed from autosomal and mitochondrial sequence illustrate the relationships between the individuals sequenced in this study. The autosomal tree is constructed by neighbor-joining trees on genetic distance estimates of 5 Mbp autosomal segments. Confidences represent the proportion of autosomal segments supporting the topology. The mitochondrial tree is derived by maximum likelihood from assembled mitochondrial haplotypes. Bootstrap confidence values are displayed for clades with >50% confidence. In bold we highlight captive individuals.*


Pte-Damian
Pte-Taweh
Pte-AkwayaJean
Pte-Koto
Pte-Kopongo
Pte-Paquita
Pte-Basho
Pte-Banyo
Pte-Tobi
Pte-Julie
Nigeria-Cameroon
Ptv-Jimmie
Ptvh-Donald
Ptv-Koby
Ptv-Bosco
Ptv-Clint
Western
Ptt-Doris
Ptt-Julie
Ptt-Clara
Ptt-Vaillant
Central
Pts-Vincent
Pts-Andromeda
Pts-Kidongo
Pts-Bwambale
Pts-Harriet
Pts-Nakuu
Eastern
Chimpanzee
Pp-Natalie
Pp-Salonga
Pp-Kosana
Pp-Chipita
Pp-Kumbuka
Pp-Hortense
Pp-LB502
Pp-Catherine
Pp-Kombote
Pp-Desmond
Pp-Bono
Pp-Hermien
Pp-Dzeeta
Bonobo
Hs-San
Hs-Mbuti
Hs-Mandenka
African
Hs-Papuan
Hs-French
Hs-Sardinian
Hs-Karitiana
Hs-Dai
Hs-Han
Non-African
Human
Gbg-Kaisi
Gbg-Mkubwa
Gbg-Victoria
Eastern Lowland
Ggd-Nyango
Cross River
Ggg-Mimi
Ggg-Abe
Ggg-Banjo
Ggg-Paki
Ggg-Oko
Ggg-Coco
Ggg-Anthal
Ggg-Vila
Ggg-Dolly
Ggg-Bulera
Ggg-Helen
Ggg-Katie
Ggg-Katie
Ggg-Kokamo
Ggg-Tzambo
Ggg-Delphi
Ggg-Carolyn
Ggg-Porta
Ggg-Kowali
Ggg-Dian
Ggg-Sandra
Ggg-Suzie
Ggg-Azizi
Ggg-Amani
Ggg-Choomba
Ggg-AkibaBeri
Ggg-Kolo
Western Lowland
Gorilla
Pp-Temmy
Pp-Napoleon
Pp-Tilda
Pp-Sari
Pp-Nonja
Bornean
Pa-Dunja
Pa-Elsi
Pa-Babu
Pa-Kiki
Pa-Buschi
Sumatran
Orangutan
Autosomal Tree
Pte-Julie
Pte-Taweh
Pte-AkwayaJean
Pte-Koto
Pte-Damian
Nigeria-Cameroon
Ptv-Jimmie
Ptv-Clint
Ptv-Koby
Ptv-Bosco
Ptvh-Donald
Western
Ptt-Vaillant
Ptt-Clara
Ptt-Doris
Ptt-Julie
Central
Pts-Bwambale
Pts-Andromeda
Pts-Kidongo
Pts-Vincent
Pts-Harriet
Pts-Nakuu
Eastern
Chimpanzee
Pp-Hortense
Pp-LB502
Pp-Catherine
Pp-Chipita
Pp-Natalie
Pp-Kumbuka
Pp-Dzeeta
Pp-Kosana
Pp-Kombote
Pp-Bono
Pp-Hermien
Pp-Salonga
Pp-Desmond
Bonobo
Hs-Mbuti
Hs-San
Hs-Mandenka
African
Hs-Karitiana
Hs-Papuan
Hs-Han
Hs-French
Hs-Dai
Hs-Sardinian
Non-African
Human
Gbg-Kaisi
Gbg-Mkubwa
Ggd-Nyango
Eastern Lowland Cross River
Ggg-Oko
Ggg-Paki
Ggg-Dian
Ggg-Mimi
Ggg-Anthal
Ggg-Abe
Ggg-Choomba
Ggg-Kolo
Ggg-Amani
Ggg-AkibaBeri
Ggg-Kokamo
Ggg-Dolly
Ggg-Sandra
Ggg-Banjo
Ggg-Vila
Ggg-Helen
Ggg-Bulera
Ggg-Suzie
Ggg-Katie
Ggg-Katie
Ggg-Tzambo
Ggg-Porta
Ggg-Delphi
Ggg-Carolyn
Ggg-Azizi
Ggg-Kowali
Ggg-Coco
Western Lowland
Gorilla
Pp-Nonja
Pp-Napoleon
Pp-Temmy
Pp-Tilda
Pp-Sari
Bornean
Pa-Kiki
Pa-Elsi
Pa-Buschi
Pa-Babu
Pa-Dunja
Sumatran
Orangutan
Mitochondrial Tree

8.6. Cross River gorilla

From a phylogenetic perspective, the Cross River gorilla genome is located at the base of all Western lowland gorillas. 42% of the autosomal trees suggest it as a potential outgroup of all Western lowland gorillas. In contrast, the mtDNA tree positions the Cross River individual within the diversity of the other Western gorillas. From the point of view of PCA, all the Western lowland gorillas are distributed along a gradient on the PC2 axis. The Cross River gorilla is more similar to the Cameroon Western lowland gorillas. However, it is the third component that clearly separates the Cross River gorilla from most Western gorillas. However, using ancestry (STRUCTURE), a separate signal is not observed for the Cross River gorilla subspecies until we establish six clusters (K=6) suggesting that the signal that separates the Cross River gorilla from the Western lowland gorillas is not clearly distinguished considering the genetic diversity in Western lowland gorillas. From the point of view of population dynamics, we infer a smaller effective population size for the Cross River gorilla when compared with the Western lowland gorillas beginning ~100 kya. Evidence of historical gene flow between Eastern lowland and Cross River gorillas is observed according to the D-statistic, ABC and divergence time estimators. Finally, we find that the Cross River sample shows reduced heterozygosity as a result of long and thus recent runs of homozygosity events as corresponds from a very limited number of individuals (300 individuals (IUCN Red List of Threatened Species)).

Section 9: Inbreeding

Javier Prado-Martinez, Krishna Veeramah, Tomas Marques-Bonet

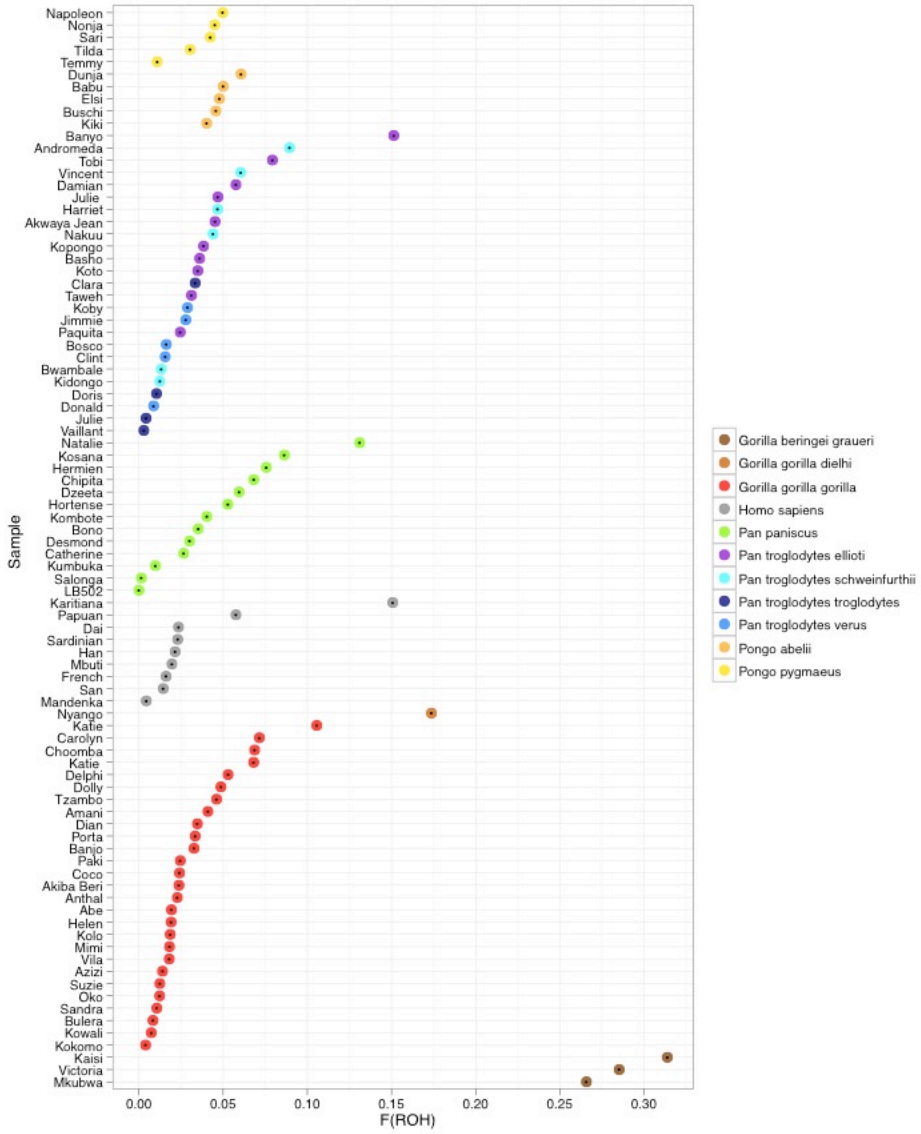
To determine the amount of inbreeding, we calculated the genome-wide heterozygosity in windows of 1 Mbp with 200 Kbp sliding windows. We detected an excess of windows with very low heterozygosity in the density plots (**Suppl. Figure 6.2**) pointing to some extent of inbreeding. In order to estimate the cutoff values for inbreeding coefficients, we calculated a different local minima for different species since this value is variable depending in the heterozygosity and divergence to the human genome:

- Human, bonobo and Western chimpanzee: 0.00015
- Eastern, Central and Nigeria-Cameroon chimpanzees: 0.00025
- Gorillas and Bornean orangutan: 0.0003
- Sumatran orangutan: 0.0004

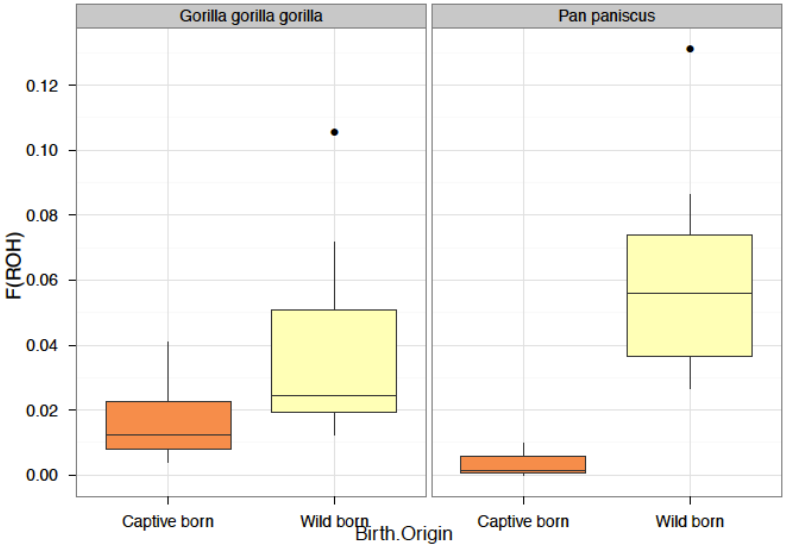
We then clustered together the neighboring regions to account for runs of homozygosity (ROH). From there we could estimate the percentage of the genome that is autozygous (F_{ROH}), which is a good measure of inbreeding as previously calculated[25]. We chose 1 Mbp as a threshold to consider a region as autozygous according to previous estimations[26], which identified that regions smaller than 0.5 Mbp are the result of background relatedness, while tracts larger than 1.6 Mbp are evidence of recent parental relatedness.

In general we observe a degree of background relatedness among all the groups studied with more recent inbreeding events in almost all the populations (**Suppl. Figure 9.1**). In humans around 2.5% of the genome shows a certain degree of autozygosity while the Karitiana and Papuan samples show a high degree of inbreeding due to the smaller population sizes among these groups. Bonobos and Nigeria-Cameroon chimpanzees show the highest amount of autozygosity among the *Pan* genus and can be explained by the patchy distribution and small population respectively. Lower levels of inbreeding are shown in Western and Central chimpanzees, but the small sample sizes in these populations may not reflect the actual trend of inbreeding. Among gorillas and bonobos, the captive-born have lower autozygosity than wild-born gorillas. We tested this trend with a Wilcoxon rank-sum test and the difference between the groups was significant in both species (p-values = 0.0067 and 0.0035, respectively) (**Suppl. Figure 9.2**). Moreover, the gorilla samples that have been geographically located in the Congo seem to have a higher amount of inbreeding than the rest of the gorillas. Eastern and Cross River gorillas show strikingly high levels of inbreeding as a possible result of population decline and habitat fragmentation[27]. Both orangutan populations show a moderate degree of inbreeding, comparable to those in Nigeria-Cameroon chimpanzees and

bonobos, suggesting that the habitat loss in the Borneo and Sumatra islands may be having an effect in the random mating among this species.



Suppl. Figure 9.1 – *Summary of inbreeding coefficients (F_{ROH}) clustered by subspecies and ordered by inbreeding rank.*



Suppl. Figure 9.2 – *Comparison of inbreeding between captive and wild individuals.*

Section 10: Loss-of-function variants

Javier Prado-Martinez, Peter H. Sudmant, Maika Malig, Carl Baker, Belen Lorente-Galdos, Marcos Fernandez-Callejo, Can Alkan, Evan E. Eichler, Arcadi Navarro, Tomas Marques-Bonet

Loss-of-function SNVs

To characterize mutations in the coding sequence of the human gene models, we annotated the variants with the ANNOVAR software[28] with the RefSeq human gene models. In order to assign the precise lineage where the mutations occurred, we clustered the species sharing these mutations and assigned mutations to the human lineage in case all the species but human carry the mutation and to the human-Pan branch in case the mutations are shared between gorilla and orangutan species.

In total we obtained 806 stop-gain/stop-loss mutations (**Suppl. Figure 10.6, Table S4**). We compared the extant studies of premature termination codons and found 91 were previously reported[5,29–32]. We compared our predictions with the EPO alignment between human-chimpanzee-gorilla-orangutan and found 63% (506) of the variants were supported by all the assemblies. We performed 151 capillary Sanger validations for which 150 were correctly predicted (99.3%) (**Table S2**).

Indel variants

Frameshift mutations may account for a large proportion of gene disruptions, doubling the number of premature termination codons (PTCs) compared to SNVs in previous studies[29]. Starting from the mappings to the human reference assembly (hg18) and with the reads previously realigned with GATK in a multi-sample fashion (see **Suppl. Section 2.1**), we used GATK Unified Genotyper to produce an initial set of indel candidates. Then we applied the following filters for indels:

- QD < 2.0, ReadPosRankSum < -20.0, FS > 200.0
- Variants overlapping segmental duplications and tandem repeats (TRF from UCSC).

We finally removed indels clustering within 10 bp to remove possible artifacts on problematic regions. We focused on indels in coding regions and performed quality controls on this subset. We first assessed the distribution of sizes, expecting triplet multiplicity as a result of purifying selection to preserve the reading frame dividing our variants in fixed and polymorphic to account for selective pressure (**Suppl. Figure 10.1**). Interestingly, if we divide the indels into fixed and segregating, we can observe how purifying selection is stronger in fixed indels, given that most of the variants are triplet multiple. See **Suppl. Section 3.5** for the validations on indels.

Gene large deletions

Larger lineage-specific deletions were identified by discordant read-pair analysis using the VariationHunter software[33]. The approach that followed was similar to the procedure used by Ventura et al[34]. Calling was performed on a per-individual basis and the resulting individual callsets were merged within species by 50% reciprocal overlap criteria. All calls were then genotyped by read depth to confirm the deletions were indeed fixed among the individuals assessed in this study. Deletion events were further confirmed by a custom designed array comparative genomic hybridization (aCGH). A total of 30 aCGH experiments were performed using all nonhuman primate species and subspecies against the human reference NA12878. In total 374/382 (97.9%) aCGH experiments successfully confirmed the lineage-specific deletion event. Excluding aCGH experiments with fewer than two probes, 100% of experiments successfully validated. We identified a total of 96 lineage-specific deletions partially or completely overlapping coding exons.

We thus screened all mutations leading to loss-of-function events using the human gene models (RefSeq) and classified them into their evolutionary context within the great ape lineage (**Suppl. Figure 10.6**). For the first time we can provide a complete picture of events leading to significant changes in genes at different branches. In total we detected 1,982 fixed loss-of-function events in 1,481 genes. (**Table S4).**

We then studied the position along the gene of these events. We divided the relative position of these mutations in bins of 5% of the human gene model and we plotted the amount of variants leading to frameshift/in-frame mutations (**Suppl. Figure 10.7, Suppl. Figures 10.2 and 10.3**). We detected an enrichment of disrupting events in the beginning and the end of the genes. These regions of the gene have been associated with a lower selective constraint given that the proportion of functional domains is smaller[29].

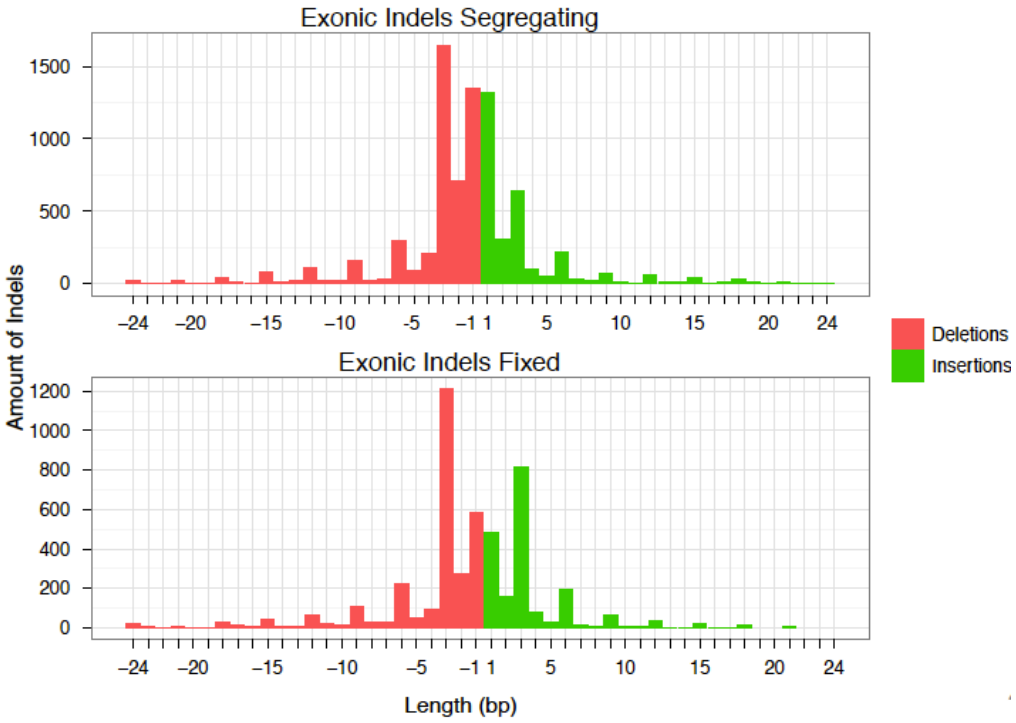
The genes we are reporting in **Table S4** may account for a significant number of important genes during great ape evolution and the gain/loss or modification of these proteins may have had a crucial impact during the phenotypic differentiation between the different lineages. Further work may be needed to perform functional characterization of the effect of these mutations, but this characterization of events is a step towards a better understanding of recent speciation of the human and great ape lineages.

Less is more hypothesis

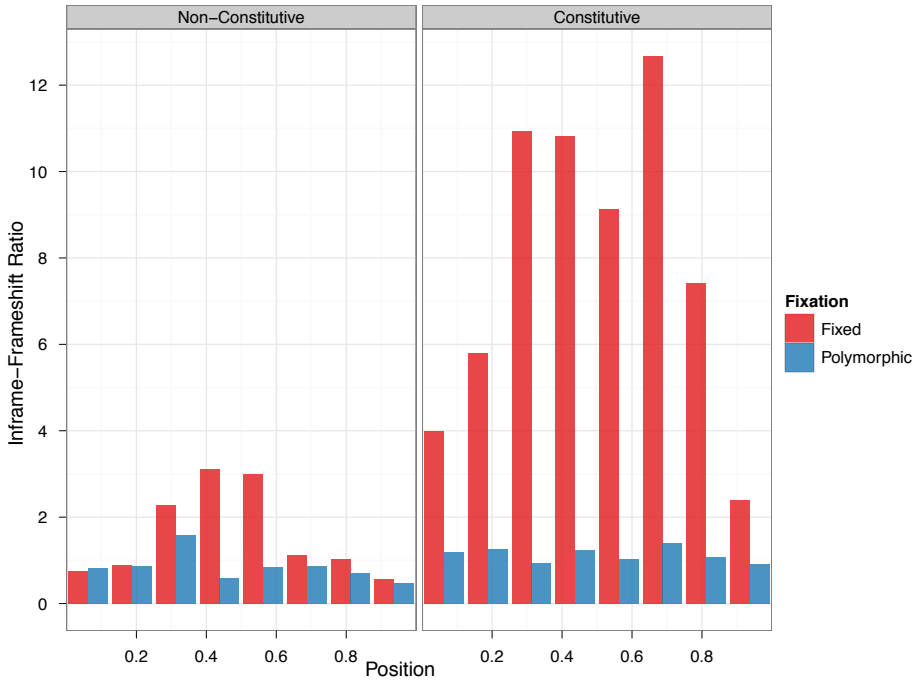
After the appearance of the less is more hypothesis[35], it has been long debated whether gene losses have an important role in the evolution. This has been hypothesized as being a very important engine of evolutionary change in the human lineage, based on several adaptations and phenotypes by using this mechanism[36].

For this reason we have tested this hypothesis along the hominid phylogeny. We studied the number of events detected in the different branches and compared them to the evolutionary distance as a function of the genetic divergence. **Suppl. Figure 10.4** shows the correlation between genetic distance and the amount of gene losses.

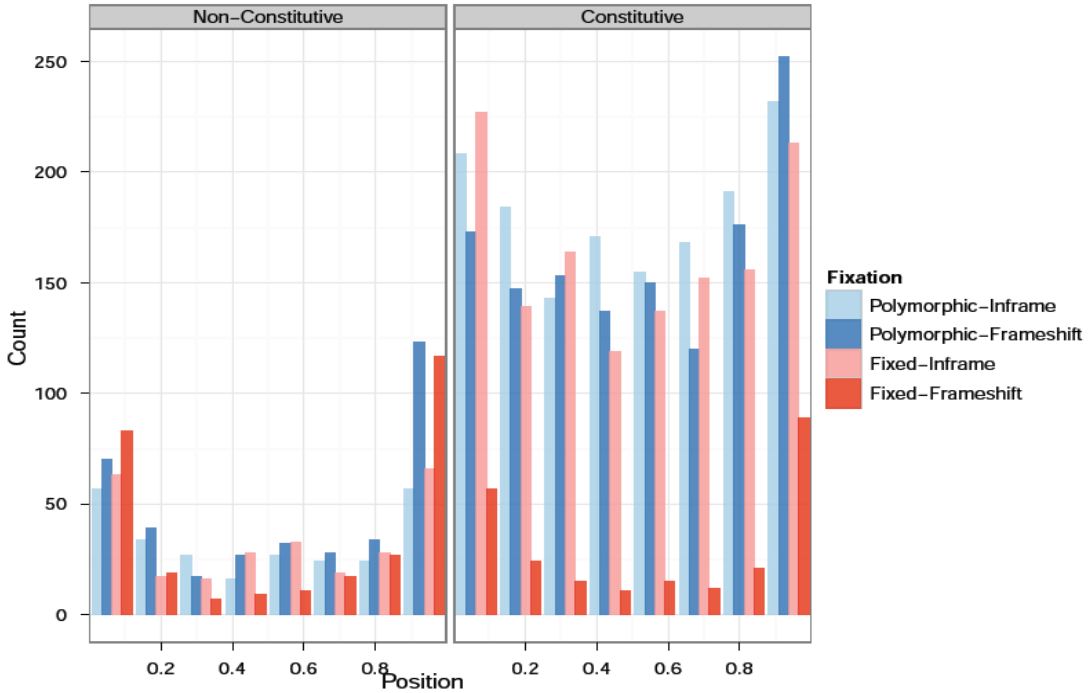
The main problem of this analysis is that we rely on the human gene models. This introduces a bias given that the pseudogenes that are specific to the human lineage are underrepresented and this can lead to an underestimation of the number of gene losses in the human lineage. We accounted for this including 67 human-specific pseudogenes from Wang et al[36]. We performed the same correlation accounting for the branches with more than 0.3% of divergence (Pongo-Homininae, Gorilla, Pan and Human branches), and we did not observe an excess of loss-of-function events during the human lineage (**Suppl. Figure 10.5**). We used a Maximum Likelihood Ratio test, to test whether the human branch has an excess of loss-of-function events or if all four lineages follow a single rate. We found that a single rate (39.7 losses per mut/Kbp) fits better these data.



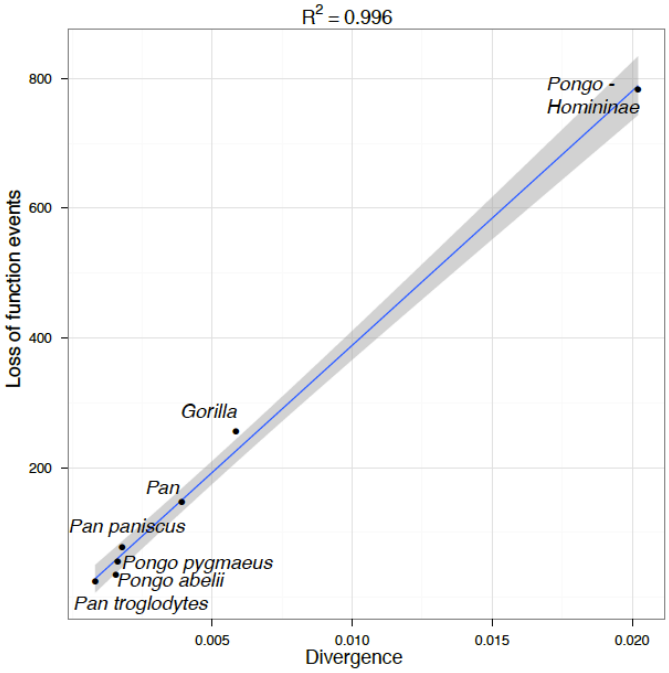
Suppl. Figure 10.1 – *Indel size distribution in coding regions. The effect of selection is stronger in fixed events; this is noticed by the larger proportion of events maintaining the reading frame compared to disruptive frameshift mutations.*



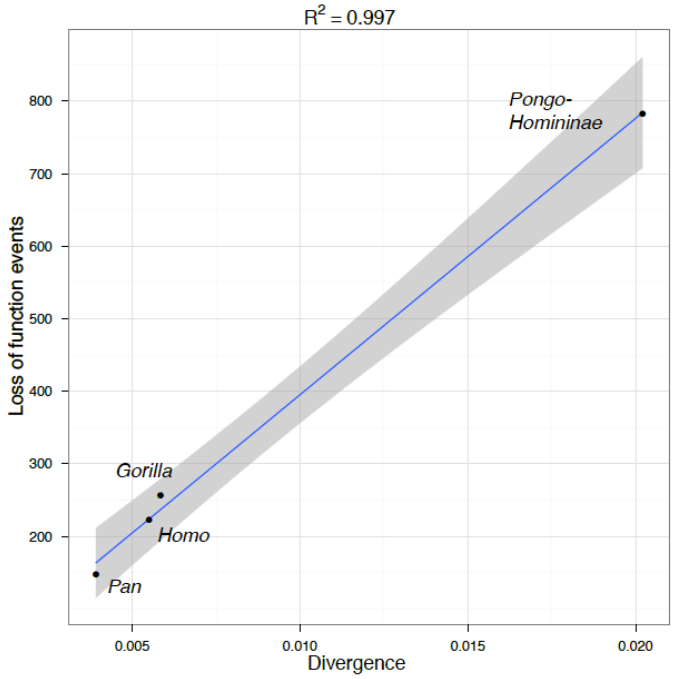
Suppl. Figure 10.2 – *Distribution of in-frame/frameshift ratio along the gene in 1:1 orthologous genes in primates. We consider constitutive the exons that appear in all the isoforms of the gene. Fixed variants are in red and polymorphic in blue.*



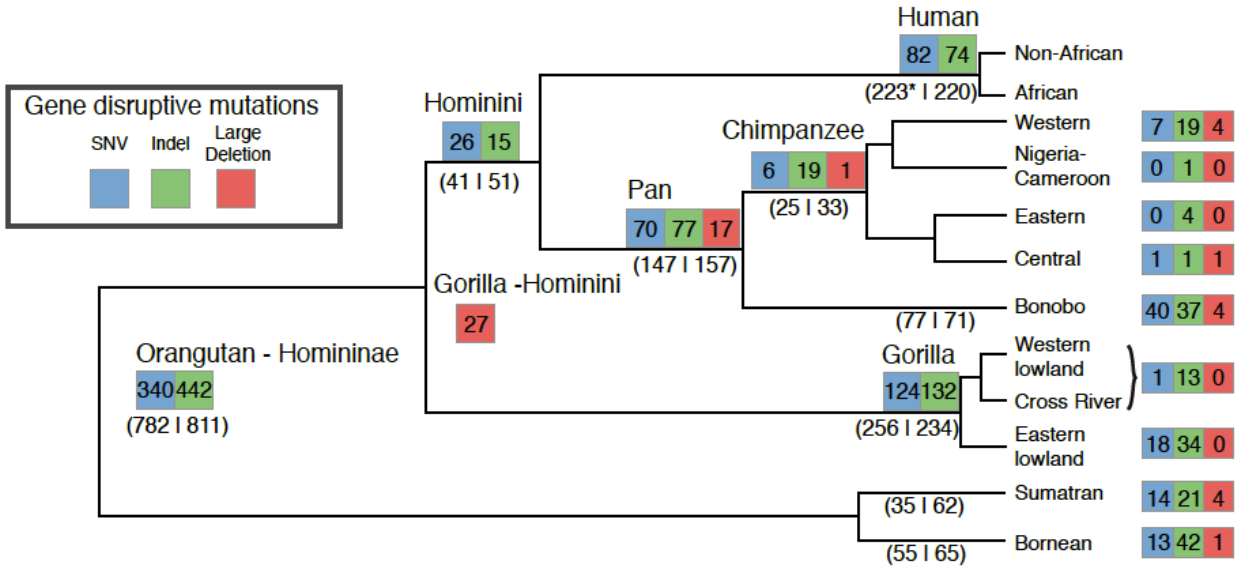
Suppl. Figure 10.3 – *Distribution of frameshift mutations across the gene positions. Both polymorphic and fixed mutations show an increase towards 5' and 3' of the gene models (1:1 orthologous used).*



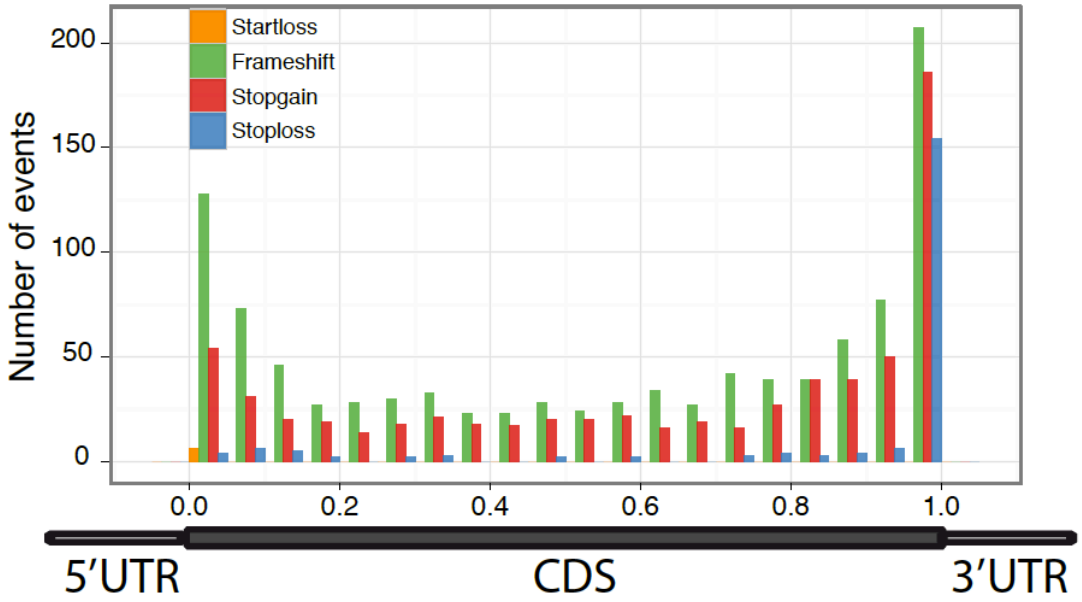
Suppl. Figure 10.4 – *Correlation between amount of gene loss events and genetic distance. The amount of loss-of-function events appears to appear at the same rate as the SNV divergence between species.*



Suppl. Figure 10.5 – *Correlation between amount of gene loss events and genetic distance in the branches larger than 0.003. We account for the Homo branch correcting for 67 non-processed pseudogenes. The human lineage does not appear to have accumulated an excess of loss-of-function events.*



Suppl. Figure 10.6 – *Mutations resulting in stop codon gains and losses fixed through great ape evolution are superimposed onto the great ape phylogeny. The numbers on each branch correspond to fixed substitutions, indels, or large deletions resulting in disruptions of the coding sequence. To correct for the human reference, mutations seen in all species, except human, were assigned to the human branch. Numbers in parenthesis show the observed and expected number of events per branch.*



Suppl. Figure 10.7 – *Genic position of disruptive events. Distributions are shown for startloss, frameshift, stopgain, and stoploss events in 5% bins throughout the gene. Both the beginning and end of the genes show an excess of loss-of-function mutations. Stoplosses in the middle of the gene are only predicted in the human and the hominini branches and indicate gene extensions in these lineages.*

We formally tested whether the rate of fixation of LoFs in the lineage of the great apes is 1) similar to that of substitutions per Kbp; and 2) shows any signs of being different in the human internal branch vs. the internal branches other great apes.

We conducted two different analyses. First, we followed the approach of Marques-Bonet *et al*.[37] and built a Likelihood Ratio Test framework that tests whether seven independent rates explain our observations significantly better than a single one. Secondly, we used a frequentist approach and presented rates in terms of observed vs. expected.

As a unit of time, we use, for each branch, either the number of substitutions per Kbp obtained from the same individuals or an estimate in units of Myrs of the length of each branch. The following table indicates the units of indel and time in the branches being tested. We use internal branches to avoid several problems, overfitting amongst them.

	Indels	**SNVs**	**Divergence*** Kbp	**Time**
Terminal abelii	21	14	1.55	0.88
Terminal Bonobo	37	40	1.78	1.3
Internal Chimpanzees	19	6	0.81	0.8
Internal Gorilla	132	124	5.84	8.06
Internal *Pongo*	442	340	20.21	15.92
Internal *Pan*	77	70	3.91	3.8
Terminal pygmaeus	42	13	1.62	0.88

To perform an LRT, we first obtained maximum-likelihood estimates for two different models. The simplest one assumes a single rate of accumulation of indels everywhere and the other one assumes that every branch has its own rate. Afterwards, we perform an LTR between the two models. We use 6 degrees of freedom since the second model has 6 more parameters.

Tabulated below are the results of the estimates and the p-value of the LOFS test that can be performed with the two units of time, Myrs or substitutions per Kbp.

LOFS / Myrs			
	Model 1 (all identical rate)	Model 2 (Seven different rates)	LTR p-value

One vs. Seven rates	λ = 43.52 LoFs/ Myrs	λ_{abe} = 39.77 λ_{bon} = 59.23 λ_{chi} = 31.25 λ_{gor} = 31.76 λ_{ora} = 49.12 λ_{pan}= 38.68 λ_{pyg} = 62.50	$1.25x10^{-10}$

LOFS / NumSubstperKb			
	Model 1 **(all identical rate)**	**Model 2** **(Seven different rates)**	**LTR** **p-value**
One vs. Seven rates	λ = 38.55 LoFs/ NumSubstperKb	λ_{abe} = 22.58 λ_{bon} = 43.26 λ_{chi} = 30.86 λ_{gor} = 43.83 λ_{ora} = 38.69 λ_{pan}= 37.60 λ_{pyg} = 33.95	0.00344822

Using both units of time, several rates explain things better. Naturally, all the ML estimated rates in each branch are equivalent to their observed rate. Admittedly, however, we do have some overfitting.

Now we want to see if in any particular branch the observed rate is larger or smaller than the expected rate (considering as the expected the overall rate). We indicate the difference and the p-values corresponding to a test assuming a Poisson distribution of events per branch. This would be equivalent to asking: To which branches do we owe a larger deviation from a common expectation?

LOFS / NumSubstperKb		
Branches	O/E ratio	p-value
P.abelii	0.59	0.00275
P.paniscus	1.12	0.20972
Internal chimpanzees	0.80	0.09373
Internal gorilla	1.14	0.20972
Internal orangs	1.00	0.50753
Internal Pan	0.98	0.44321
P. pygmaeus	0.88	0.21053

Human branch

We applied the same test for human vs. the other three internal branches.

	Divergence* Kbp	**LOFS**
Human	5.55	223

GGO	5.84	256
Pan	3.91	147
Orang	20.21	782

The overall expected rate would be λ = 39.71 LOFS/ NumSubstperKb, which is not significantly different than the other four different rates (p-value 0.7349). In terms of O/E, it is again self-explanatory.

O/E approach		
LOFS / NumSubstperKb		
Branches	O/E ratio	p-value
Human terminal	1.02	0.43983
Shared with gorilla	1.10	0.26817
Shared with orang	0.97	0.43395
Shared with Pan	0.95	0.37174

RNAseq validation

To provide further evidence of whether these variants are found on expressed genes and carry the same mutations previously predicted on genomic data, we used RNAseq data[38]. We mapped these data to the hg18 allowing up to an indel size of 8 using TopHat splice junction mapper with the human gene models. Then we called variants using SAMtools roughly, without applying further filtering. We analyzed only the LoF events where the region is covered by least five reads to have enough support in the SNP and indel calling. We obtained >97% of validation in all LoF predicted with genomic data (**Suppl. Table 10.1**). This validation rate is concordant with the previous validations using Sanger capillary sequencing.

	SNV (%)	SNV (#)	INDEL (%)	INDEL (#)
Bonobo	100%	108 of 108	97.40%	75 of 77
Chimpanzee	98.90%	90 of 91	97.33%	73 of 75
Gorilla	100%	122 of 122	98.89%	95 of 96
Orangutan	97.84%	136 of 139	97.43%	76 of 78

Suppl. Table 10.1 – *Validation of fixed LoF mutations in RNAseq data from Brawand et al. where coverage (>4X) allow us to genotype according to transcriptome expression data.*

Section 11: X versus autosomes

August Woerner, Krishna R Veeramah, Michael F Hammer

For both autosomes and the X chromosome, genic regions for hg18 were defined based on the set-union (to take into account overlapping and alternative transcripts, etc.) of all genes defined in the RefSeq Gene Collection. 20 Kbp loci were then identified as previously reported[39]. In brief, a central nongenic locus was first found whose 5' and 3' tips are maximally distant, in genetic units, from the nearest genes. Successive non-overlapping 20 Kbp loci were identified walking towards genic boundaries from this central locus in both the 5' and 3' directions until the gene boundaries were reached. Genetic distances were defined based on the fine-scale recombination map of Hinch et al.[9] estimated using ancestry switches detected in African Americans (http://www.well.ox.ac.uk/~anjali/AAmap/). Loci in the pseudo-autosomal region of the X chromosome were excluded from the analysis.

Nucleotide diversity/divergence (π/D) was calculated for each locus for each subspecies. Species-specific masks were applied to the loci and thus not all loci contained 20 Kbp of callable sequence for this calculation and the total sequence considered may change between species. Loci with less than 5 Kbp of callable sequence were dropped from the analysis. Note that 20 Kbp was found to be of sufficient size to estimate π/D at a reasonable level of accuracy (i.e., contain sufficient segregating sites) but still represent a relatively confined interval of genetic distance.

For a given subspecies, D was taken as the average divergence of all individuals relative to the ancestral node of the primate phylogeny defined for this dataset (see section on ancestral allele calls). For each nucleotide position considered, an ancestral allele was chosen randomly, weighted by the relative probabilities of the four alleles at the ancestral node. For nucleotide positions segregating with two alleles in a subspecies where one of the alleles was the same as the ancestral node, a divergence value of 1 was assigned to all individuals with the derived allele. For nucleotide positions segregating where neither allele was the same as the ancestral node, we assumed that two mutations had occurred and that one mutation occurred on the background of the other (rather than occurring independently on the background of the ancestral allele). The first mutation to occur was chosen randomly, weighted by the relative allele frequencies (i.e., the major allele is more likely to have mutated first). Individuals inferred to have two changes from the ancestral allele were assigned a divergence value of 2. As the sample sizes were fairly restrictive, male X chromosomes were used in the estimate of π and D. This was achieved by assigning male genotypes to the most-frequent allele for SNPs on the X chromosome.

For each subspecies, loci were grouped into six bins ([0−0.05], [0.05−0.1], [0.1−0.2], [0.2−0.4], [0.4−0.8], [0.8−2.0]) of increase cM distance from the nearest genic regions. Mean π/D was then calculated for all loci within each bin. Bins were of increasing interval size (0.05cM in the first bin to 1.2cM in the final bin) to account for the number of loci available being reduced moving further away from genes, as described in Gottipati et al[40]. 95% Confidence intervals (CIs) for each bin were calculated using standard bootstrapping of loci, though we note that this will not fully take into account the interdependence between neighboring loci in the same bin and thus are likely somewhat anti-conservative. The ratio of X to autosomal diversity at each of the bin positions was calculated by dividing the mean π_X/D_X by the mean π_A/D_A.

Data quality filtering

For the initial analysis, no coverage filters were applied to the data and for male samples haploid calls on the X chromosome were based on the allele with highest read depth (AD field from GATK Unified Genotyper). We also explored applying 5X coverage filters on the autosomes and X chromosomes and found no noticeable different in results without filters.

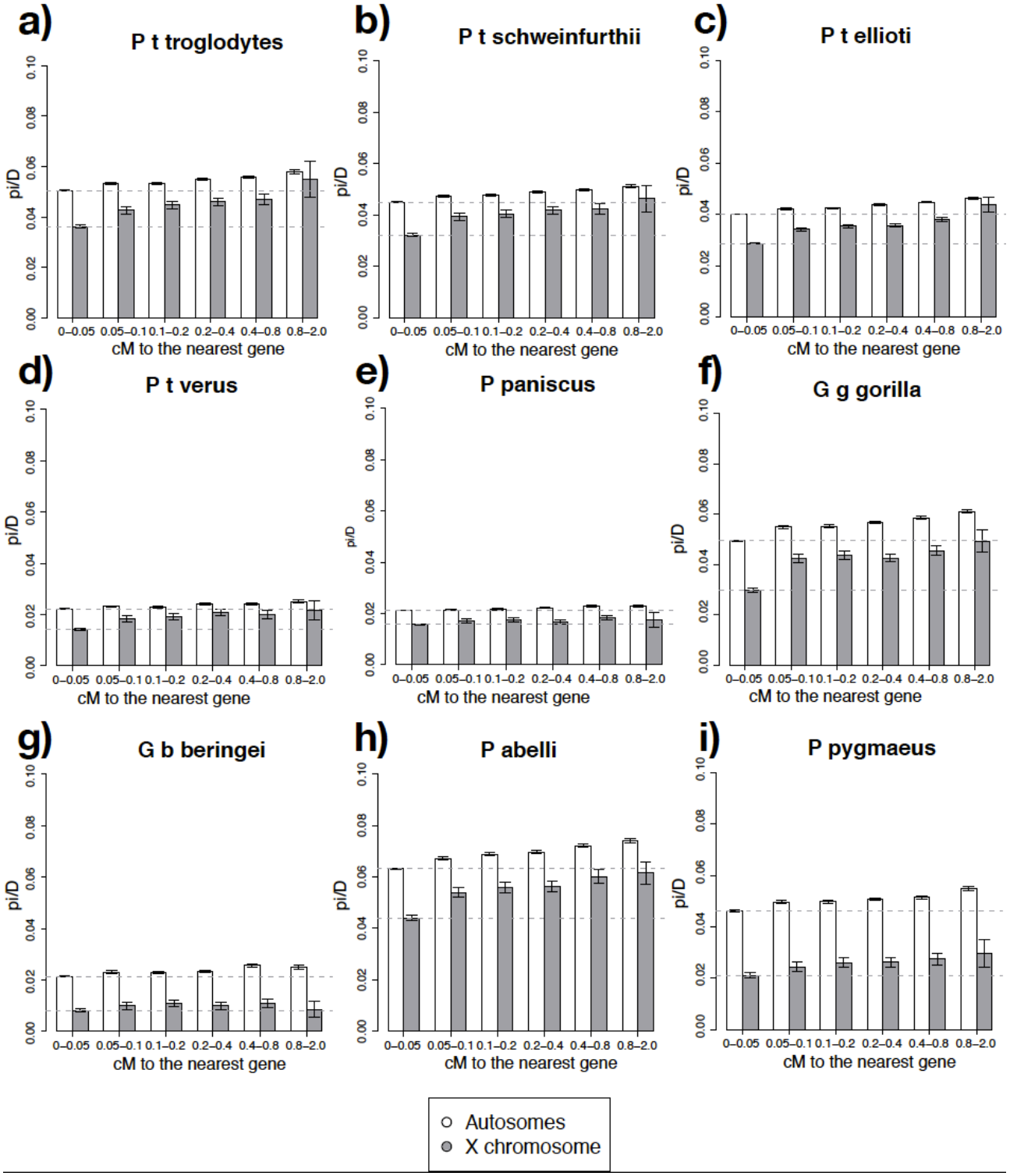
Contrasting autosome and X chromosome variation can inform us about population demography, sex-specific behaviors, and selective constraint. Therefore, we examined the level of both autosomal and X chromosomal diversity as a function of genetic distance from genes within each great ape subspecies. To control for variation in mutation rate among loci, diversity was divided by divergence from a common ancestral primate node. Consistent with previous studies in humans[39–41], both autosomal and X chromosome diversity increase when moving away from genes in nearly all subspecies (**Suppl. Figure 11.1**). In addition, X chromosome diversity generally increases at a faster rate, as seen previously in humans. Close to genes, X/A diversity is lower than the expected neutral value of 0.75 (**Suppl. Figure 11.1**), suggesting that the effect of purifying or positive selection is greater on the X chromosome. Such an effect is expected if novel mutations tend to be partially recessive because they are more quickly exposed to selection in hemizygous males. In contrast, X/A diversity is usually greater than 0.75 at regions far from genes. Given that these regions are expected to be less affected by selection, this pattern of diversity is consistent with an increased variance in male reproductive success for seven of the nine subspecies considered here.

The overall rate of increase of X/A diversity is similar across all four chimpanzee subspecies despite a wide range of N_e estimates, suggesting similar distributions of fitness effects among species. Central, Eastern, and Nigeria-Cameroon chimpanzees show very similar levels of X/A diversity close to genes. Interestingly, the subspecies with the lowest effective population size, Western chimpanzees, exhibits

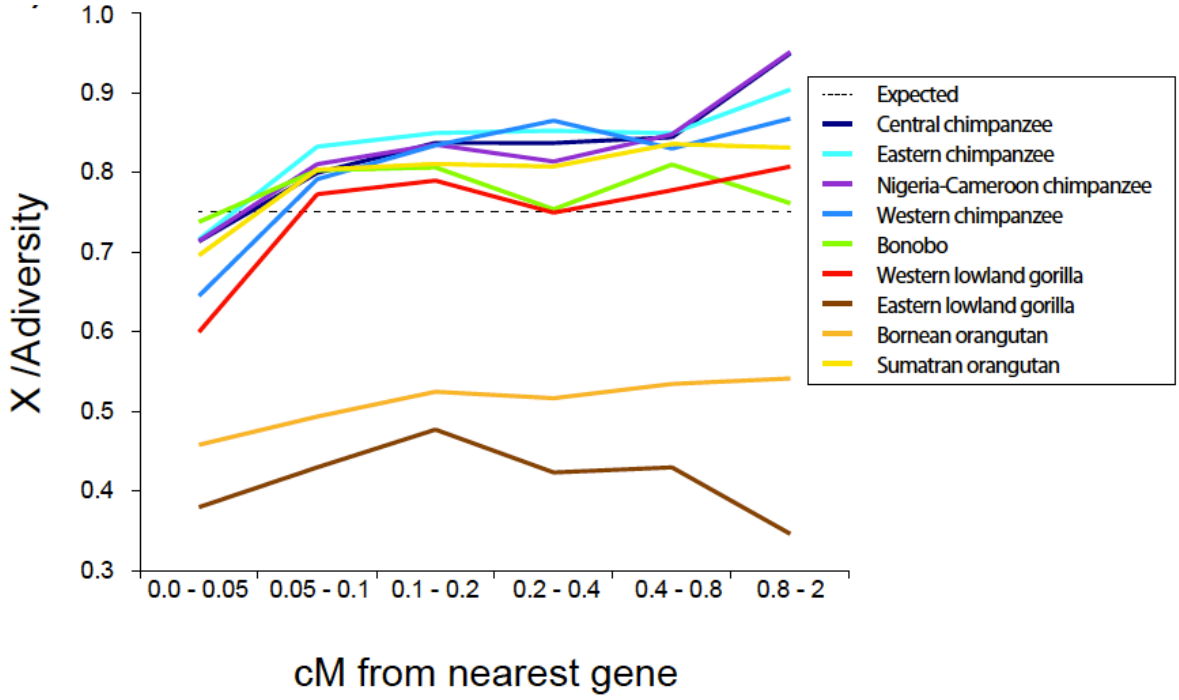
substantially lower X/A diversity near genes (the CIs of Western chimpanzee X/A diversity do not overlap with any of the other three subspecies). Demographic effects are unlikely to be the cause of this as X/A diversity levels quickly recover as one moves away from genes. One possible explanation is that Western chimpanzees have experienced stronger selection on the X chromosome relative to the autosomes.

Patterns of X/A diversity are similar in Western gorillas and Sumatran orangutans. However, both Bornean orangutans and Eastern gorillas demonstrate substantially reduced levels of X/A diversity compared with expectations, regardless of distance from genes (e.g., mean values are between 0.54 and 0.35, respectively) (**Suppl. Figure 11.1.a, g, i**). This is consistent with theoretical predictions for a recent reduction in population size[42]. Another possible contributing factor is male-specific migration into these populations from other subspecies following their divergence, which would increase the number of breeding males (a similar hypothesis has been proposed in humans to explain the reduction in X/A diversity in non-Africans relative to Africans[43]).

Interestingly, bonobos exhibit little or no evidence of reduced variation close to genes on either the X chromosome or the autosomes (**Suppl. Figure 11.1.e**). Prüfer et al.[30] also noted a poor correlation between bonobo diversity and regions defined in humans to be under different levels of background selection. This suggests that levels of selective constraint might be reduced in bonobos compared with most great apes, which would be consistent with the finding of a higher ratio of deleterious nonsynonymous to synonymous mutations (**Suppl. Figure 11.2**). Alternative (and perhaps less likely) explanations include a dramatic shift in rates of recombination or position of hotspots on the bonobo lineage, or complex demographic factors that effectively erased signals of positive or purifying selection in the bonobo genome.



Suppl. Figure 11.1 – *Mean estimates of π/D on the autosomes and X chromosome as a function of distance from genic regions for all primate subspecies.*



Suppl. Figure 11.2 – *Levels of X/A diversity as a function of distance from genes. The dotted line represents the expected X/A ratio under neutrality.*

Section 12: Recent demography

Estimating gene flow in chimpanzees

Whether gene flow has occurred between populations that are genetically distinct is major question for those interested in the demographic history of populations[44,45]. Gene flow can be estimated in different ways—for instance, by evaluating a model where effective population size, population divergence, and gene flow occur simultaneously and are parameterized. Models derived from a rich body of population genetic theory such as those based on the backwards-in-time coalescent (e.g., isolation-migration (IM) methods[46] or similar Approximate Bayesian Computation (ABC) frameworks[47]) and the forward-in-time diffusion approximation[48,49] present elegant and powerful approaches for obtaining actual estimates for these parameters. However, as in all explicit modeling inference, the investigator must guard against the curse of dimensionality and balance the number of parameters estimated (and thus the complexity of the model) with inferential power; this balance is particularly important to consider in the case of ABC methods where extremely complicated models can be proposed based on simulations[50] (IM and diffusion approximation models are somewhat already limited by certain analytical constraints). Both oversimplified and overcomplicated models can produce parameter estimates that lack accuracy, precision, or both and the effect is not always predictable or easily quantifiable. Even the relatively simple scenario of gene flow between two populations after divergence can be affected by factors such as variable strength of migration over time, the time migration began and ended after population divergence, and the presence of asymmetric migration in one direction.

Therefore, methods that can infer gene flow without invoking a particular model of demography are attractive alternatives, especially in scenarios where there is very little information about the underlying demography to appropriately parameterize a model with any confidence. An example of such a method that is particularly applicable to whole-genome data is the D-statistic[51], which is based on a relatively simple summary of allele sharing between three populations of interest and an outgroup. For example, when two populations are known to show a close phylogenetic relationship compared to a more distantly diverged third, the D-statistic can produce compelling evidence of unbalanced gene flow between the external population and the two internal populations (e.g., evidenced by the Neanderthal introgression into non-Africans[52]). However, while the D-statistic can provide evidence that such gene flow may have occurred, the quantification and timing of this gene flow is not possible or extremely difficult to infer without invoking explicit models (e.g., how much gene flow has there been and was the gene flow recent, old, continuous, or instantaneous). The interpretation of a D-statistic usually requires the assumption of some underlying model of divergence (e.g., the D-statistic result in Neanderthals can also be explained assuming a model of ancient population structure in Africa[52,53]).

We additionally examine gene flow using TreeMix[54], which estimates population splits and migrations between populations using a graph. The method works by first building a bifurcating tree to represent the relationships among populations. Populations that are a poor fit to the inferred tree are then identified, and gene flow events involving such populations are incorporated to improve the fit of the model to the observed data. This statistical framework allows us to estimate the presence and amount of gene flow between divergent populations with the caveat that inferred migrations is limited by the number of populations considered (in our case, we can detect only one migration).

Previous studies on chimpanzee gene flow

Our analysis of genetic variation in chimpanzees demonstrated the presence of distinct populations that correspond to the four known subspecies. The major patterns of genetic differentiation between the four subspecies can be parameterized by a model involving a series of population divergence events beginning approximately ~500,000 years ago (**Figure 2**). However, a number of papers examining autosomal loci in chimpanzees[46,55–57] have demonstrated that it is also important to consider gene flow occurring subsequent to population divergence to fully explain patterns of genetic variation amongst chimpanzees. All four previous studies invoked a coalescent-based modeling approach that assumed some topology of population splitting with subsequent gene flow. Becquet and Prezorwski examined Central, Western and Eastern chimpanzees as pairs of populations with symmetric migration and identified the strongest signal of gene flow between Western and Eastern chimpanzees. Both Hey[46] and Wegmann and Exoffier[57] examined the same populations in a single analysis, but while the former allowed asymmetric migration between all pairs of extant and ancestral populations, the latter restricted asymmetric migration to Central and Western chimpanzees (based on the results of Won and Hey) and did not allow migration between Eastern and Western chimpanzees. The more parameterized model of Hey found migration into Eastern chimpanzees from both neighboring Central chimpanzees as well the more geographically distant Western chimpanzees and also found posteriors with non-zero peaks for all other pairwise comparisons except from Eastern to Western chimpanzees, though some results were dependent on the particular priors applied. Even more interesting was the identification of statistically significant migration from Western chimpanzees into the ancestors of Central and Eastern chimpanzees. Wegmann and Excoffier, who used an ABC approach with a more limited parameterization of migration, also estimated population growth parameters within a single framework. Therefore, the estimates of migration were less nuanced in this study, though they did identify strong asymmetric migration from Western to Central chimpanzees.

It is important to note, however, that all these previous studies were based on limited amounts of sequence data and did not incorporate Nigeria-Cameroon chimpanzee population genetic data. Though the phylogeny for bonobos and Western, Central and Eastern common chimpanzees is well established, there is still uncertainty regarding their relationship to Nigeria-Cameroon chimpanzees (*P.t. ellioti*)[58]. Thus, we examined the

relationship among all four chimpanzee subspecies by inferring the pattern of sequence divergence using classical phylogenetic methods as well as population divergence using a coalescent-based ABC analysis (as the former does not always reflect the latter[59]). Regional neighbor-joining trees and a maximum-likelihood tree, estimated from allele frequencies, show that Nigeria-Cameroon chimpanzees and Western chimpanzees form a clade at the sequence divergence level. This topology is also supported at the population divergence level in our ABC analysis as well as a pairwise PSMC divergence analysis (**Figure 3**). Though we note that the inclusion of a relatively simple model of symmetrical migration between all pairs of populations complicates this inference and it becomes harder to discriminate this balanced topology to the unbalanced topology previously inferred from microsatellite data[58] with Western chimpanzees as an outgroup to the other three chimpanzee subspecies. This suggests that a simple model of divergence with isolation cannot fully explain our whole-genome data and indicates the presence of complex patterns of post-divergence migration and admixture as suggested by the previous studies described above.

The strongest signal we identified of symmetric migration between adjacent populations from the ABC analyses was between the Eastern and Central chimpanzees, as previously observed in an analysis using a similarly specified model of symmetrical migration[56]. A second signal of migration is also observed in the other parapatric (**Figure 1**) comparison involving Central and Nigeria-Cameroon chimpanzees (the latter population of which was not included in Becquet and Prezworski 2007).

However, as described above, two previous analyses have identified asymmetrical gene flow patterns amongst chimpanzees[46,57] that our ABC analysis is not parameterized to infer.

Given the best-supported topology of chimpanzees inferred above, we applied the D-statistic to test whether unequal levels of gene flow have occurred between an out group subspecies and two subspecies that have more recently diverged. Consistent with the observation of Hey (2010), this analysis shows that Western chimpanzees are genetically closer to Eastern than to Central chimpanzees (D(H,W;E,C)>16SD). Yet, an even larger D-statistic was found that is suggestive of gene flow between Nigeria-Cameroon chimpanzees and Eastern chimpanzees, while TreeMix (which is only able to model the strongest gene flow event) also identified this signal ($P=2 \times 10^{-300}$) and orientated the event from Nigeria-Cameroon into Eastern chimpanzees. Finally, we also find that Eastern and Central chimpanzees are both closer to Nigeria-Cameroon than to Western chimpanzees (D(H,E;W,N)>25SD, D(H,C;W,N)>17SD).

As noted by Hey (2010)[46], direct migration from Western to Eastern chimpanzees seems geographically unlikely (today they are separated by 3,000 km), but indirect

gene flow through an intermediary population such as Nigeria-Cameroon chimpanzees, as hinted by our analyses, may provide a more plausible mechanism to explain previous inferences of gene flow (though more complicated scenarios involving ancestral migration likely also contribute). Clearly the inclusion of Nigeria-Cameroon samples will be key in future studies that focus in more depth on teasing apart migration and admixture patterns amongst chimpanzees.

12.1. D-statistic

Heng Li

To formally test unbalanced gene flows between species, we performed a D-statistic test. For four haploid sequences U, V, X and Y, a site is classified as BABA if at the site U=X≠Y=V, or classified as ABBA if U=Y≠V=X. Define:

$$D'(U,V;X,Y) = \frac{\#BABA - \#ABBA}{\#BABA + \#ABBA}$$

The D-statistic equals the ratio of D' to its standard deviation estimated with block jackknife. A positive D indicates that the genetic distance between U and X is larger than between V and Y, while a negative D indicates that the distance between U and Y is larger than V and X. In particular, if U is a known outgroup of other samples, D(U,V;X,Y)>0 if V is genetically closer to Y, while D(U,V;X,Y)<0 if V is genetically closer to X. The D-statistic provides a formal and model-free test for testing unbalanced gene flows between the (U,V) clade and the (X,Y) clade.

Suppl. Table 12.1.1 shows the D-statistics between different subspecies. Let d(X,Y) be the genetic distance between two populations X and Y. The table suggests that the following inequalities: d(pte-Koto,pts-Nakuu)<d(pte-Koto,ptt-Vaillent)<d(ptv-Clint,ptt-Vaillent) and d(pte-Koto,pts-Nakuu)<d(ptv-Clint,pts-Nakuu)<d(ptv-Clint,ptt-Vaillent). Combining the two inequalities we know that d(pte-Koto,pts-Nakuu) is the closest pair between the two chimpanzee clades, which is also identified by TreeMix. Interestingly, the smallest D-statistic among common chimpanzees is achieved for D(ptv,pte;pts,ptt). A parsimonious scenario that is consistent with an insignificant D(ptv,pte;pts,ptt) but significant other D values would be that a branch of ancestral population of Western and Nigeria-Cameroon later admixed into Eastern, while a branch of ancestral population of Eastern and Central admixed into Nigeria-Cameroon chimpanzee. Because these two gene flows do not affect the balance of the (ptv,pte) and (pts,ptt) clades, they will result in an insignificant D(ptv,pte;pts,ptt).

Sample U	Sample V	Sample X	Sample Y	$D'(U,V;X,Y)$	D
Human	pp-Dzeeta	ptv-Clint	pte-Koto	0.53%	1.24
		ptv-Clint	ptt-Doris	0.83%	2.11
		ptv-Clint	pts-Kidongo	-0.57%	-1.43
		pte-Koto	ptt-Doris	0.43%	1.14
		pte-Koto	pts-Kidongo	-0.85%	-2.19
		ptt-Doris	pts-Nakuu	-1.05%	-3.00
Human	ptv-Clint	pts-Nakuu	ptt-Vaillent	-5.33%	-16.35
	ptv-Koto	pts-Nakuu	ptt-Vaillent	-6.61%	-20.15
	pts-Nakuu	ptv-Clint	pte-Koto	9.06%	25.72
	ptt-Vaillent	ptv-Clint	pte-Koto	7.34%	21.83
ptv-Clint	pte-Koto	pts-Nakuu	ptt-Vaillent	-2.54%	-6.66
ptv-Clint	pte-Julie	pts-Nakuu	ptt-Vaillent	0.50%	1.30
ptv-Clint	pte-Tobi	pts-Nakuu	ptt-Vaillent	0.83%	2.08
ptv-Clint	pte-Banyo	pts-Nakuu	ptt-Vaillent	-0.92%	-2.27
Human	gbg-Mkubwa	ggg-Delphi	ggd-Nyango	2.82%	8.23

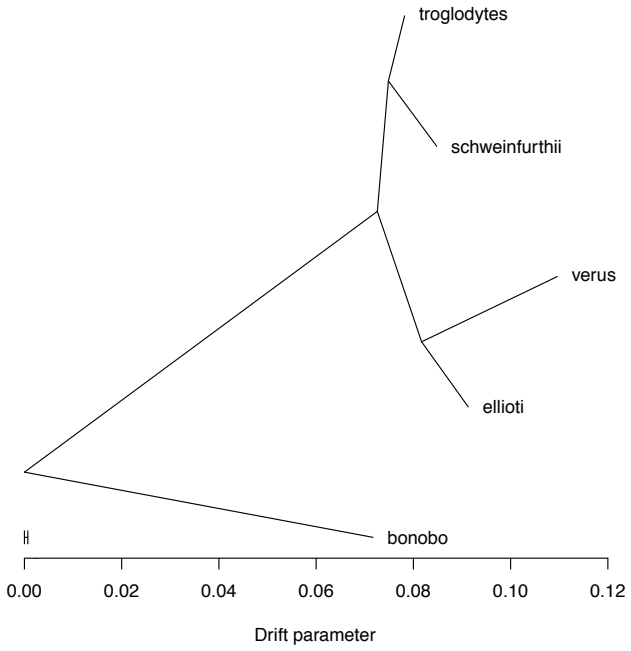
Suppl. Table 12.1.1 – *D-statistic for four haploid sequences, A, B, X and O, a site is classified as ABBA if at the site base A = O B = X, or classified as BABA if B = O= A = X. Define D'(A, B, X; O)== (#ABBA –#BABA)/(#ABBA +#BABA). D(A, B, X; O) equals the ratio of the mean of D' to its standard deviation, estimated by block Jack-knife. A positive D value indicates that sample B is genetically closer to X, while a negative vale indicates A closer to X.*

12.2 TreeMix

Joanna L. Kelley, Jeffrey M. Kidd

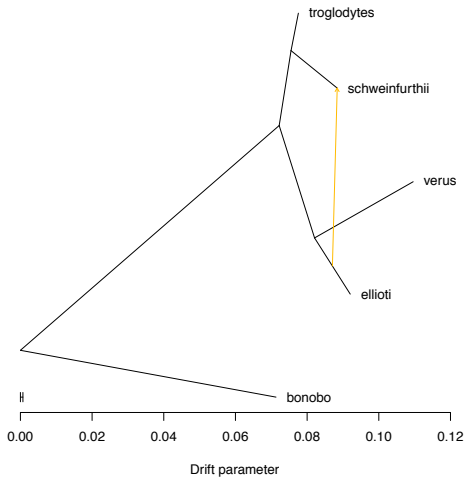
To provide a more complete picture among chimpanzee demographic history, we applied a method to infer population splits and gene flow events, TreeMix[54]. We based our analysis on a set of ~5 million polymorphic sites that were randomly selected from the total set of sites using PLINK[18] (--thin 0.31). We used the SNP calls derived from mapping to the human genome reference (NCBI Build 36). We only considered autosomal SNPs, and removed variants on random or unassigned chromosomes.

We first assessed the tree topology using TreeMix with the randomly thinned data we infer a ML tree and residual fit (**Suppl. Figure 12.2.1**) with all 38 individuals. The tree model explains 99.92% of the variance in the data.



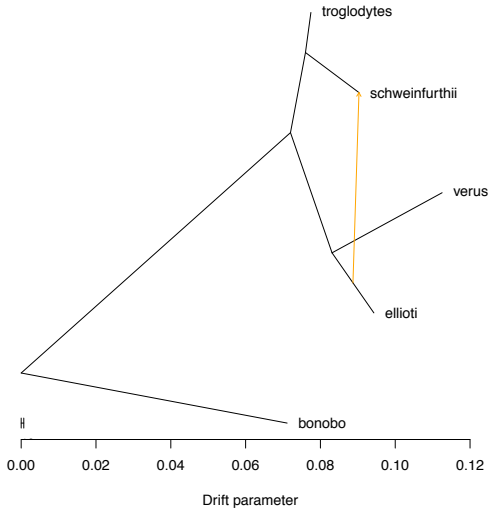
Suppl. Figure 12.2.1 – *ML tree inferred with TreeMix. Five million randomly selected SNPs without regard to MAF or LD. LD grouping with 5,000 SNPs per bin.*

Using the ML tree as the user specified tree, we use TreeMix to infer one migration event (**Suppl. Figure 12.2.2**). The model with one migration event explains 99.98% of the variation. The migration event is inferred from the *P.t.ellioti* branch to *P.t.schweinfurthii*. The TreeMix migration event weight for the *P.t.ellioti* ancestry into the *P.t.schweinfurthii* population is 9.2% ± 0.2% standard errors ($P<2.22 \times 10^{-308}$), using a block jackknife to obtain standard errors and P-values (with 5,000 sites in each block).



Suppl. Figure 12.2.2 – *ML tree with one migration event. It shows a significant migration between ellioti and schweinfurthii populations.*

The results are robust to removing individuals with potential admixture (**Suppl. Figure 12.2.3**). 34 individuals remain after removing those with evidence of admixture. The tree without migration explains 99.93% of the variance. With one migration event, 99.98% of the variance is explained. The migration estimate is 13.6% ± 0.6% (P<2.22x10-308), using a block jackknife to obtain standard errors and P-values (with 5,000 sites in each block). This provides further evidence in addition to the D-statistic to suggest gene flow between Nigeria-Cameroon and Eastern chimpanzees.



Suppl. Figure 12.2.3 – *ML tree with one migration event for non-admixed individuals. The gene flow between ellioti and schweinfurthii is still maintained.*

12.3 Approximate Bayesian Computation analysis of chimpanzee demography

Krishna R Veeramah, August Woerner, Michael F Hammer

Methods

In order to characterize *Pan troglodytes* demographic history at the whole-genome level we developed an Approximate Bayesian Computation (ABC) approach based on coalescent simulations[60] amenable to handling large scale data (tens of millions of bases for multiple individuals). The aim of this analysis was primarily to identify the appropriate topology describing *p.t.* subspecies divergence but the general framework also allowed us to estimate parameters such effective population sizes, population divergence times and migration rates.

Sequence data

Performing ABC in the context of the data assembled in this study is complicated by the following factors; mapping of chimpanzees to a human reference (and thus no usable recombination map), a lack of information on haplotype phase and potential intra and inter species contamination (and the subsequent application of an allele balance filter resulting in a ~10% loss in heterozygosity). Therefore we chose to perform the analysis on a subset of ~37,000 1 Kbp loci previously identified for inference of population genetic parameters in humans (and that show good synteny with chimpanzees)[61] and that had at least 90% of bases called in all individuals with no or little evidence of contamination (after applying filters previously described as well as the masks described in **Section 2**) This resulted in sequence data from 3 *p.t.t*, 3 *p.t.s*, 7 *p.t.e* and 2 *p.t.v* individuals at 10,008 loci, with no allele balance filter applied. Ancestral alleles at segregating sites were identified by orientating against the bonobo and human reference genomes. False heterozygotes caused by high error rates in next generation sequencing could present a problem in demographic inference with regard to singletons[62]. However, as the average coverage for these chimpanzee genomes was high and the total sequence data considered relatively low (~10MB per individual), we do not consider this to be a major issue.

Mutation rate estimates

Mutation rates for loci were estimated using previously aligned human data[61] as previously described[62], with a Jukes Cantor correction applied to sequence divergence estimates and the assumption of a human chimpanzee split of 6 million years, 25 years per generation in humans, and an ancestral population size of 83,000 (estimated by CoalHMM analysis in **Section 13**). The mean mutation rate across loci was 1.08×10^{-8} per base per generation (stdev = 5.6×10^{-9}), in line with recent estimates of the human mutation rate[63,64]. While the absolute estimates of mutation rates at these loci may not be directly applicable to chimpanzees, the

relative rates are likely to be captured amongst loci, at which point the inferential problem becomes one of scaling parameter estimates appropriately.

Summary statistics

We computed the following summary statistics to describe the data for every pair of populations: number of shared polymorphisms, number of private polymorphisms in each populations and the number of private fixed sites in each population. These statistics are known to contain substantial information about population demography[65] and are utilized in the program MIMAR which has previously been used to estimate Chimpanzee demography under an isolation-migration model[56]. These statistics are particularly useful as they are ambiguous to the requirement of haplotype inference. We use the mean and variance of these summary statistics across all loci to describe the data (unlike MIMAR where these summaries are used to calculate a likelihood for each locus individually, which is computationally intensive for the amount of data considered here). Other summary statistics that might traditionally be considered useful for demographic inference such as Tajima's D were not utilized due to the low sample size for some subspecies considered in the analysis. Therefore our method is unlikely to capture the inference of parameters such as exponential growth rates for population size and thus are not considered here (though they have been estimated previously[57]).

Demographic models

We consider two main subspecies topologies when constructing models of *p.t* demography (**Suppl. Figure 12.3.1**). The first is an unbalanced topology with *p.t.v* representing the earliest diverged subspecies that has been described previously[58] and is consistent with the PCA, Frappe (**Figure 1 and Section 8**) and diploid PSMC analysis (**Figure 3**). The second is a balanced topology, with *p.t.v* and *p.t.e* forming a distinct clade from *p.t.t* and *p.t.s*, and matches the neighbor-joining tree and sequence divergence patterns and was also inferred by the haploid PSMC analysis. However the latter analysis is problematic when coalescence times overlap speciation time, as would almost certainly be the case for the time frame of *p.t* subspecies divergence, with populations with the largest N_e being underestimated the most.

Parameters (and associated priors) describing both topologies are indicated in **Suppl. Figure 12.3.1, 12.3.4** and **Suppl. Table 12.3.1**. Prior distributions are motivated by Wegmann and Excoffier[57] and all are uniform distributed on a log10 (x) scale. The following classes of parameters are considered: effective population size (N), time of population divergence (t) and, for later iterations of the analysis, number of migrants per generation ($M=Nm$, with m equal to the migration rate). As priors for absolute times of population divergence would likely overlap, in order to obtain flat priors we considered the time of an internal branching event with

respect to the more recent branching event above in the topology, a divergence scheme used in the likelihood-based method MCMCcoal[66].

ABC analysis

ABC analysis was performed using two different regression adjustments depending on their application. When estimating model parameters we utilized ABCtoolbox[67], which implements a general linear model (GLM) adjustment[68] on retained simulations. To maximize sufficiency but limit dimensionality, the full set of summary statistics was transformed into partial least squares (PLS) components[47] and we used the change in Root Mean Square Error (RMSE) to guide the choice of number components. These PLS components were then used to estimate parameters. When performing model choice ABCtoolbox can be used to find the marginal density of each model in order to calculate a Bayes Factor. However there are concerns about biases resulting from a lack of summary statistic sufficiency when applying Bayes Factors in ABC[69]. Therefore we used the logistic regression (LR) method previously described[70] to perform model choice using an adapted version of the R function calmod.r as well a more naïve method (the direct method, DM) of the proportion of retained simulations from each model[71]. We used the Kruskal-Wallis-based ranking method described in Veeramah et al[72] to identify the set of and number summary statistics most relevant to model choice. Simulations were performed using a version of ms[73] adapted for Python to allow fast, parallel processing. Individual locus mutation rates were incorporated into theta and any sites missing in the real data (including non-segregating sites) were also masked in simulated data. 1% of simulations were retained for the GLM (parameter estimation) and LR (model choice) adjustments. Principal component analysis (PCA) was used for comparing the multidimensional distribution of summary statistics using the ‘‘prcomp’’ function in R.

Results

Models without migration

We examined the relative probabilities of two likely branching models involving *p.t.* Model 1 results in *p.t.v* branching off earliest from the *p.t.* lineage followed by *p.t.e* (unbalanced model, **Suppl. Figure 12.3.1A**), while Model 2 involves the ancestors of *p.t.v* and *p.t.e* branching off together (balanced model, **Suppl. Figure 12.3.1B**). Initially, in order to reduce the number of parameters examined, no migration was considered in these models. Unlike Model 1 (which is restricted by the divergence order), the divergence time T2 in Model 2 was chosen without regard to T1 and was free to be larger or smaller.

PCA visually demonstrated a good multidimensional fit between the observed summary statistic data and simulated data generated under both models (**Suppl.**

Figure 12.3.2). We then identified the set of summary statistics that best distinguished the two models based on simulated data, and were able to identify the correct model in 97% of cases using LR and 82% using DM[71]. Using this tuned set of summary statistics we obtained a LR and DM posterior probability (*P*) for Model 2 of 96% and 64% respectively. Based on these two estimates of the posterior probability there was a 100% (LR) and 80% (DM) probability respectively of Model 2 being the true model using the method of Fagundes et al.[16] (i.e., $\Pr(P_{Md2}=0.96|Md2)/[\Pr(P_{Md2}=0.96|Md2)+ \Pr(P_{Md2}=0.96|Md1)]$).

Summary statistics were then transformed into PLS components in order to infer parameters from Model 2. 10 PLS components were used to infer parameters based on a total of 300K simulations. A p-value for the fit of the GLM based on the fraction of retained simulations with a smaller or equal likelihood to the observed data was 0.558, indicating a good fit of the local adjustment to the observed data. In addition the 95% CIs appeared relatively reliable from simulated data, with 94-97% of known true parameter values falling within them from 1000 simulated pseudo-observed sets. Posterior probabilities are shown in **Suppl. Table 12.3.2** and posterior distributions visualized in **Suppl. Figure 12.3.3**. GLM-fitted posterior distributions generally showed good peakedness and were congruent with the raw retained distributions.

The estimates of N_e in present day populations, at least with regard to the order of magnitude, were compatible with other estimates described in this and other studies, with *p.t.t* highest and *p.t.v* lowest. The estimates of ancestral N_es (N_{T1}, N_{T2} and N_{anc}) were also compatible with previous work and the PSMC analysis (**Figure 3**), which suggests a decrease of N_e backwards in time along the p.t. lineage, with a relatively small N_e before all subspecies started diverging. The divergence time estimates suggested an older split for the *p.t.e*/*p.t.v* clade than the *p.t.s*/*p.t.t* clade by about 80K years (albeit with large CIs), which would be compatible with the current geographic distribution and the large genetic distance we observe between *p.t.e* and *p.t.v* via PCA analysis as well as being consistent with **Figure 8.5.1 and Figure 2**. The timing of the *p.t.s*/*p.t.t* split (3-36K and 1-30K generations with and without singletons respectively) was also in line with that of Wegmann and Excoffier [2] (8-25K generations) (we obtained wider CIs, in part as we allow a larger prior while Wegmann and Excoffier hit the limit of their prior within their 95% CI) while our additive median estimate of N_{T1}, N_{T2} and N_{anc} of 27K generations (30K without singletons) was also compatible with Wegmann and Excoffier's [2] estimate of the split of *p.t.v* from *p.t.t*/*p.t.s* (16-47K generations, 300K-940K years).

Models with symmetrical migration

Inference of migration parameters between Chimpanzee subspecies has previously been performed by Wegmann and Excoffier[57] and Hey[46]. Though our low sample

sizes limit our power to infer these parameters with great confidence we updated our divergence models to included symmetric migration between all pairs of present day and ancestral populations (an additional 9 migration parameters to the original models). By adding migration between ancestral populations it was necessary to further parameterize the balanced model by assuming one branching event occurred before another. Therefore Model 2 with migration was split into two balanced models with migration, Model 4A, where *p.t.t* and *p.t.s* diverged most recently and Model 4B, where *p.t.e* and *p.t.v* diverged most recently. Model 1 with symmetric migration was named Model 3 (**Suppl. Figure 12.3.4**).

PCA of the two original models against the three migration models showed a substantially better fit of the latter to the observed data (**Suppl. Figure 12.3.4**). This was confirmed in a model choice analysis using the DM, which consistently showed a much higher posterior probability for each migration model against its non-migration model counterpart (P>0.88-0.97) (**Suppl. Table 12.3.3**) (Model 2 was re-simulated as two models, 2A and 2B, to match the branching order restrictions of Models 4A and 4B respectively). The average power to distinguish migration from non-migration models using the DM was 86% (min 75%, max =100%) based on simulated data. Therefore migration appears to be an important factor to consider when inferring Chimpanzee demography. Though generally in good agreement with the DM estimates, the LR method on one occasion gave unusual results, which, having manual examined the data, is likely due to poor fit of the logistic regression model because of the lack of representation of the non-migration model in the retained dataset, and thus we discarded results using LR in this instance.

However, simulations demonstrated that there was almost no power to distinguish between the three migration models. When comparing Models 4A and 4B there was at least some (but not particularly strong) evidence that Model 4A (i.e., a more recent *p.t.t*/*p.t.s* split) is the most likely with a posterior probability of 57% and probability that this is the correct model of 63% given the posterior probability. This is consistent with the estimates of T1 being slightly more recent than T2 in Model 2 and the general patterns of diversity in the whole-genome data and previous work. In addition, when we attempted parameter inference for Model 4B, T1 was much older than for Models 3 and 4A while the posterior distribution for T2 for this model was extremely flat and non-informative unlike the other two models, indicating the method may be finding it difficult to identify a good divergence time for the second divergence event because of an incorrect branching order.

Models 3 and 4A were almost completely indistinguishable, with the probability of choosing the correct model being 50% (i.e., no better than random chance). This is perhaps not surprising as the likely recent population divergence events coupled with subsequent migration makes this a 'hard' phylogenetic problem that may need

more samples and/or larger or different loci to resolve (for example STRs to distinguish recent and ancient migrations). Therefore we consider the topology of Chimpanzee divergence with regard to *p.t.e* somewhat uncertain with both unbalanced and balanced models with migration fitting the data equally well. As a consequence we report parameter estimates for both models (**Suppl. Tables 12.3.4** and **12.3.5** and **Suppl. Figure 12.3.6** and **12.3.7**).

Despite not being able to distinguish the correct divergence topology there is still substantial information about the demographic processes connecting the *p.t* subspecies when conducting parameter inference with migration regardless of the model chosen. P-values examining the fit of the GLM to the observed data were again good (>0.69) while 95% CIs were generally reliable (though somewhat more noisy than for the non-migration model, not surprising given the increase in the number of parameters). N_e estimates were still in line with the non-migration models but median divergence times became older and the CIs more diffuse, as would be expected from adding migration after population divergence events. However, more interestingly, despite the low sample size historical migration was detected with relatively peaked posteriors regardless of which model of divergence topology was examined. Substantial migration was observed between *p.t.t* and *p.t.s,* as would be expected by their overlapping geographic ranges but migration was also detected between *p.t.e* and *p.t.t* (i.e., parapatric populations). There is also some indication of migration between the *p.t.t*/*p.t.s* ancestors and *p.t.v*, which is consistent with the finding of Hey[46].

In order to not over-parameterize the model we did not consider asymmetrical migration rates between pairs of populations as in Hey[46] or population growth like in Wegmann and Excoffier[57]. However, this will be an interesting question to tackle in the future in a similar ABC framework using larger numbers of whole genomes and longer loci (which will require mapping to the chimpanzee reference genome and a recombination map).

Parameter	Min	Max	Min[a]	Max[a]	Distribution	Md1	Md2	Md3	Md4A	Md4B
log(N1)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N2)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N3)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N4)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N_{T1})	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(N_{T2})	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(Nanc)	3.0	5.4	1,000	251,189	Uniform	X	X	X	X	X
log(T1)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log(T2)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log(T3)	3.0	5.0	20,000	2,000,000	Uniform	X	X	X	X	X
log($M_{1\text{-}2}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{2\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{2\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{3\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	X
log($M_{1,2\text{-}3}$)	-2.0	1.0	0.01	10	Uniform			X	X	
log($M_{1,2\text{-}4}$)	-2.0	1.0	0.01	10	Uniform			X	X	
log($M_{1\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform					X
log($M_{2\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform					X
log($M_{1\text{-}2\text{-}3,4}$)	-2.0	1.0	0.01	10	Uniform			X		
log($M_{1\text{-}2,3\text{-}4}$)	-2.0	1.0	0.01	10	Uniform				X	X

Note. [a]Values converted from log10 scale to real world estimates, with divergence time assuming a 20 year generation time. N1=*p.t.s*, N2=*p.t.t*, N3=*p.t.e*, N4=*p.t.v*

Suppl. Table 12.3.1 - *Priors for the various ABC models (X marks where prior is relevant).*

Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log(N1)	0.97	4.62	4.54	3.73	5.26	34,532	5,315	181,238
log(N2)	0.97	5.34	5.15	4.62	5.40	140,443	41,611	251,177
log(N3)	0.96	4.53	4.48	3.73	5.18	30,319	5,315	150,998
log(N4)	0.96	4.21	4.12	3.18	4.93	13,212	1,523	84,475
log(N_{T1})	0.95	5.13	5.02	4.43	5.40	104,342	27,029	251,189
log(N_{T2})	0.95	4.68	4.67	4.02	5.37	46,739	10,496	235,055
log(N_{anc})	0.94	4.42	4.42	4.37	4.47	26,459	23,668	29,693
log(T1)	0.95	4.01	4.01	3.42	4.56	205,334	52,899	725,421
log(T2)	0.95	4.19	4.15	3.60	4.62	281,203	79,365	825,353
log(T3)	0.96	3.38	3.51	3.00	4.15	65,232	20,000	284,564

Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 300K simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

Suppl. Table 12.3.2 - *Posterior estimates for Model 2.*

Best Fit Model (BFM)	Alternative Model	Probability of BFM (*Pr*)	Power from 1000 sims	Pr(BFM true | *Pr*)
Md 3	Md 1	96%	98%	98%
Md 4A	Md 2A	97%	88%	99%
Md 4B	Md 2B	87%	82%	82%
Md 4A	Md 4B	57%	62%	63%
Md 4B	Md 3	54%	54%	57%

Suppl. Table 12.3.3 - *Posterior probabilities comparing various combinations of model.*

Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log(N1)	0.99	3.70	3.96	3.00	4.97	9,220	1,000	93,319
log(N2)	0.93	5.07	4.88	4.01	5.40	76,268	10,210	251,189
log(N3)	0.99	3.78	3.82	3.05	4.60	6,584	1,123	39,372
log(N4)	1.00	3.64	3.65	3.14	4.15	4,423	1,394	14,150
$\log(N_{T1})$	0.91	4.91	4.44	3.30	5.40	27,234	1,975	251,189
$\log(N_{T2})$	0.94	4.17	4.27	3.17	5.35	18,490	1,481	224,880
$\log(N_{anc})$	0.96	4.09	4.02	3.09	4.85	10,543	1,234	71,162
log(T1)	0.95	4.62	4.35	3.31	5.00	450,837	41,243	2,000,000
log(T2)	0.92	4.67	4.25	3.23	5.00	352,128	34,298	2,000,000
log(T3)	0.93	4.61	4.14	3.19	5.00	278,535	30,990	2,000,000
$\log(M_{1\text{-}2})$	0.98	0.08	-0.07	-1.41	1.00	0.85	0.04	10.00
$\log(M_{1\text{-}3})$	0.91	-1.42	-1.22	-2.00	-0.13	0.06	0.01	0.74
$\log(M_{1\text{-}4})$	0.98	-1.72	-1.50	-2.00	-0.68	0.03	0.01	0.21
$\log(M_{2\text{-}3})$	0.96	-0.72	-0.77	-1.98	0.38	0.17	0.01	2.37
$\log(M_{2\text{-}4})$	0.92	-1.54	-1.35	-2.00	-0.50	0.04	0.01	0.32
$\log(M_{3\text{-}4})$	0.94	-1.04	-1.05	-1.71	-0.40	0.09	0.02	0.40
$\log(M_{1,2\text{-}3})$	0.91	-0.75	-0.63	-1.96	0.77	0.23	0.01	5.83
$\log(M_{1,2\text{-}4})$	0.94	-1.41	-0.93	-2.00	0.61	0.12	0.01	4.07
$\log(M_{1\text{-}2\text{-}3,4})$	0.90	-1.17	-0.51	-1.90	0.86	0.31	0.01	7.18

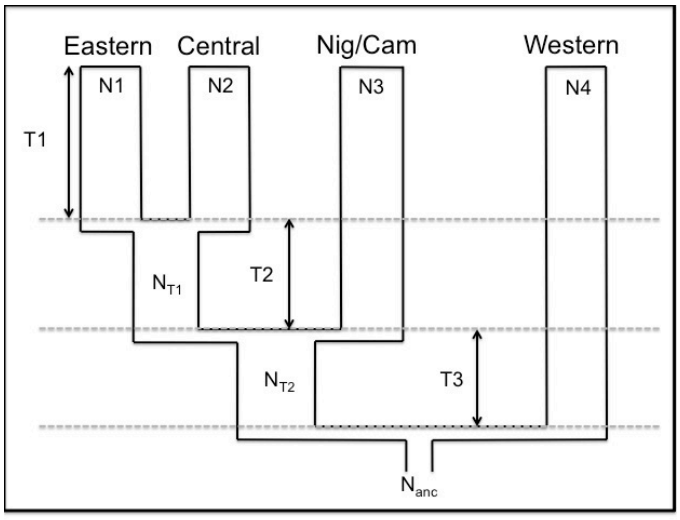
Suppl. Table 12.3.4 - *Posterior estimates for Model 3.*

Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 1M simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

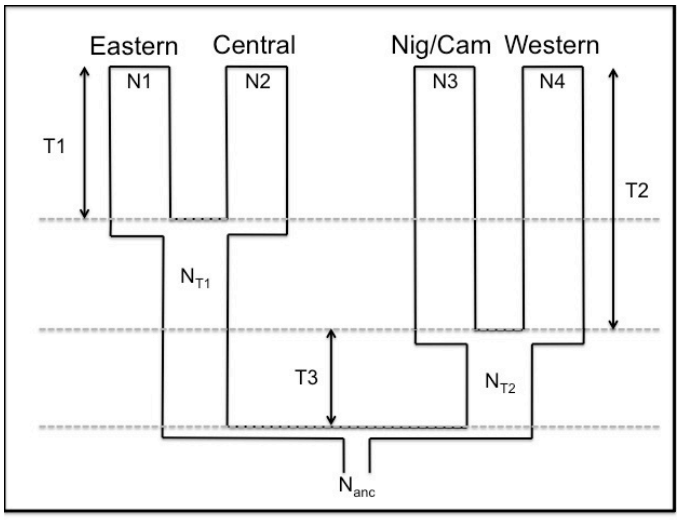
Parameter	HDPI 95% fit[a]	Posterior Estimation[b]				Real World Estimates[c]		
		Mode	Median	HDPI 95		Median	95% CI	
				Lower	Upper		Lower	Upper
log(N1)	0.94	4.01	4.02	3.03	5.02	10,528	1,069	104,814
log(N2)	0.98	5.06	4.86	4.00	5.40	72,001	9,986	251,189
log(N3)	0.98	3.67	3.78	3.01	4.59	6,033	1,022	38,511
log(N4)	0.99	3.76	3.76	3.11	4.39	5,710	1,283	24,739
$\log(N_{T1})$	0.93	4.71	4.48	3.31	5.40	30,241	2,041	251,189
$\log(N_{T2})$	0.91	4.60	4.21	3.08	5.28	16,275	1,200	188,395
$\log(N_{anc})$	0.96	4.02	3.97	3.12	4.72	9,240	1,319	52,786
log(T1)	0.98	4.52	4.25	3.26	5.00	358,517	36,416	2,000,000
log(T2)	0.95	4.60	4.19	3.21	5.00	309,557	32,452	2,000,000
log(T3)	0.87	4.64	4.18	3.20	5.00	301,655	31,567	2,000,000
$\log(M_{1\text{-}2})$	0.97	0.11	-0.11	-1.47	1.00	0.77	0.03	10.00
$\log(M_{1\text{-}3})$	0.96	-1.50	-1.12	-2.00	0.03	0.08	0.01	1.06
$\log(M_{1\text{-}4})$	0.93	-1.71	-1.49	-2.00	-0.68	0.03	0.01	0.21
$\log(M_{2\text{-}3})$	0.97	-0.65	-0.72	-1.98	0.44	0.19	0.01	2.76
$\log(M_{2\text{-}4})$	0.94	-1.57	-1.35	-2.00	-0.44	0.05	0.01	0.36
$\log(M_{3\text{-}4})$	0.97	-1.29	-1.26	-1.95	-0.58	0.06	0.01	0.26
$\log(M_{1,2\text{-}3})$	0.90	-1.31	-0.61	-1.96	0.77	0.25	0.01	5.95
$\log(M_{1,2\text{-}4})$	0.92	-1.41	-0.92	-2.00	0.57	0.12	0.01	3.72
$\log(M_{1\text{-}2,3\text{-}4})$	0.90	-1.22	-0.54	-1.88	0.89	0.29	0.01	7.74

Suppl. Table 12.3.5 - *Posterior estimates for Model 4A.*

Note. [a]A metric demonstrating how often known simulated values (n=1000) fell within the calculated 95% CI, which gives a guide to the reliability of these CI's for real data. [b]Calculated using 10PLS components, 1M simulations and retaining 1%. [c]Values converted from log10 scale to real world estimates, with divergence time assuming a 20-year generation time.

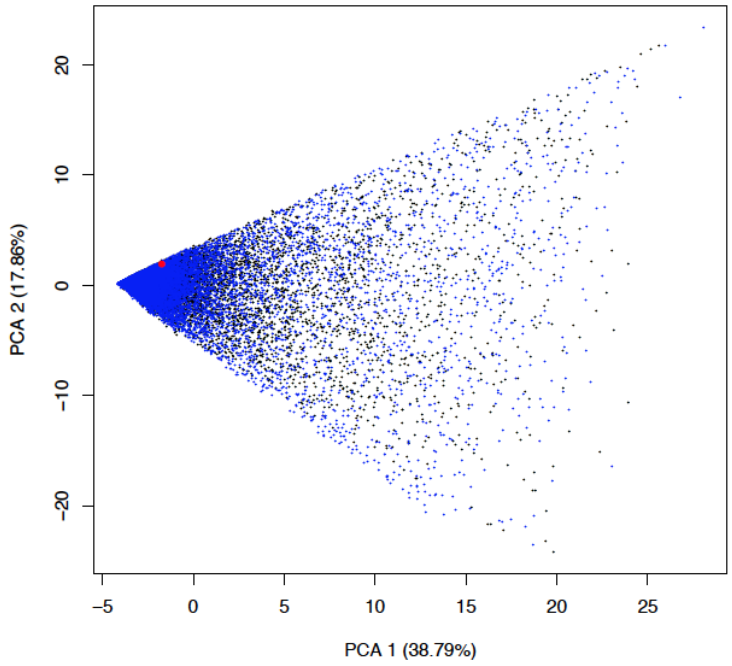


Model 1

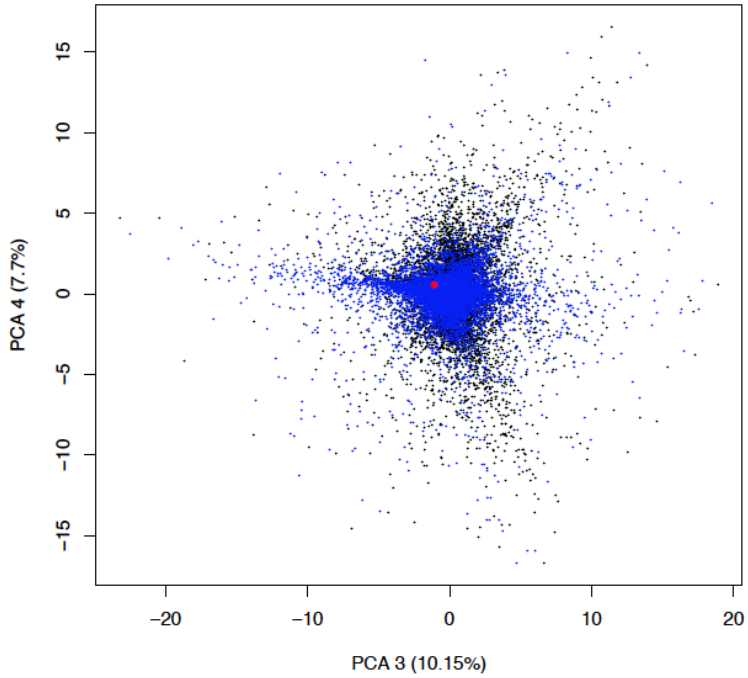


Model 2

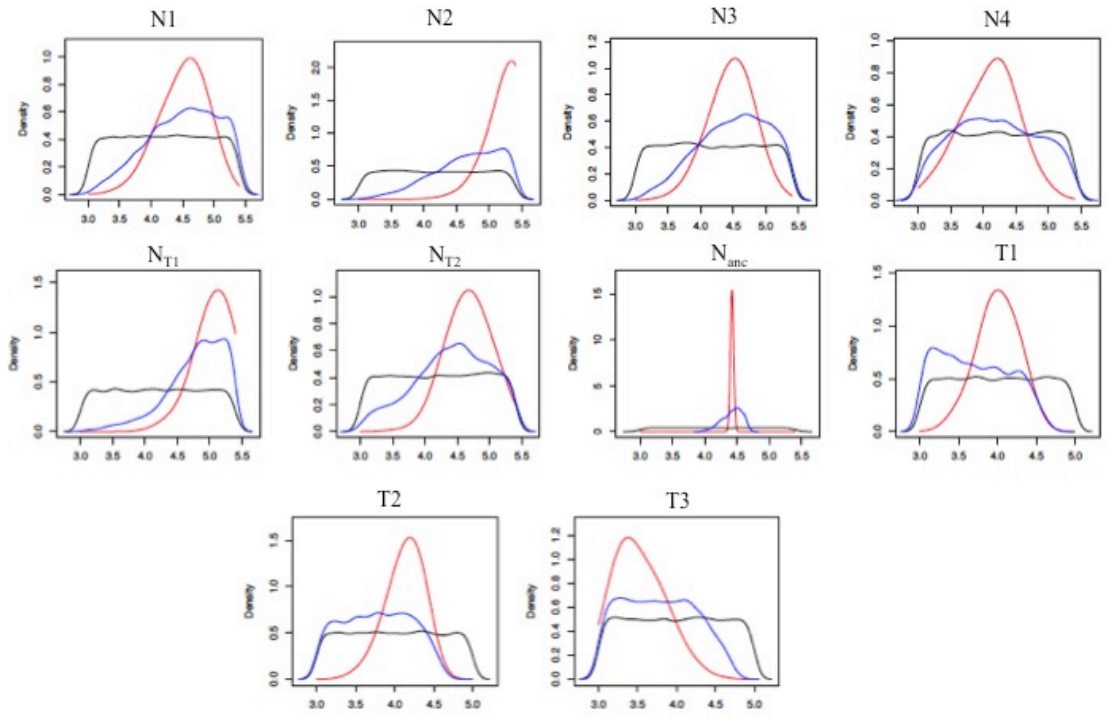
Suppl. Figure 12.3.1 - *Demographic Models 1 and 2 for P.t divergence without migration.*



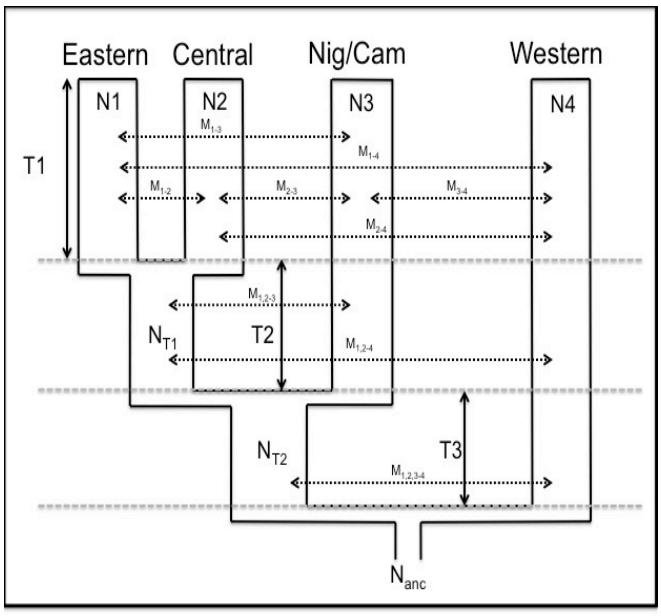
Suppl. Figure 12.3.2A - *PCA 1 and 2 of summary statistics for Model 1 (black), Model 2 (blue) and observed data (red).*



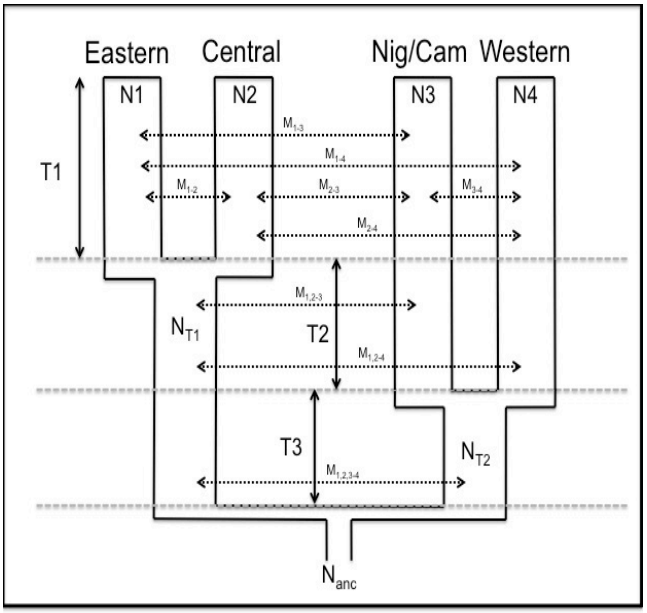
Suppl. Figure 12.3.2B - *PCA 3 and 4 of summary statistics for Model 1 (black), Model 2 (blue) and observed data (red).*



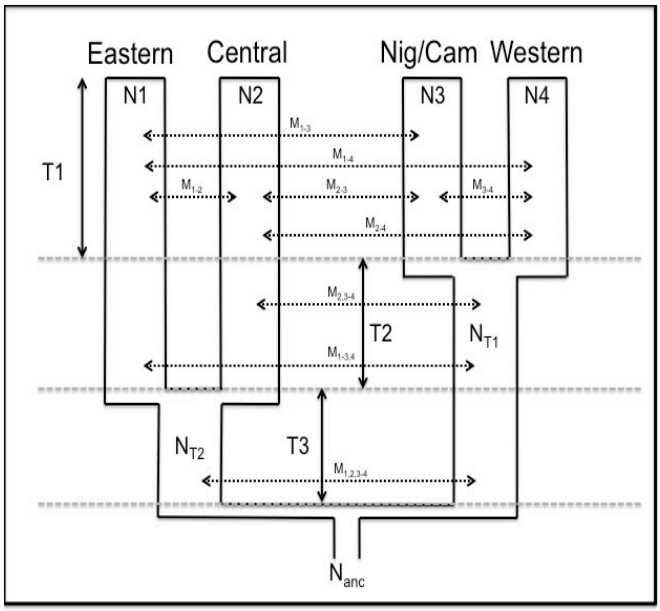
Suppl. Figure 12.3.3 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 2.*



Model 3

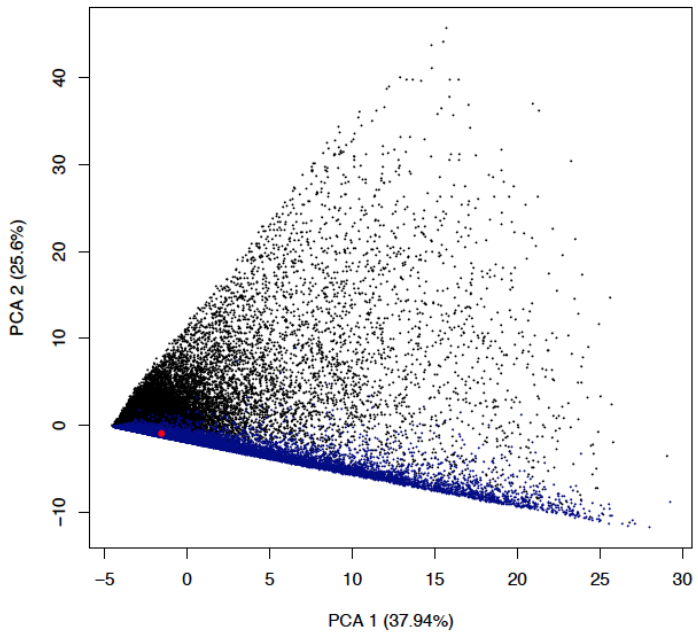


Model 4A

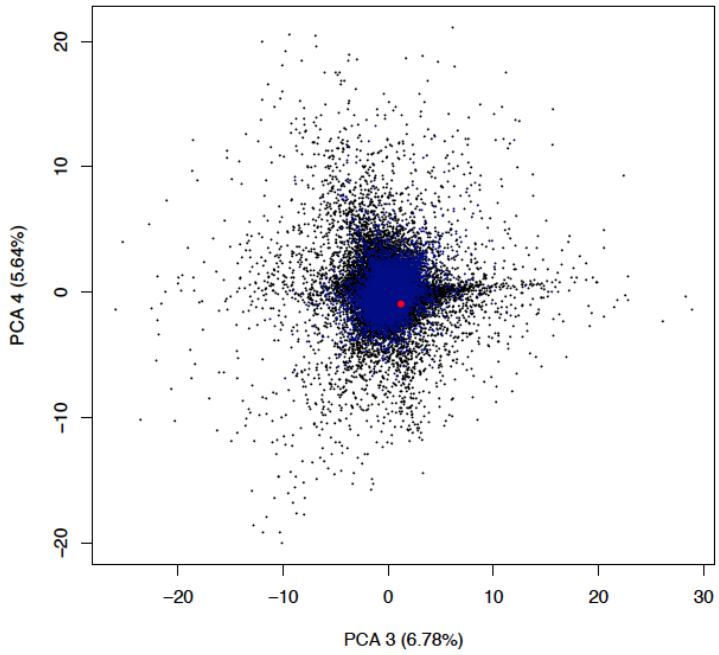


Model 4B

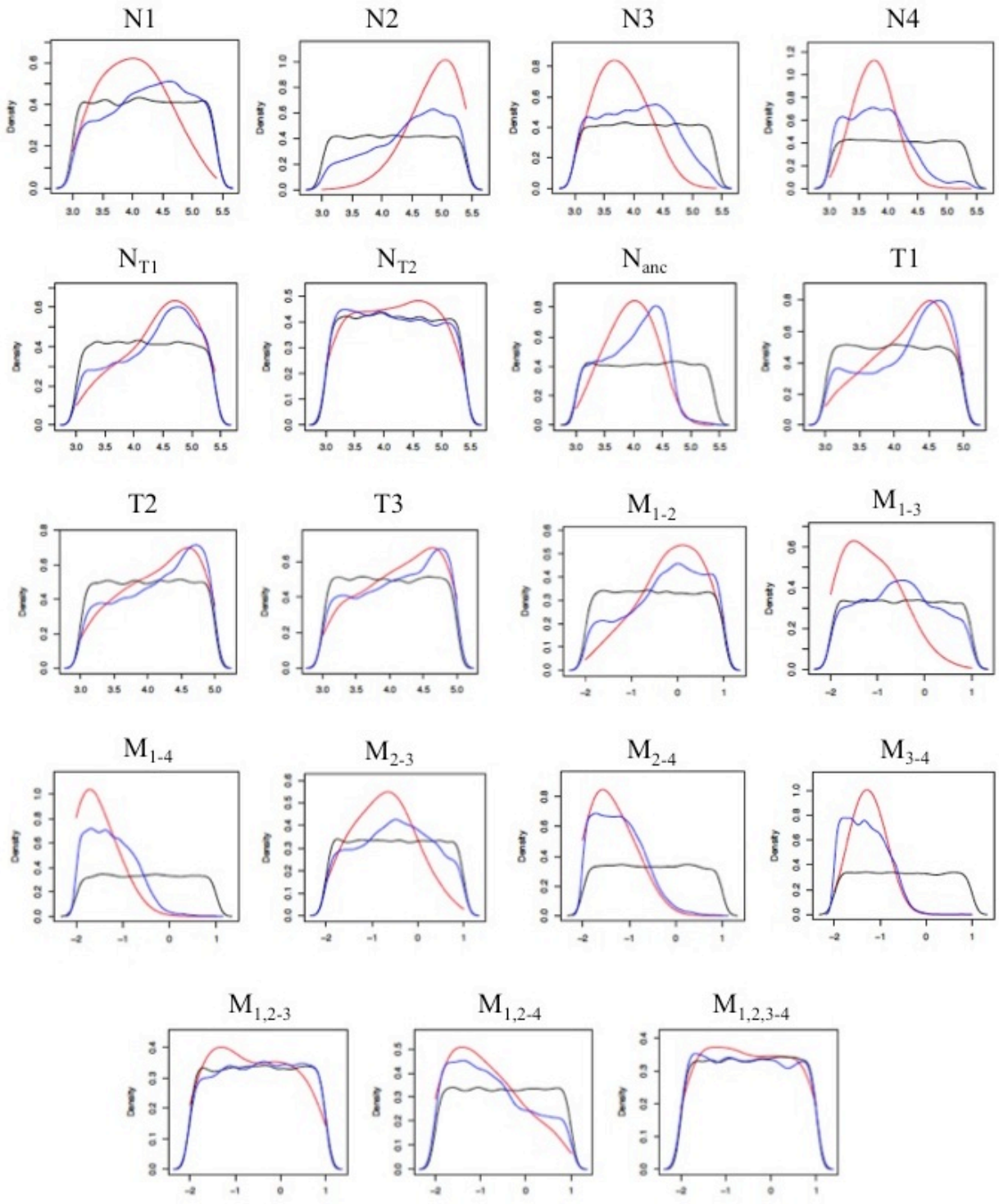
Suppl. Figure 12.3.4 - *Demographic Models 3, 4A and 4B for P.t divergence with symmetric migration.*



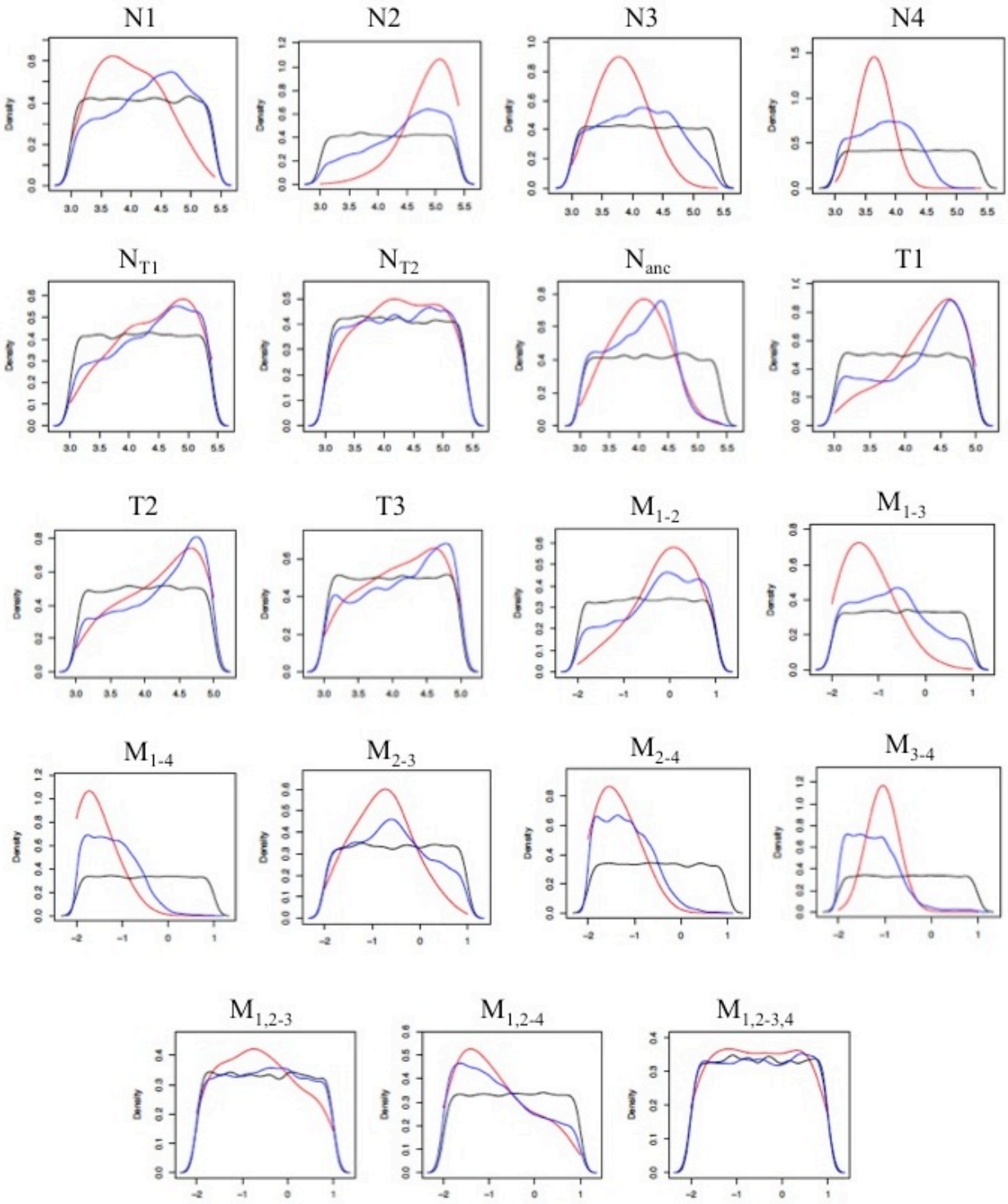
Suppl. Figure 12.3.5A - *PCA 1 and 2 of summary statistics for Models 1 and 2 (black), Model 3, 4A and 4B (blue) and observed data (red).*



Suppl. Figure 12.3.5B - *PCA 3 and 4 of summary statistics for Models 1 and 2 (black), Model 3, 4A and 4B (blue) and observed data (red).*



Suppl. Figure 12.3.6 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 3.*



Suppl. Figure 12.3.7 - *Prior (black), retained (blue) and GLM adjusted posterior distributions under Model 4A.*

12.4 PSMC

Heng Li, Jeffrey M. Kidd, Joanna L. Kelley, Carlos D. Bustamante, David Reich

Methods

1.1 Calling consensus sequence

We aligned great ape short reads to the hg18 human genome with BWA[2]. We called the consensus using SAMtools[741]. For the PSMC analysis[75], we selected samples from each subspecies (**Suppl. Table 12.4.1**) such that 1) each sample has relatively high read depth in each subspecies; 2) each sample is known to have a low contamination level and without evident hybridization according to the principle component analysis (PCA); and 3) if allowed, at least three samples, including one male sample, are chosen from each subspecies. A few samples were dropped after the PSMC analysis due to excessively large inferred population size, which is typically an indication of poor consensus calling or contamination.

1.2 Scaling population parameters

We use d, the number of substitutions per base between a pair of sequences (i.e., pairwise sequence divergence), to measure time, and use θ, the scaled mutation rate, to measure the effective population size. The advantage of such scaling is that both d and θ can be directly inferred from sequence data without using any additional scaling parameters that cannot be determined by a coalescent model. When we know the generation time g and the mutation rate per base per generation μ from other sources, $d/(2\mu/g)$ gives the time in years and $\theta/(4\mu)$ gives the effective population size. For primates, μ/g is typically ranged from 10^{-9} to 0.5×10^{-9} per base per year. This value may differ slightly across species and might have been changed over the past 20 million years[76].

1.3 Inferring population size history

We inferred the historical population size with PSMC[22]. We measured the variance of the estimate by bootstrapping: we selected one sample from each subspecies (**Suppl. Table 12.4.1**), split its consensus into 10 Mbp segments, randomly resampled about 300 segments with replacement, and then rerun PSMC on the resampled segments. We repeated the procedure 100 times. The fluctuation of the 100 rounds of inferences suggests the variance.

1.4 Inferring divergence time

---

[1] Command line: "SAMtools mpileup -Euf ref.fa aln.bam | bcftools view -c -| vcfutils.pl vcf2fq -d min depth -D max depth" where min depth is set to 1/3 of the average read depth and max depth set to twice of the average.

[2] Command line: "psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o result.psmc cns.psmcfa".

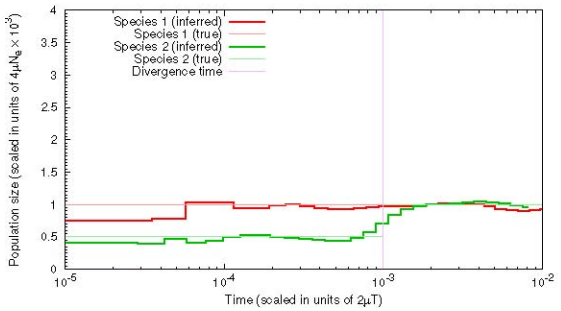
When we plot the PSMC inferences of two subspecies together, the time point where the two historical population sizes diverge approximates the divergence time. However, although this approach is intuitive and works apparently well, it has several problems. Firstly, the plot does not provide quantified time. Telling where sizes of two subspecies diverge is not always obvious and, at times, subjective. Secondly, it is possible that the two subspecies had the same size after the split, which will lead to an underestimate. Thirdly, there might be considerable gene flows between the two subspecies after the initial split. Differentiation in population sizes may not correspond to the final split. At last, PSMC has a known artifact where it may smooth out sudden size changes and push back the divergence time (**Suppl. Figure 12.4.1**). The PSMC plot only gives us a qualitative sense of the divergence time.

A second approach to infer divergence time is to hybridize two haploid sequences from each species and then run PSMC on the pseudo-diploid sequence. The time point where the inferred population goes to infinity corresponds to the divergence time. As the samples we use are diploid, we randomly select an allele at a heterozygous site to derive a haploid sequence. This approximation works well if the speciation time is much deeper than the average the most recent common ancestor (TMRCA) of each species such that most heterozygotes arose after the speciation. This method gives a stronger signal of speciation than the first method, but it does not quantify the time either.

Sample	Depth	Callable (Gbp)	Heterozygosity	Comment
P.t.ellioti-Koto	23.7	2.091	1.33×10^{-3}	
P.t.ellioti-Jean	21.0	2.076	1.27×10^{-3}	
P.t.ellioti-Taweh	20.8	2.088	1.34×10^{-3}	
P.t.ellioti-Damian	19.8	2.066	1.27×10^{-3}	
P.t.schweinfurthii-Kidongo	43.7	2.049	1.61×10^{-3}	
P.t.schweinfurthii-Bwambale	41.4	2.046	1.68×10^{-3}	
P.t.schweinfurthii-Nakuu	37.4	2.039	1.56×10^{-3}	
P.t.troglodytes-Vaillant	29.1	2.034	1.96×10^{-3}	
P.t.troglodytes-Doris	27.6	2.028	1.89×10^{-3}	
P.t.troglodytes-Julie	21.3	1.965	1.84×10^{-3}	Large recent population size
P.t.verus-Clint	32.6	2.049	0.84×10^{-3}	
P.t.verus-Jimmie	26.4	2.037	0.83×10^{-3}	Large recent population size
P.t.verus-Koby	18.0	1.935	0.77×10^{-3}	
P.t.verus-Bosco	16.4	2.080	0.78×10^{-3}	
P.paniscus-Dzeeta	36.2	2.023	0.78×10^{-3}	
P.paniscus-Desmond	35.6	2.022	0.82×10^{-3}	
P.paniscus-Hermien	33.8	2.032	0.76×10^{-3}	
G.b.graueri-Victoria	28.1	1.797	0.84×10^{-3}	
G.b.graueri-Kaisi	29.5	2.017	0.80×10^{-3}	Large recent population size
G.b.graueri-Mkubwa	15.8	2.054	0.76×10^{-3}	
G.g.gorilla-Delphi	30.9	2.011	1.72×10^{-3}	
G.g.gorilla-Amani	29.8	2.013	1.76×10^{-3}	
G.g.gorilla-Banjo	22.9	1.999	1.76×10^{-3}	
G.g.dielhi-Nyango	17.7	1.952	1.18×10^{-3}	
P.abelii-Dunja	31.1	1.961	2.42×10^{-3}	
P.abelii-Elsi	29.7	1.960	2.51×10^{-3}	
P.abelii-Babu	26.4	1.917	2.43×10^{-3}	
P.abelii-Buschi	26.0	1.944	2.44×10^{-3}	
P.abelii-Kiki	25.6	1.938	2.44×10^{-3}	
P.pygmaeus-Tilda	27.9	1.971	1.71×10^{-3}	
P.pygmaeus-Napoleon	27.4	1.973	1.64×10^{-3}	
P.pygmaeus-Sari	24.9	1.905	1.59×10^{-3}	

Suppl. Table 12.4.1 – *Samples used for PSMC analysis. The average read depth (second column) is estimated at HapMap3 sites. The third column gives the number of sites in hg18 where a genotype can be called confidently. A site is masked as 'uncalled' if at the site: 1) the read depth is more than twice or less than one-third of the average depth; 2) the site is within 5 bp around a predicted short indel; 3) the root-mean-square mapping quality is below 10; 4) the estimated consensus quality is below 30; and 5) fewer than 18 out of 35 overlapping 35-mer from hg18 can be mapped elsewhere with zero or one mismatch. Heterozygosity is estimated in callable regions only. PSMC bootstrapping is applied to samples in the bold font face*.

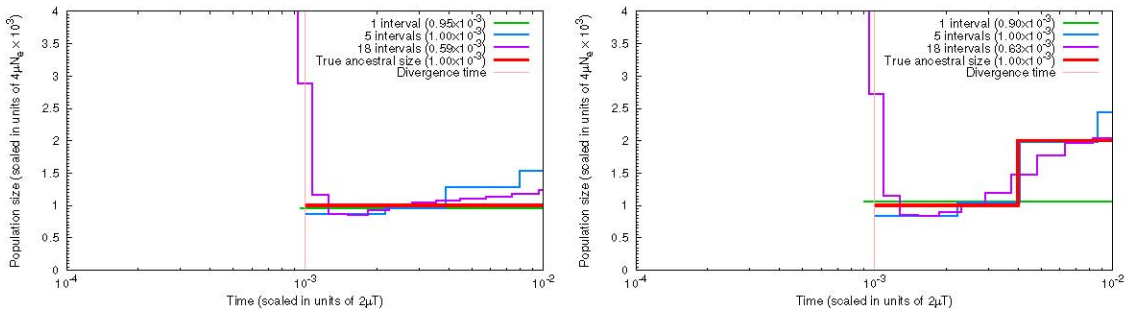
We will propose a third approach to quantify divergence time with a small modification to the original PSMC model; we assume no coalescences after divergence. The divergence time is just another parameter of the PSMC model that can be estimated together with population sizes. A caveat is that divergence has a similar effect to infinite population size. When we use many small time intervals, PSMC will be confused by the two scenarios and underestimate the time.



Suppl. Figure 12.4.1 – *PSMC inference given sudden population size changes. In simulation, species 1 keeps a constant population size, while species 2 has halved its population size immediately after the divergence at x = 10^{-3}. The ms command line in use is: "ms4500-t1000-r300 1000000 -I2 22 -en011-en02 0.5 -ej 0.5 2 1 -eN 0.500001 1".*

Our temporary solution is to use fewer size parameters to force the population size around divergence to be small. This is not an ideal solution, but it seems to work on simulated data.

In **Suppl. Figure 12.4.2**, we estimate the divergence time by fitting the ancestral population with 1, 5 or 18 time intervals[3]. With many intervals, PSMC infers excessively large population size around the speciation. Although this is also an indication of speciation, the time is underestimated. Using fewer intervals gives accurate estimates, especially when PSMC can estimate the ancestral population size changes well.



[3] The '-p' parameter used by psmc is set to '40', '20+4*5' and '6+17*2', respectively.

Suppl. Figure 12.4.2 – *Estimating divergence time on simulated data. Numbers in parentheses give the inferred divergence time. The thick red line shows the simulated ancestral population size.*

Results

2.1 Population size history and divergence time

Figure 3 shows the population size history for human and primates. Overall, samples from the same subspecies/population agree well with each other and the fluctuation between samples largely falls within the variance of a single sample. The ancestral population sizes of different species before speciation also match well. These are expected.

Figure 3 already hints at the divergence time between subspecies, but as we discussed earlier, the time is not quantitative and may be subjected to artifacts. To quantify the divergence time, we applied PSMC with divergence time as an extra parameter. **Suppl. Table 12.4.2** shows the inferred divergence time ds, the ancestral population size θs before the speciation. The standard deviation estimated by bootstrapping for the *P.t.verus*-Clint and *P.t.schweinfurthii*-Kidongo sample pair is 0.012×10^{-3}. The consistency between sample pairs from same species pairs also suggests a small variance predicted by the model.

It is worth noting that for a constant ancestral population with a clean speciation, we would expect dg = ds + θs[4]. This is not true in the table because: 1) we fit the ancestral population using a piece-wise constant function with five intervals; 2) segmental duplications in great apes are shared between samples and will inflate dg; and 3) a real history deviating from a simple speciation model will break the equality. If we simulate a constant ancestral population size and fit the ancestral population with one time interval, dg = ds + θs approximately stands.

[4]For a population with a constant effective population size Ns, the average coalescent time between two sequences from the population is 2Ns generations. If two species diverged Ts generations ago, the average coalescent time between two sequences from each species is Tg = Ts +2Ns. Multiplying 2μ to each side of the equation yields dg = ds + θs.

For most sample pairs, **Suppl. Table 12.4.2** broadly agrees with **Figure 3**, with a few exceptions. Firstly and most strikingly, P.t.verus and P.t.ellioti diverged more recently than what we see from **Figure 3**. This implies considerable genetic exchanges between the two subspecies after the initial split. Secondly, P.t.verus and P.t.ellioti seem to be closer to P.t.schweinfurthii than to P.t.troglodytes, but this is not obvious from **Figure 3**. This observation might suggest more gene flow between Western and Eastern chimpanzees than between Western and Central several hundred thousand years ago. Thirdly, PSMC predicts G.b.graueri to be closer to G.g.diehli than to G.g.gorilla. Similarly this may imply unbalanced gene flows.

To formally test unbalanced gene flows between species, we performed a D-statistic test (**Suppl. Table 12.4.3**). This test also suggests that Western chimpanzees are genetically closer to Eastern chimpanzees than to Central and that Eastern lowland gorillas are closer to Cross River gorillas than to Western lowland gorillas. The test prefers to put Western and Nigeria-Cameroon chimpanzees in one clade, though it also implies the latter is closer to Central and Eastern chimpanzees.

2.2 Comparison to the previous studies

When we compare the divergence time from different studies, a major complication is time scaling. There are typically two approaches to time scaling. The first is to assume a fixed mutation rate μ and generation time g and to use the two parameters to scale time to years.

Sample 1	Sample 2	d_g ($\times 10^{-3}$)	d_s ($\times 10^{-3}$)	θ_s ($\times 10^{-3}$)
P.paniscus-Dzeeta	P.t.troglodytes-Vaillant	3.82	1.75	1.08
P.paniscus-Dzeeta	P.t.schweinfurthii-Kidongo	3.84	1.77	1.09
P.paniscus-Dzeeta	P.t.verus-Clint	3.85	1.76	1.07
P.paniscus-Dzeeta	P.t.ellioti-Koto	3.83	1.77	1.08
P.t.verus-Clint	P.t.schweinfurthii-Kidongo	2.14	0.81	0.87
P.t.verus-Clint	P.t.schweinfurthii-Bwambale	2.14	0.83	0.85
P.t.verus-Clint	P.t.schweinfurthii-Nakuu	2.14	0.81	0.86
P.t.verus-Clint	P.t.troglodytes-Vaillant	2.17	0.92	0.77
P.t.verus-Clint	P.t.troglodytes-Doris	2.20	0.94	0.76
P.t.verus-Clint	P.t.ellioti-Koto	1.78	0.47	1.00
P.t.ellioti-Koto	P.t.schweinfurthii-Kidongo	2.06	0.78	0.88
P.t.ellioti-Koto	P.t.troglodytes-Vaillant	2.11	0.90	0.77
P.t.schweinfurthii-Kidongo	P.t.troglodytes-Vaillant	2.02	0.75	0.85
G.b.graueri-Mkubwa	G.g.gorilla-Delphi	2.22	0.30	1.54
G.b.graueri-Mkubwa	G.g.dielhi-Nyango	2.09	0.22	1.45
G.g.gorilla-Delphi	G.g.dielhi-Nyango	1.86	0.16	1.33
P.abelii-Elsi	P.pygmaeus-Tilda	3.33	0.97	1.81

Suppl. Table 12.4.2 – *Inferred speciation time and ancestral population size. For each sample, a pseudo-haploid sequence was derived by choosing a random allele at a heterozygote. dg is the average sequence divergence between a pair of pseudo-haploid sequences from two samples, excluding uncalled regions in either sample. ds is the PSMC inferred speciation time and θs is the inferred scaled mutation rate (proportional to the effective population size) right before the speciation. ds and θs were estimated from a pseudo-diploid sequence generated by hybridizing two haploid*

sequences from different samples. Five population size parameters were used to fit the ancestral population size history.

For time scaled this way, it is easy to convert the time back to sequence divergence ds. The second approach to time scaling is to scale the inferred time by the human-chimpanzee sequence divergence. If we assume the average human-chimpanzee genetic divergence to be 7 million years ago (mya) or so, we can derive time in years. The latter method aims to account for variable mutation rates in different regions. However, due to the large ancestral population size, the human-chimpanzee sequence divergence may vary greatly around 7 mya. The effectiveness of the second approach might be debatable. In addition, for the second approach, it is still possible to scale time back to sequence divergence ds if the divergence between human and chimpanzee is given.

A second complication is that when the speciation is not clean, it may not be straightforward to precisely define the divergence time. A population model may fold other factors, such as present and ancestral population sizes, migration, and structure into divergence time. For example, a model without considering migration will prefer more recent divergence time in comparison to an isolation-migration model. The divergence time estimate from different studies may not be strictly comparable. **Suppl. Table 12.4.4** shows the speciation time between pairs of subspecies in the previous studies[55–57,77–79]. Our PSMC estimates tend to be close to the majority.

Sample A	Sample B	Sample X	$D(A, B, X;$ human$)$
P.t.v-Clint	P.t.e-Koto	P.p-Dzeeta	1.58
P.t.v-Clint	P.t.t-Vaillant		1.45
P.t.v-Clint	P.t.s-Kidongo		-1.34
P.t.e-Koto	P.t.t-Vaillant		0.61
P.t.e-Koto	P.t.s-Kidongo		-2.29
P.t.t-Vaillant	P.t.s-Kidongo		-3.13
P.t.e-Koto	P.t.t-Vaillant	P.t.v-Clint	-116.71
P.t.e-Koto	P.t.s-Kidongo		-100.22
P.t.t-Vaillant	P.t.s-Kidongo		16.94
P.t.t-Vaillant	P.t.s-Kidongo	P.t.e-Koto	20.53
P.t.v-Clint	P.t.e-Koto	P.t.t-Vaillant	21.79
P.t.v-Clint	P.t.e-Koto	P.t.s-Kidongo	25.28
G.g.g-Delphi	G.g.d-Nyango	G.b.g-Mkubwa	8.23

Suppl. Table 12.4.3 – *D-statistic for four haploid sequences, A, B, X and O, a site is classified as ABBA if at the site base A = O B = X, or classified as BABA if B = O= A = X. Define D'(A, B, X; O)== (#ABBA −#BABA)/(#ABBA +#BABA). D(A, B, X; O) equals the ratio of the mean of D' to its standard deviation, estimated by block Jack-knife. A positive D value indicates that sample B is genetically closer to X, while a negative vale indicates A closer to X.*

Species 1	Species 2	d_s ($\times 10^{-3}$)	μ/g ($\times 10^{-9}$)	Reference
P.paniscus	P.t.schweinfurthii	1.57	1.00	Becquet and Przeworski (2007)
P.paniscus	P.t.verus	1.75	1.00	Becquet and Przeworski (2007)
P.paniscus	P.t.verus	1.75	1.02	Won and Hey (2005)
P.paniscus	P.t.troglodytes	1.75		This study
P.paniscus	P.t.verus	1.76		This study
P.paniscus	P.t.ellioti	1.77		This study
P.paniscus	P.t.schweinfurthii	1.77		This study
P.paniscus	P.troglodytes	1.83	0.71	Caswell et al. (2008)
P.paniscus	P.t.troglodytes	1.84	1.00	Becquet and Przeworski (2007)
P.paniscus	P.troglodytes	1.98	1.00	Prüfer et al. (2012)
P.paniscus	P.troglodytes	2.56	0.80	Wegmann and Excoffier (2010)
P.t.verus	P.t.schweinfurthii	0.56	1.00	Becquet and Przeworski (2007)
P.t.verus	P.t.schwein./troglo.	0.72	0.71	Caswell et al. (2008)
P.t.verus	P.t.schweinfurthii	0.81		This study
P.t.verus	P.t.troglodytes	0.88	1.00	Becquet and Przeworski (2007)
P.t.verus	P.t.schwein./troglo.	0.88	0.80	Wegmann and Excoffier (2010)
P.t.verus	P.t.troglodytes	0.92		This study
P.t.verus	P.t.troglodytes	0.92	1.08	Won and Hey (2005)
P.t.troglodytes	P.t.schweinfurthii	0.44	1.00	Becquet and Przeworski (2007)
P.t.troglodytes	P.t.schweinfurthii	0.70	0.80	Wegmann and Excoffier (2010)
P.t.troglodytes	P.t.schweinfurthii	0.78		This study
G.beringei	G.gorilla	0.15	0.96	Thalmann et al. (2007)
G.b.graueri	G.g.dielhi	0.22		This study
G.beringei	G.gorilla	0.24	1.33	Becquet and Przeworski (2007)
G.b.graueri	G.g.gorilla	0.30		This study
G.beringei	G.gorilla	0.60	0.60	Scally et al. (2012)
P.abelii	P.pygmaeus	0.74	1.00	Mailund et al. (2011)
P.abelii	P.pygmaeus	0.89	1.00	Locke et al. (2011)
P.abelii	P.pygmaeus	0.97		This study
P.abelii	P.pygmaeus	2.78	1.00	Becquet and Przeworski (2007)

Suppl. Table 12.4.4 – *Speciation time in the literature. The speciation time in all the previous studies is converted to sequence divergence ds where possible. μ/g is the per-base per-year mutation rate used in the corresponding study. dHC is the average human-chimpanzee sequence divergence.*

Comparison of PSMC Results

To assess the effect of reference sequence divergence on our analysis, we applied the PSMC method using the mappings to each species reference. The mappings were processed using BWA, Picard, and GATK as described above. We made diploid consensus sequence calls using SAMtools and ran PSMC as previously described (with psmc -N 25 -t 15 -r 5 -p "4+25*2+4+6"). To avoid potential artifacts, we limited analysis to a subset of samples with high coverage and showing low evidence of potential contamination. These results recapitulate the basic patterns observed from analysis using hg18 (see **Suppl. Figure 12.4.3**) both in terms of relative effective population sizes among subspecies and apparent times in which demographic trajectories diverge. We note, however, that for the orangutans, mapping to the orangutan genome shows a qualitative difference in the inferred effective population sizes for *Pongo abelii*.

Suppl. Figure 12.4.3 – *PSMC analysis based on mappings to species reference genome assemblies. We limited analysis to a subset of samples with high coverage and low evidence of contamination including eight orangutans (A), four gorillas (B), and four chimpanzees and one bonobo (C).*

Section 13: Analysis of demography and incomplete lineage sorting (ILS)

13.1 Incomplete lineage sorting (ILS)

Kasper Munch, Thomas Mailund, Mikkel H. Schierup

The coalescent hidden Markov model (CoalHMM)

A consequence of ILS is that segments of a genomic alignment have evolutionary relationships different from the species tree. The CoalHMM framework[80] allows for inference of population genetic parameters and patterns of ILS and the model is based on a hidden Markov chain with hidden states representing gene trees with separate topologies and separate coalescent times. The model is thus able to represent incomplete lineage sorting along a genomic alignment. The model applied is a three species isolation model (**Suppl. Figure 13.1**) with the following demographic parameters: two ancestral population sizes, Ne_1, Ne_2, and two speciation times, T_1, T_2. These parameters are all scaled with the substitution rate.

Suppl. Figure 13.1 – *Isolation model used in the analysis. T_1: speciation time of species X and Y. T_2: speciation time of species Y and Z. Ne_1: effective population size of the population size ancestral to species X and Y. Ne_2: effective population size of the ancestor to all three species.*

The CoalHMM operates with four different trees connecting three species: species X and Y may find a common ancestor in their ancestral population (**Suppl. Figure 13.2** top left) or in the population ancestral to all three species (**Suppl. Figure 13.2** top right), and species Z may find a common ancestor with either X or Y in the population ancestral to all three species (**Suppl. Figure 13.2** bottom left and right).

The model is applied in turn to genomic alignments of four species of apes of which one only serves as outgroup. Individuals used are listed in **Suppl. Table 13.1** and combinations of individuals used in each analysis are listed in **Suppl. Table 13.2**.

Suppl. Figure 13.2 – *The four hidden states in the HMM. The four states correspond to the four different trees describing the ancestry of an alignment column. Arrows indicate possible transitions.*

P. tro. tro. I	*Pan troglodytes_troglodytes* A959 Julie
P. tro. tro. II	*Pan troglodytes_troglodytes* A960 Clara

P. tro. sch.	*Pan troglodytes_schweinfurthii* 9729 Harriet
P. tro. ver.	*Pan troglodytes_verus* 9668 Bosco
P. tro. eli.	*Pan troglodytes_ellioti* Koto
P. pan.	*Pan paniscus* A915 Kosana
H. sap.	*Homo sapiens* San HGDP01029
G. gor. gor.	*Gorilla gorilla gorilla* A933 Dian
G. gor. ber.	*Gorilla beringei_graueri* Victoria
P. pan. abe.	*Pongo abelii* A950 Babu

Suppl. Table 13.1 – *Key for individuals included in analyses.*

1	P. tro. tro. I	P. pan	H. sap.	P. pan. abe.
2	P. tro. tro. II,	P. pan	H. sap.	P. pan. abe.
3	P. tro. sch.,	P. pan.	H. sap.	P. pan. abe.
4	P. tro. ver.,	P. pan	H. sap.	P. pan. abe.
5	P. tro. eli.,	P. pan	H. sap.	P. pan. abe.
6	P. tro. tro. I	H. sap.	G. gor. gor	P. pan. abe.
7	P. tro. tro. I	H. sap.	G. gor. ber.	P. pan. abe.

Suppl. Table 13.2 – *Combinations of individuals used in analyses.*

Preparation of input alignments

To generate input alignments we map all called SNPs to the human reference genome (NCBI Build 36). Bases not called are substituted with N to indicate missing data. This results in a genomic sequence in human coordinates for each individual. These sequences are further masked using the RepeatMasker track from the UCSC genome browser. Genomic sequences from four individuals from different species are chosen. The implicit alignment that these form is filtered removing consecutive runs of more than 100 alignment columns of all N characters splitting alignment accordingly. To aid the computation, an upper bound on the length of consecutive alignment of 100,000 bases is imposed and the alignment is split accordingly. An individual CoalHMM analysis is performed on ~1 Mbp of such alignment to estimate demographic model parameters and proportions of ILS.

Estimates of model parameters

The distributions of speciation times estimated on individual 1 Mbp alignments are shown in **Suppl. Figure 13.3**. Estimates for individual species combinations are listed in **Suppl. Table 13.3-13.6** along with confidence intervals and standard error of the mean. CoalHMM measures time in mutations rather than in years but the estimated model parameters can be rescaled to years using a per-year mutation rate and a generation time. Here we have used a per-year mutation rate of 0.6e-9 per site[64] and a generation time of 25 years. All parameters scale linearly with mutation rate and population sizes scale linearly with generation time, making it

straightforward to obtain split times and population sizes from an alternative choice of rescaling.

We found a general agreement between analyses when including different subspecies of chimpanzee and gorilla, but also when including chimpanzee, human and gorilla compared to chimpanzee, bonobo and human. The chimpanzee-bonobo speciation time is slightly lower, but significantly, for those analyses including Central chimpanzee than for the analyses including Nigeria-Cameroon, Western or Eastern chimpanzee. The p-value of a Kruskal-Wallis test on estimates from all analyses is 3.4e-12, whereas it is 0.94 and 0.76 when performing the test on the analyses including and excluding Central chimpanzee, respectively. This would be in line with a notion that bonobos split from a population ancestral to Central chimpanzees.

Suppl. Figure 13.3 – *Distributions of estimated split times (five outliers not shown).*

Chimpanzee subsp.	**Split time**	**1.96 * SE of mean**
P. tro. eli.	1,285,736	23,507
P. tro. ver.	1,284,132	23,564
P. tro. sch.	1,294,088	23,681
P. tro. tro. II	1,204,658	23,539
P. tro. tro. I	1,203,596	23,622

Suppl. Table 13.3 – *Chimpanzee-bonobo speciation time estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Chimpanzee subsp.	**Split time**	**1.96 * SE of mean**
P. tro. eli.	5,244,082	53,789
P. tro. ver.	5,262,635	55,058
P. tro. sch.	5,235,109	54,378
P. tro. tro. II	5,205,198	49,153
P. tro. tro. I	5,201,148	46,426

Suppl. Table 13.4 – *Chimpanzee-human speciation time estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Gorilla subsp.		
G. gor. ber.	5,684,228	62,533
G. gor. gor.	5,648,467	61,730

Suppl. Table 13.5 – *Chimpanzee-human speciation time estimated from chimpanzee, human, gorilla, orangutan analyses.*

Gorilla subsp.	Split time	1.96 * SE of mean
G. gor. ber.	8,669,901	65,141
G. gor. gor.	8,644,655	62,319

Suppl. Table 13.6 – *Chimpanzee-gorilla speciation time estimated from chimpanzee, human, gorilla, and orangutan analyses.*

Ancestral population sizes

The distributions of estimated ancestral population sizes on individual 1 Mbp of alignment are shown in **Suppl. Figure 13.4**. Estimates for individual species combinations are listed in **Suppl. Tables 13.7-13.10** along with standard error of the mean.

Suppl. Figure 13.4 – *Distributions of estimated effective population sizes.*

Chimpanzee subsp.	Ne	1.96 * SE of mean
P. tro. eli.	32,000	525
P. tro. ver.	32,189	515
P. tro. sch.	32,036	542
P. tro. tro. II	31,532	501
P. tro. tro. I	31,573	477

Suppl. Table 13.7 – *Chimpanzee-bonobo population size estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Chimpanzee subsp.	Ne	1.96 * SE of mean
P. tro. eli.	83,133	1,868
P. tro. ver.	83,015	1,798
P. tro. sch.	83,413	1,843
P. tro. tro. II	83,400	1,860
P. tro. tro. I	84,032	1,858

Suppl. Table 13.8 – *Chimpanzee-Human population size estimated from chimpanzee, bonobo, human, and orangutan analyses.*

Gorilla subsp.	Ne	1.96 * SE of mean
G. gor. ber.	86,230	1,804
G. gor. gor.	86,967	1,747

Suppl. Table 13.9 – *Chimpanzee-Human population size estimated from chimpanzee, human, gorilla, orangutan analyses*.

Gorilla subsp.	Ne	1.96 * SE of mean
G. gor. ber.	67,123	1,274
G. gor. gor.	66,747	1,267

Suppl. Table 13.10 – *Chimpanzee-Gorilla population size estimated from chimpanzee, human, gorilla, orangutan analyses*.

Proportions of ILS

The most likely hidden state is assigned to each column of analyzed alignment using posterior decoding of the optimized hidden Markov model (HMM). The proportion of each type of ILS is calculated in 1 Mbp windows in human reference coordinates. Only windows covering at least 500 Kbp of alignment are included. The distribution of inferred proportions of ILS is shown in **Suppl. Figure 13.5** with summary statistics listed in **Suppl. Table 13.11**. The two types of ILS, ((x,z),y) and ((y,x),x), should be in equal proportions. In the two chimpanzee-human-gorilla analyses, however, the estimated proportion of ILS where chimpanzee and gorilla coalesce first is larger than the proportion of ILS with human and gorilla coalescing first. CoalHMM assumes an ultrametric tree and a violation of this assumption will produce such an effect—in this case if the human branch is longer. SNPs are called by mapping reads to the human reference genome and this may potentially result in higher sensitivity in calling human SNP, thus producing the observed effect.

The estimated proportions agree well with the theoretically expected proportion that is readily calculated from rescaled estimates as exp[-(T2-T1)/(25*2*Ne1)]. This yields 2.9% for chimpanzee-bonobo-human ILS assuming a chimpanzee-bonobo speciation at 1.3 mya, a chimpanzee-human speciation at 5.2 mya, and a chimpanzee-bonobo ancestral population size of 32,000. The proportion of chimpanzee-human-gorilla ILS is expected to be 16.6% assuming a chimpanzee-human speciation at 5.6 mya, a human-gorilla speciation at 8.6 mya, and a chimpanzee-human ancestral population size of 86,000. Note that the theoretically expected proportions of ILS are not dependent on the mutation rate and generation time used for rescaling. The expected proportions are shown in **Suppl. Figure 13.5** as gray lines.

Suppl. Figure 13.5 – *Distributions of the proportion of ILS in 1 Mbp windows in human reference coordinates.*

Species: ((x,y),z)	% ((x,z),y)	% ((y,z),x)
((P. tro. eli., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. ver., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. sch., P. pan), H. sap)	2.7 -/+ 0.001	2.5 -/+ 0.001
((P. tro. tro. I,P. pan), H. sap)	2.7 -/+ 0.001	2.2 -/+ 0.001
((P. tro. tro. II, P. pan), H. sap)	2.7 -/+ 0.001	2.2 -/+ 0.001
((P. tro. tro. I, H. sap.), G. gor. ber.)	17.5 -/+ 0.004	14.7 -/+ 0.004
((P. tro. tro. I, H. sap.), G. gor. gor.)	17.6 -/+ 0.004	14.8 -/+ 0.004

Suppl. Table 13.11 – *Proportion of ILS in 1 Mbp windows in human reference coordinates. Confidence interval is calculated as -/+ 1.96 * standard error of mean.*

13.2 Isolation model CoalHMM

Kasper Munch, Thomas Mailund, Anders E. Halager, Mikkel H. Schierup

Model

The coalescence with recombination is a model of the local genealogy along a sample of genomes. For two haploid genomes, it models how the time to TMRCA changes as one scans along a genome alignment.

Suppl. Figure 13.6 – *Example ancestral recombination graph for a sequence segment from two different species. Recombinations can occur at any point in time but coalescence can only occur once in the common ancestor, i.e., further back in time than the split tau1. Right graph shows TMRCA along the sequence.*

The distribution of TMRCA, and the distribution of segment lengths sharing TMRCA, is determined by the split time between the populations/species the genomes are taken from, the effective population sizes in the two populations and the ancestral population, and the recombination rate along the genomes (**Suppl. Figure 13.6**). By modeling the changes in TMRCA along the genome alignment in an HMM, we can thus infer these parameters[79].

Data preparation

We obtained genomes from each individual as in **section 12.** We then translated these diploid genomes into haploid genomes by, for each heterozygotic site, picking an allele at random. This means that the phase of the haploid genomes changes as we scan along the genome, but if the level of shared polymorphism between species is low (see simulations[79] supplemental text S1), it does not change TMRCA between the genomes, which is what the method uses for its inference.

We then further filtered the data by running a sliding window along the genome, with a window size of 1 Kbp and a jump of 100 bp, removing windows with more than 10% missing data, since we found that regions with missing data display a greater variance in divergence than regions without missing data. We analyzed the genome in chunks of 10 Mbp of alignment, with the parameters of the HMM reinitiated whenever a gap in the alignment has been introduced by the data filtering.

Parameter inference

We applied the method[79] with three parameters: a coalescence rate C, a split time tau, and a recombination rate R, all measured in a time scale of substitutions per base pair. This means that the model assumes that the effective population size in the two extant species is the same as in the ancestral population, but as previously

shown[81], this means that the model will estimate the ancestral coalescence rate rather than the extant rates, since this is the most important rate for the likelihood of the model.

If the mutation rate per year, u, is known, and the generation time, g, is known, these parameters can be rescaled since tau/u is then the split time in years, $2/C = 4N\ u\ g$, and $R/(u\ g)$ is the recombination rate per base pair per generation.

Results

With the filtering described above, >95% of 10 Mbp fragments converge to reasonable values except for the two most ancient split events, i.e., between human and gorilla and between human and orangutan. In these cases we have removed obvious outliers and report on the remaining fragments.

Suppl. Figure 13.7 – *Coalescence rates C=1/4Nu and split times (x 1e3) for different species comparisons, with two different sets of individuals compared in each case (blue and red). A. Eastern and Western gorilla, B. Bornean and Sumatran orangutan, C. Chimpanzee and bonobo, D. Human and bonobo, E. Human and gorilla, F. Human and orangutan.*

From the results of **Suppl. Figure 13.7**, we have calculated the mean (with standard error of the mean, SEM) for each comparison (**Suppl. Table 13.12**. These numbers can be translated into years and population sizes if we make assumptions about the generation time and the mutation rate per year. We show results assuming a generation time of 25 years (20 years for the comparison of gorillas) and three different mutation rates, either 1e-9 or 0.6e-9 per year or 1.1e-8 per generation.

Suppl. Table 13.12 – *Coalescence rates C=1/4Nu and split times (x 1e3) for different species comparisons with standard errors of the mean (SEM). Effective population sizes of ancestral species and split times are shown for two different mutation rates per year and a generation time of 25 years (20 years when comparing the two gorillas). Mutation rate calibrations are 1e-9 and 0.6e-9 per year and 1.1e-8 per generation.*

References

1. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–6 (2012).

2. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

3. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–8 (2011).

4. Consortium, C. S. and A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).

5. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).

6. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).

7. Mckenna, A. *et al.* The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).

8. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Research* **11**, 1725–1729 (2001).

9. Browning, B. L. & Yu, Z. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *The American Journal of Human Genetics* **85**, 847–861 (2009).

10. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* **26**, 2867–73 (2010).

11. Bowden, R. *et al.* Genomic Tools for Evolution and Conservation in the Chimpanzee : Pan troglodytes ellioti Is a Genetically Distinct Population. *PLoS Genetics* **8**, 1–10 (2012).

12. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41**, 1061–1067 (2009).

13. Donmez, N. & Brudno, M. Hapsembler : An Assembler for Highly Polymorphic Genomes. *Research in Computational Molecular Biology* 38–52 (2011).doi:10.1007/978-3-642-20036-6_5

14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

15. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511–518 (2005).

16. Fischer, A. *et al.* Bonobos Fall within the Genomic Variation of Chimpanzees. *PLoS ONE* **6**, e21605 (2011).

17. Behar, D. M. *et al.* A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. *The American Journal of Human Genetics* **90**, 675–684 (2012).

18. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).

19. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: Analytical and study design considerations. *Genetic epidemiology* **28**, 289–301 (2005).

20. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2**, e190 (2006).

21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).

22. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).

23. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* **5**, e1000519 (2009).

24. Auton, A. *et al.* A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science* **336**, 193–198 (2012).

25. Mcquillan, R. *et al.* Runs of Homozygosity in European Populations. *American journal of human genetics* **83**, 359–372 (2008).

26. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *American journal of human genetics* **91**, 275–92 (2012).

27. Liengola, H. J. and & I Post-Conflict Inventory of Kahuzi-Biega National Park,. *Gorilla Journal* **30**, 3–5 (2005).

28. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164 (2010).

29. Wetterbom, A., Gyllensten, U., Cavelier, L. & Bergström, T. F. Genome-wide analysis of chimpanzee genes with premature termination codons. *BMC genomics* **10**, 56 (2009).

30. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* 1–6 (2012).doi:10.1038/nature11128

31. Hahn, Y. & Lee, B. Human-specific nonsense mutations identified by genome sequence comparisons. *Human genetics* **119**, 169–78 (2006).

32. Hahn, Y. & Lee, B. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics* **21 Suppl 1**, i186–94 (2005).

33. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19**, 1270–1278 (2009).

34. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Research* **21**, 1640–1649 (2011).

35. Olson, M. V When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics* **64**, 18–23 (1999).

36. Wang, X., Grus, W. E. & Zhang, J. Gene Losses during Human Origins. *PLoS Biology* **4**, e52 (2006).

37. Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877–81 (2009).

38. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–8 (2011).

39. Hammer, M. F. *et al.* The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature genetics* **42**, 830–1 (2010).

40. Gottipati, S., Arbiza, L., Siepel, A., Clark, A. G. & Keinan, A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature genetics* **43**, 741–3 (2011).

41. Hernandez, R. D. *et al.* Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science* **331**, 920–924 (2011).

42. Pool, J. E. & Nielsen, R. Population size changes reshape genomic patterns of diversity. *Evolution: International Journal of Organic Evolution* **61**, 3001–3006 (2007).

43. Keinan, A. & Reich, D. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Molecular Biology and Evolution* **27**, 2312–2321 (2010).

44. Wakeley, J. Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology* **49**, 369–386 (1996).

45. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–96 (2001).

46. Hey, J. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Molecular biology and evolution* **27**, 921–33 (2010).

47. Wegmann, D., Leuenberger, C. & Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**, 1207–18 (2009).

48. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics* **5**, e1000695 (2009).

49. Lukic, S. & Hey, J. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192**, 619–39 (2012).

50. Aeschbacher, S., Futschik, A. & Beaumont, M. A. Approximate Bayesian computation for modular inference problems with many parameters: the example of migration rates. *Molecular ecology* **22**, 987–1002 (2013).

51. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* **28**, 2239–52 (2011).

52. Green, R. R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science (New York, N.Y.)* **328**, 710–722 (2010).

53. Eriksson, A. & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13956–60 (2012).

54. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* **8**, e1002967 (2012).

55. Won, Y.-J. & Hey, J. Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* **22**, 297–307 (2005).

56. Becquet, C. & Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Research* **17**, 1505–1519 (2007).

57. Wegmann, D. & Excoffier, L. Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* **27**, 1425–1435 (2010).

58. Gonder, M. K. *et al.* Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *PNAS* **108**, 4766–4771 (2011).

59. Degnan, J. H. & Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS genetics* **2**, e68 (2006).

60. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).

61. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* **43**, 1031–4 (2011).

62. Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–24 (2008).

63. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science (New York, N.Y.)* **328**, 636–9 (2010).

64. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–5 (2012).

65. Wakeley, J. & Hey, J. Estimating ancestral population parameters. *Genetics* **145**, 847–55 (1997).

66. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–56 (2003).

67. Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC bioinformatics* **11**, 116 (2010).

68. Leuenberger, C. & Wegmann, D. Bayesian computation and model selection without likelihoods. *Genetics* **184**, 243–52 (2010).

69. Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. S. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15112–7 (2011).

70. Fagundes, N. J. R. *et al.* Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17614–9 (2007).

71. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* **16**, 1791–8 (1999).

72. Veeramah, K. R. *et al.* An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Molecular biology and evolution* **29**, 617–30 (2012).

73. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)* **18**, 337–8 (2002).

74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

75. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–6 (2011).

76. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745–53 (2012).

77. Caswell, J. L. *et al.* Analysis of chimpanzee history based on genome sequence alignments. *PLoS genetics* **4**, e1000057 (2008).

78. Thalmann, O., Fischer, A., Lankester, F., Pääbo, S. & Vigilant, L. The complex evolutionary history of gorillas: insights from genomic data. *Molecular Biology and Evolution* **24**, 146–158 (2007).

79. Mailund, T. *et al.* A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS genetics* **8**, e1003125 (2012).

80. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS genetics* **3**, e7 (2007).

81. Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G. & Schierup, M. H. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genetics* **7**, 15 (2011).

Appendix C - Supplementary Material for Chapter 4

Supplementary Material

Evolution and diversity of copy number variation in the great ape lineage

Peter H. Sudmant, John Huddleston, Claudia R. Catacchio, Maika Malig, LaDeana W. Hillier, Carl Baker, Kiana Mohajeri, Ivanela Kondova, Ronald E. Bontrop, Stephan Persengiev, Francesca Antonacci, Mario Ventura, Javier Prado-Martinez, Tomas Marques-Bonet, and Evan E. Eichler

Additional resources available online:
http://eichlerlab.gs.washington.edu/greatape-cnv

Section 1: Genome mapping and quality control 318
Section 2: Nonhuman primate reference genome comparisons 325
Section 3: Copy number variant calling methods 327
Section 4: Validations 332
Section 5: Lineage-specific great ape segmental duplications and deletions 336
Section 6: Lineage-specific deletions and diversity of sequence not represented in the human reference genome 344
Section 7: Gene duplication analysis 362
Section 8: Distribution of duplication and deletion events 379
Section 9: Rates of segmental duplication and deletion and underlying genes 390
Section 10: Copy number variation and duplication diversity 395
Section 11: A human genomic disorder identified in a nonhuman primate 406
Section 12: References 318

Section 1: Genome mapping and quality control

Genome mapping for read-depth-based copy number analysis: Genomes were mapped to the human reference assembly Build 36 (UCSC HG18) using the mrsfastc mapping software(Sudmant et al. 2010; Hach et al. 2010). Reads were first divided into their 36 bp constituents and mapped with a maximum edit distance of 2 to a repeat-masked reference. Subdividing reads into their 36 bp constituents served firstly to allow us to align reads with an ungapped alignment algorithm (mrsfastc) quickly. Additionally, subdividing reads served to normalize the varied read lengths between genomes. Masking was performed with RepeatMasker (default UCSC masking version 3 Repeat Masking) and Tandem Repeats Finder(Marques-Bonet et al. 2009; Benson 1999; Cheng et al. 2005) using the following parameters: *match 2, mismatch 3, delta 5, PM 80, PI 10, minscore 30, maxperiod 1000*. Masked regions were extended out by 36 bp on either side after mapping to eliminate mapping edge effects adjacent to masked regions.

Read-depth-based GC correction and copy number prediction: Copy numbers were estimated in tiled 500 bp, 1 kbp, and 3 kbp windows of unmasked sequence across the genome of each individual as described in Sudmant *et al* 2010(1000 Genomes Project Consortium et al. 2012; Sudmant et al. 2010). First, a GC correction step was performed to eliminate biases in sequence coverage introduced during library construction and associated with the GC content of loci. For each individual a multiplicative GC correction factor was calculated and applied to the genome. This correction factor was calculated by determining the average read-depth across putative invariant diploid regions of the human genome (potentially copy number variable loci subtracted from HG18, namely, the Database of Genomic Variants, gaps, segmental duplications (SDs), and copy number variants identified by Conrad *et al*(Meyer et al. 2012; Conrad et al. 2010) binned by GC content as computed in 401 base-pair windows across the genome. A correction factor k_{GC} was then calculated as $k_{GC}=\mu_{total}/\mu_{GC}$ where μ_{total} is the total average read-depth and μ_{GC} is the read-depth within a particular GC content bin. The corrected read-depth at a base *x* where *d(x)* is the read-depth at base *x* is then calculated as $d'(x) = d(x)*k_{GC}$ where k_{GC} is chosen to match the particular GC content at base *x.* Copy numbers were then estimated by using a linear regression model fit to regions of known copy number 2 in all primates.

Genome-wide copy number 2 assessment: Read-depth-based copy number estimates are based on the assumption that the number of reads shotgun sequenced over a particular locus will be directly proportional to the copy number of that locus. As such, extreme biases introduced during sequencing and library construction or resulting from DNA degradation may reduce the power of this technique even after corrections are applied. To test the quality of each of our genomes, we focused on 4836 regions encompassing ~1.1 Gbp of sequence, which we predict to be largely devoid of any structural variation and thus fixed at copy number 2. These loci were selected by removing from the human reference genome regions of known or likely structural variation, including the Database of Genomic Variation, genomic gaps, SDs, and variants detected by Conrad *et al*(Miller et al. 1993; Conrad et al. 2010). We then further eliminated any regions <100 kbp in length to avoid over-fragmenting. The resulting set of loci should largely be fixed at copy number 2 and invariant among all human genomes. Among different species of nonhuman primates, some these loci may indeed be copy number variable. However, we can assume that not only will the number of such variable sites will be low but that among individuals of the same species the same number of regions will be variable and therefore the total number of invariant loci within a species should be consistent. We thus estimated the copy number of these loci in tiled 3 kbp windows of unmasked sequence. 131242 such windows were considered within each individual genome (**Figure 1.1**). Additionally, the proportion of loci correctly predicted as copy number 2 as a function of the GC content of a region was also quantified to assess the effect of library construction induced GC biases (**Figure 1.2**).

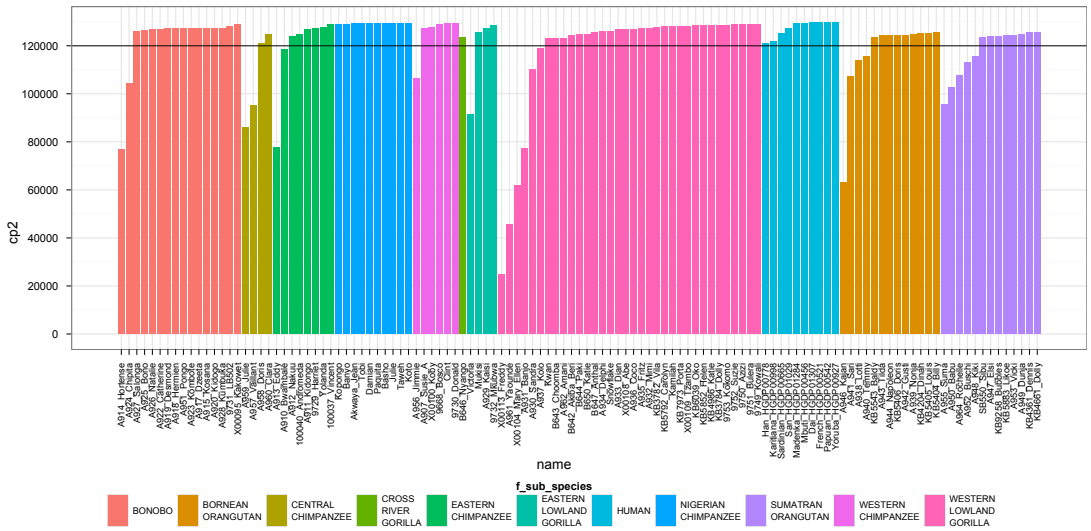


Figure 1.1: 131,242 3 kbp windows encompassing of 1.1 Gbp of putative diploid invariant sequence were analyzed in each sequenced genome. The number of 3 kbp windows correctly determined as copy number 2 is displayed for each genome sorted and colored by species and subspecies. The black line indicates a cutoff of 120,000, below which corresponds to genomes with poor copy number estimation across the entire spectrum of GC content. Genomes failing this cutoff are plotted with dotted lines in Figure 1.2 below.

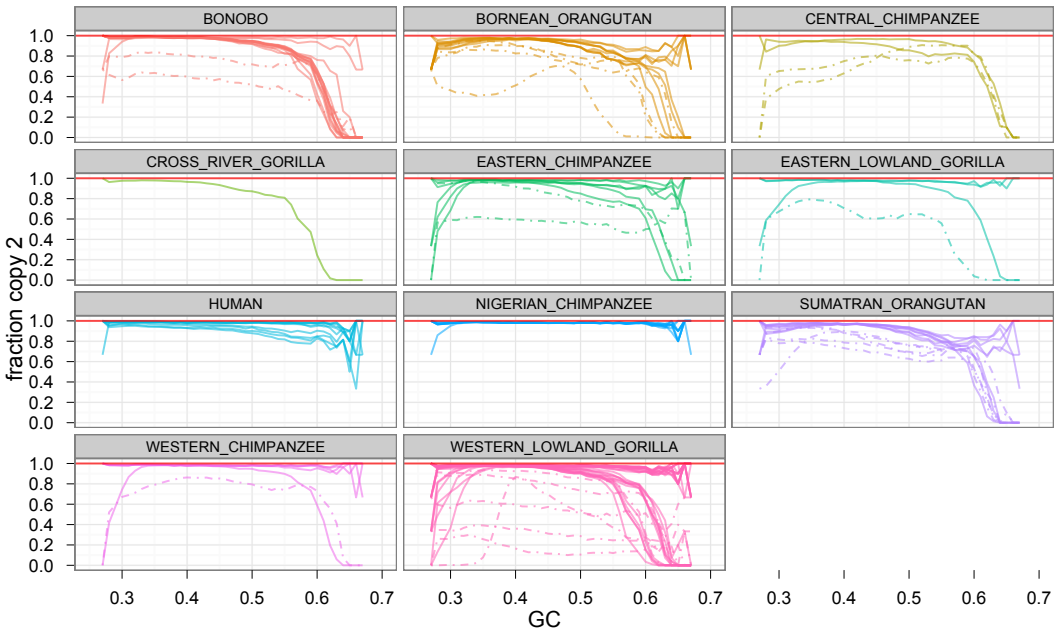


Figure 1.2: The fraction of 3 kbp windows correctly determined to be copy number 2 is plotted as a function of GC content grouped and colored by subspecies. Regions of increased GC content are expected to have reduced coverage and increased variance and, thus, are more difficult to accurately assay. Reduced power to accurately estimate copy number across the entire spectrum of GC content is, however, an indication of a genome that has been subject to multiple confounding factors during

sequencing or poor starting DNA quality. Genomes exhibiting a trend of reduced copy number prediction power across the entire spectrum of GC content were discarded and are plotted with dotted lines. These genomes correspond to those falling below the cutoff threshold indicated in Figure 1.1 and are listed in Table 1.1.

Table 1.1: Individuals filtered out by copy 2 region analyses (23).

name	sub_species	fraction of diploid regions correctly estimated
A914_Hortense	BONOBO	0.59
A924_Chipita	BONOBO	0.80
A959_Julie	CENTRAL_CHIMPANZEE	0.66
A957_Vaillant	CENTRAL_CHIMPANZEE	0.73
A913_Eddy	EASTERN_CHIMPANZEE	0.59
A910_Bwambale	EASTERN_CHIMPANZEE	0.90
A956_Jimmie	WESTERN_CHIMPANZEE	0.81
Victoria	EASTERN_LOWLAND_GORILLA	0.70
X00113_Freddy	WESTERN_LOWLAND_GORILLA	0.19
A937_Kolo	WESTERN_LOWLAND_GORILLA	0.90
X00104_Mary_Ellen	WESTERN_LOWLAND_GORILLA	0.47
A931_Banjo	WESTERN_LOWLAND_GORILLA	0.59
A961_Yaounde	WESTERN_LOWLAND_GORILLA	0.35
A930_Sandra	WESTERN_LOWLAND_GORILLA	0.84
A941_Sari	BORNEAN_ORANGUTAN	0.82
A938_Lotti	BORNEAN_ORANGUTAN	0.87
A940_Temmy	BORNEAN_ORANGUTAN	0.88
A946_Kajan	BORNEAN_ORANGUTAN	0.48
A952_Buschi	SUMATRAN_ORANGUTAN	0.86
A948_Kiki	SUMATRAN_ORANGUTAN	0.88
A955_Suma	SUMATRAN_ORANGUTAN	0.73
A964_Rochelle	SUMATRAN_ORANGUTAN	0.82
A950_Babu	SUMATRAN_ORANGUTAN	0.78

We previously reported that at regions of increased GC content, Illumina-sequenced genomes often demonstrate reduced and more variable depth-of-coverage(Olson 1999; Sudmant et al. 2010; Alkan et al. 2009). To a certain degree this can be corrected for and additionally only a small fraction of the human genome lies within these high GC regions; >90% of the human genome is at <50% GC **(Figure 1.3)**. 4.2%, 1.7% and 0.056% of the genome are at >55%, >60% and >65% GC content, respectively. Our analysis of those 1.1 Gbp of putative copy number 2 loci among all of the genomes indicated, however, that a small fraction of these genomes displayed reduced power to accurately estimate copy number across the entire spectrum of GC content **(Figure 1.2** dashed lines, **Table 1.1)**. The genomes demonstrating this trend corresponded to those for which the fewest number of copy number 2 loci were accurately assessed **(Figure 1.1)** and thus a cutoff of correctly estimating the copy of 120,000/131,242 windows (91.4%) was selected to flag any genomes of reduced quality due to processes in the library construction and sequencing or perhaps in the quality of the DNA sequenced. These genomes (n=23) were discarded from the read-depth-based analyses of copy number and SDs. Additionally, regions of extreme GC content (>0.57% GC corresponding to 69,119,920 bp, 2.23%) were masked from the analysis to avoid regions of potential biased sequencing.

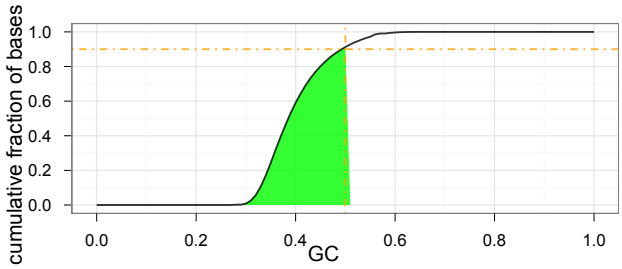


Figure 1.3: A curve of the fraction of base pairs represented in the human genome cumulatively as a function of GC content. >90% of the human genome is represented at 50% GC or less as exhibited in the shaded green portion below the curve.

Genomes passing quality-control dataset: In total, 97 genomes were assessed in this study, including 10 human genomes and the genome of an archaic hominid Denisovan individual (referred to as Homo Denisova here). 75 of the individuals assessed here were sequenced as part of the Great Ape Genome Project—of which 13 showed low levels of contamination (Prado/Sudmant 2012, under review, **Table 1.2**). Ten of the individuals demonstrated <2% contamination while the remaining three showed <4% contamination. Our read-depth-based copy number prediction methodology is highly robust to such low levels of contamination as read-depth typically fluctuates with a mean and variance proportional to the underlying copy number and sequencing coverage of the sample(Brawand et al. 2011; Scally et al. 2012; Sudmant et al. 2010; Ventura et al. 2011; Meyer et al. 2012; Locke et al. 2011). To ensure no biases were introduced, all analyses which may be sensitive to cross-species contamination were performed/confirmed in the subset of non-contaminated individuals.

Table 1.2: Individuals assessed in this study. In addition to 75 genomes sequenced as part of the Great Ape Diversity project, 22 previously published samples were analyzed(Siepel et al. 2005; Scally et al. 2012; Ventura et al. 2011; Meyer et al. 2012; Locke et al. 2011)**.**

Species/sub-species	individual	Source	estimated % contamination
Gorilla beringei graueri	9732_Mkubwa	GAGP	
Gorilla beringei graueri	A929_Kaisi	GAGP	0.2
Gorilla beringei graueri	Mukisi	GAGP	
Gorilla gorilla diehli	B646_Nyango	GAGP	
Gorilla gorilla gorilla	9749_Kowali	GAGP	
Gorilla gorilla gorilla	9750_Azizi	GAGP	
Gorilla gorilla gorilla	9751_Bulera	GAGP	
Gorilla gorilla gorilla	9752_Suzie	GAGP	
Gorilla gorilla gorilla	9753_Kokomo	GAGP	
Gorilla gorilla gorilla	A932_Mimi	GAGP	0.093
Gorilla gorilla gorilla	A933_Dian	GAGP	
Gorilla gorilla gorilla	A934_Delphi	GAGP	0.06
Gorilla gorilla gorilla	A935_Fritz	GAGP	3.8
Gorilla gorilla gorilla	A936_Coco	GAGP	
Gorilla gorilla gorilla	A962_Amani	GAGP	
Gorilla gorilla gorilla	B642_Akiba_Beri	GAGP	
Gorilla gorilla gorilla	B643_Choomba	GAGP	
Gorilla gorilla gorilla	B644_Paki	GAGP	
Gorilla gorilla gorilla	B647_Anthal	GAGP	
Gorilla gorilla gorilla	B650_Katie	GAGP	
Gorilla gorilla gorilla	Kamilah	Scally *et al, 2012*	

Gorilla gorilla gorilla	KB3782_Vila	GAGP	
Gorilla gorilla gorilla	KB3784_Dolly	GAGP	
Gorilla gorilla gorilla	KB4986_Katie	GAGP	
Gorilla gorilla gorilla	KB5792_Carolyn	GAGP	
Gorilla gorilla gorilla	KB5852_Helen	GAGP	
Gorilla gorilla gorilla	KB6039_Oko	GAGP	
Gorilla gorilla gorilla	KB7973_Porta	GAGP	
Gorilla gorilla gorilla	Kwan	Ventura *et al, 2011*	
Gorilla gorilla gorilla	Snowflake	GAGP	
Gorilla gorilla gorilla	X00108_Abe	GAGP	
Gorilla gorilla gorilla	X00109_Tzambo	GAGP	
Homo denisova	Denisova_30x	Meyer *et al,* 2012	
Homo sapiens	Dai_HGDP01307	Meyer *et al,* 2012	
Homo sapiens	French_HGDP00521	Meyer *et al,* 2012	
Homo sapiens	Han_HGDP00778	Meyer *et al,* 2012	
Homo sapiens	Karitiana_HGDP00998	Meyer *et al,* 2012	
Homo sapiens	Madenka_HGDP01284	Meyer *et al,* 2012	
Homo sapiens	Mbuti_HGDP00456	Meyer *et al,* 2012	
Homo sapiens	Papuan_HGDP00542	Meyer *et al,* 2012	
Homo sapiens	San_HGDP01029	Meyer *et al,* 2012	
Homo sapiens	Sardinian_HGDP00665	Meyer *et al,* 2012	
Homo sapiens	Yoruba_HGDP00927	Meyer *et al,* 2012	
Pan paniscus	9731_LB502	GAGP	
Pan paniscus	A915_Kosana	GAGP	
Pan paniscus	A917_Dzeeta	GAGP	
Pan paniscus	A918_Hermien	GAGP	
Pan paniscus	A919_Desmond	GAGP	
Pan paniscus	A920_Kidogo	GAGP	
Pan paniscus	A922_Catherine	GAGP	
Pan paniscus	A923_Kombote	GAGP	
Pan paniscus	A925_Bono	GAGP	
Pan paniscus	A926_Natalie	GAGP	
Pan paniscus	A927_Salonga	GAGP	
Pan paniscus	A928_Kumbuka	GAGP	0.6
Pan paniscus	A951_Pongo	GAGP	
Pan paniscus	X00095_Kakowet	GAGP	0.018
Pan troglodytes ellioti	Akwaya_Jean	GAGP	
Pan troglodytes ellioti	Banyo	GAGP	
Pan troglodytes ellioti	Basho	GAGP	
Pan troglodytes ellioti	Damian	GAGP	
Pan troglodytes ellioti	Julie	GAGP	
Pan troglodytes ellioti	Kopongo	GAGP	
Pan troglodytes ellioti	Koto	GAGP	
Pan troglodytes ellioti	Paquita	GAGP	
Pan troglodytes ellioti	Taweh	GAGP	
Pan troglodytes ellioti	Tobi	GAGP	
Pan troglodytes schweinfurthii	100037_Vincent	GAGP	
Pan troglodytes schweinfurthii	100040_Andromeda	GAGP	0.8
Pan troglodytes schweinfurthii	9729_Harriet	GAGP	
Pan troglodytes schweinfurthii	A911_Kidongo	GAGP	
Pan troglodytes schweinfurthii	A912_Nakuu	GAGP	0.8
Pan troglodytes schweinfurthii	Yolanda	GAGP	2
Pan troglodytes troglodytes	A958_Doris	GAGP	
Pan troglodytes troglodytes	A960_Clara	GAGP	
Pan troglodytes verus	9668_Bosco	GAGP	
Pan troglodytes verus	9730_Donald	GAGP	
Pan troglodytes verus	A907_Susie_A	GAGP	3.2
Pan troglodytes verus	Clint	GAGP	
Pan troglodytes verus	X00100_Koby	GAGP	
Pongo abelii	A947_Elsi	GAGP	0.9
Pongo abelii	A949_Dunja	GAGP	
Pongo abelii	A953_Vicki	GAGP	3.8
Pongo abelii	KB4361_Dennis	Locke *et al*, 2011	
Pongo abelii	KB4661_Dolly	Locke *et al*, 2011	
Pongo abelii	KB5883_Likoe	Locke *et al*, 2011	
Pongo abelii	KB9258_Bubbles	Locke *et al*, 2011	
Pongo abelii	SB550_Sibu	GAGP	
Pongo pygmaeus	A939_Nonja	GAGP	
Pongo pygmaeus	A942_Gusti	GAGP	1.2
Pongo pygmaeus	A943_Tilda	GAGP	

Pongo pygmaeus	A944_Napoleon	GAGP
Pongo pygmaeus	KB4204_Dinah	Locke *et al*, 2011
Pongo pygmaeus	KB5404_Billy	Locke *et al*, 2011
Pongo pygmaeus	KB5405_Louis	Locke *et al*, 2011
Pongo pygmaeus	KB5406_Doris	Locke *et al*, 2011
Pongo pygmaeus	KB5543_Baldy	Locke *et al*, 2011

Table 1.3: Coverage of individuals assessed in this study. Coverage calculations are shown for bwa and mrsfast kmer mappings.

indiv	bwa coverage	kmer-coverage
Denisova_30x	-	23.464668
Gorilla_beringei_graueri-9732_Mkubwa	15.22	7.5223
Gorilla_beringei_graueri-A929_Kaisi	27.095	15.252012
Gorilla_gorilla_dielhi-B646_Nyango	19.3176	14.538121
Gorilla_gorilla_gorilla-B642_Akiba_Beri	17.2405	12.863963
Gorilla_gorilla_gorilla-B643_Choomba	18.8362	13.976078
Gorilla_gorilla_gorilla-B644_Paki	18.7551	13.912575
Gorilla_gorilla_gorilla-B647_Anthal	17.8673	13.23363
Gorilla_gorilla_gorilla-B650_Katie	16.4672	12.118194
Gorilla_gorilla_gorilla-9749_Kowali	16.1971	11.872634
Gorilla_gorilla_gorilla-9750_Azizi	16.0982	9.904
Gorilla_gorilla_gorilla-9751_Bulera	16.1769	11.951256
Gorilla_gorilla_gorilla-9752_Suzie	16.4371	12.025414
Gorilla_gorilla_gorilla-9753_Kokomo	13.4676	9.667402
Gorilla_gorilla_gorilla-A932_Mimi	24.7175	1.833891
Gorilla_gorilla_gorilla-A933_Dian	26.2103	22.277507
Gorilla_gorilla_gorilla-A934_Delphi	30.2397	23.113322
Gorilla_gorilla_gorilla-A935_Fritz	26.1051	7.291716
Gorilla_gorilla_gorilla-A936_Coco	22.2133	16.636503
Gorilla_gorilla_gorilla-A962_Amani	33.9349	23.197861
Gorilla_gorilla_gorilla-Snowflake	18.9214	12.456781
Gorilla_gorilla_gorilla-X00108_Abe	15.6194	8.953561
Gorilla_gorilla_gorilla-X00109_Tzambo	18.5012	11.487812
Homo_sapiens-Dai_HGDP01307	19.2118	9.009638
Homo_sapiens-French_HGDP00521	23.5054	12.040849
Homo_sapiens-Han_HGDP00778	-	10.267159
Homo_sapiens-Karitiana_HGDP00998	17.6082	7.576881
Homo_sapiens-Madenka_HGDP01284	22.5465	11.475714
Homo_sapiens-Mbuti_HGDP00456	15.8098	8.238722
Homo_sapiens-Papuan_HGDP00542	16.2915	8.487436
Homo_sapiens-San_HGDP01029	32.8205	14.390751
Homo_sapiens-Sardinian_HGDP00665	21.9033	9.473815
Homo_sapiens-Yoruba_HGDP00927	28.5036	14.729003
Pan_paniscus-9731_LB502	9.66735	7.051574
Pan_paniscus-A915_Kosana	37.5078	28.006446
Pan_paniscus-A917_Dzeeta	39.4451	30.510164
Pan_paniscus-A918_Hermien	38.5	28.862005
Pan_paniscus-A919_Desmond	39.6459	24.840106
Pan_paniscus-A920_Kidogo	26.732	16.105185
Pan_paniscus-A922_Catherine	23.424	17.85244
Pan_paniscus-A923_Kombote	27.2995	26.391115
Pan_paniscus-A925_Bono	27.0355	16.973132
Pan_paniscus-A926_Natalie	28.0149	26.687373
Pan_paniscus-A927_Salonga	22.9217	14.590245
Pan_paniscus-A928_Kumbuka	30.7078	26.832489
Pan_paniscus-A951_Pongo	32.7053	24.546877
Pan_paniscus-X00095_Kakowet	18.2247	11.423614
Pan_troglodytes_ellioti-Akwaya_Jean	22.4677	5.745486
Pan_troglodytes_ellioti-Banyo	9.91445	6.682296
Pan_troglodytes_ellioti-Basho	14.8194	6.699436
Pan_troglodytes_ellioti-Damian	20.5135	4.298133
Pan_troglodytes_ellioti-Julie	22.3339	8.427398
Pan_troglodytes_ellioti-Koto	23.8682	5.709182
Pan_troglodytes_ellioti-Paquita	13.532	9.440522
Pan_troglodytes_ellioti-Taweh	20.6315	4.507583
Pan_troglodytes_ellioti-Tobi	-	9.009664
Pan_troglodytes_ellioti-Kopongo	-	7.946675
Pan_troglodytes_schweinfurthii-100037_Vincent	21.339	13.587553

Pan_troglodytes_schweinfurthii-100040_Andromeda	20.0961	15.853668
Pan_troglodytes_schweinfurthii-A911_Kidongo	47.1899	34.386432
Pan_troglodytes_schweinfurthii-A912_Nakuu	39.6525	26.502671
Pan_troglodytes_schweinfurthii-Yolanda	25.8635	28.46547
Pan_troglodytes_schweinfurthii-9729_Harriet	11.0701	8.636543
Pan_troglodytes_troglodytes-A958_Doris	35.7141	21.921671
Pan_troglodytes_troglodytes-A960_Clara	23.9787	16.498607
Pan_troglodytes_verus-9730_Donald	19.2662	9.404114
Pan_troglodytes_verus-A907_Susie_A	28.076	21.334752
Pan_troglodytes_verus-9668_Bosco	15.6902	7.808031
Pan_troglodytes_verus-Clint	33.7102	16.566894
Pan_troglodytes_verus-X00100_Koby	17.8774	10.910666
Pongo_abelii-A947_Elsi	31.4027	22.223994
Pongo_abelii-A949_Dunja	34.0396	23.289702
Pongo_abelii-A953_Vicki	29.8808	20.514495
Pongo_pygmaeus-A939_Nonja	25.6488	18.264823
Pongo_pygmaeus-A942_Gusti	29.6212	12.912687
Pongo_pygmaeus-A943_Tilda	30.2099	12.133029
Pongo_pygmaeus-A944_Napoleon	29.1711	7.543612
Gorilla_gorilla_gorilla-KB3782_Vila	10.8599	7.763543
Gorilla_gorilla_gorilla-KB3784_Dolly	14.7481	10.054133
Gorilla_gorilla_gorilla-KB4986_Katie	12.6238	8.955044
Gorilla_gorilla_gorilla-KB5792_Carolyn	10.4592	7.50207
Gorilla_gorilla_gorilla-KB5852_Helen	13.1221	9.352174
Gorilla_gorilla_gorilla-KB6039_Oko	13.8467	9.738127
Gorilla_gorilla_gorilla-KB7973_Porta	10.4381	7.553997
Gorilla_gorilla_gorilla-Kwan	-	8.935158
Gorilla_beringei_graueri-Mukisi	-	2.313648
Gorilla_gorilla_gorilla-Kamilah	-	22.561756
Pongo_pygmaeus-KB5404_Billy	-	12.425859
Pongo_pygmaeus-KB4204_Dinah	-	2.84881
Pongo_abelii-KB4361_Dennis	-	4.627225
Pongo_abelii-KB4661_Dolly	-	2.420915
Pongo_pygmaeus-KB5405_Louis	-	3.508526
Pongo_pygmaeus-KB5406_Doris	-	4.041577
Pongo_pygmaeus-KB5543_Baldy	-	3.784019
Pongo_abelii-KB5883_Likoe	-	2.955035
Pongo_abelii-KB9258_Bubbles	-	5.161534
Pongo_abelii-SB550_Sibu	-	3.676026

Section 2: Nonhuman primate reference genome comparisons

Mapping to nonhuman primate reference assemblies: In order to assess human-specific deletions and the copy number of genomic sequence not represented in the human reference, we identified nonhuman primate-specific sequence. The gorGor3(McLean et al. 2011; Chimpanzee Sequencing and Analysis Consortium 2005; Scally et al. 2012), panTro3(Varki et al. 2008; Locke et al. 2011; Chimpanzee Sequencing and Analysis Consortium 2005) and ponAbe2(Elsea and Girirajan 2008; Kent 2002; Locke et al. 2011) references were each mapped in 1 kbp chunks with BLAT(Varki et al. 2008; Altschul et al. 1990; Olson 1999; Kent 2002) against the human reference genome (default parameters and *–minIdentity=90*) to identify regions ≥1 kbp that could not be aligned to the human reference (**Table 2.1**). Sex chromosomes were excluded and only regions with ≥500 bp of unmasked sequence were considered in the analysis.

Table 2.1: The total sum of unmasked base pairs considered from each nonhuman primate assembly that did not place in the human reference genome.

Reference	Species	Unmasked base pairs	Homologous control sequence base pairs included in mapping
panTro3	Chimpanzee	6015364	2848731
gorGor3	Gorilla	6124188	2827964
ponAbe2	Orangutan	21110529	2797810

All genomes were mapped to this nonhuman primate-specific sequence in addition to four orthologous collections of ~2.85 Mbp of control diploid sequence present in all primate references in order to calibrate copy number counts. Copy number calling was performed on nonhuman primate sequence not present in the human reference, as explained above, adjusted to use the appropriate set of control sequence.

In total, 30,033 loci containing ≥500 bp of unmasked sequence were identified in the nonhuman primate genomes assayed; however, this set is redundant as sequence not present in the human reference genome may be present in multiple nonhuman primate reference genomes. We, thus, searched all sequences against each other using Mega BLAST(Cooper et al. 2011; Sudmant et al. 2010; Emerson et al. 2008; Altschul et al. 1990) (Mega BLAST version 2.2.11, parameters *-D 2 -p 95 -F m -U T -s 220 -R T*) and clustered together all sequences that could align over 20% of their length with ≥97% identity. A total of 22,114 clusters were then identified and from each cluster the sequence containing the maximum number of unmasked base pairs was selected for consideration. Finally, masked sequence in each of the regions considered was padded (extended out on both sides) by 36 bp to eliminate mapping edge effects(Witkin 1984; Sudmant et al. 2010) as was done in the full genome mappings described above. After repeat masking, 9855 regions with ≥500 bp remained. These loci, encompassing 25,374,943 total base pairs, were analyzed for their duplicated and deleted content (**Figure 2.1, see Section 6**).

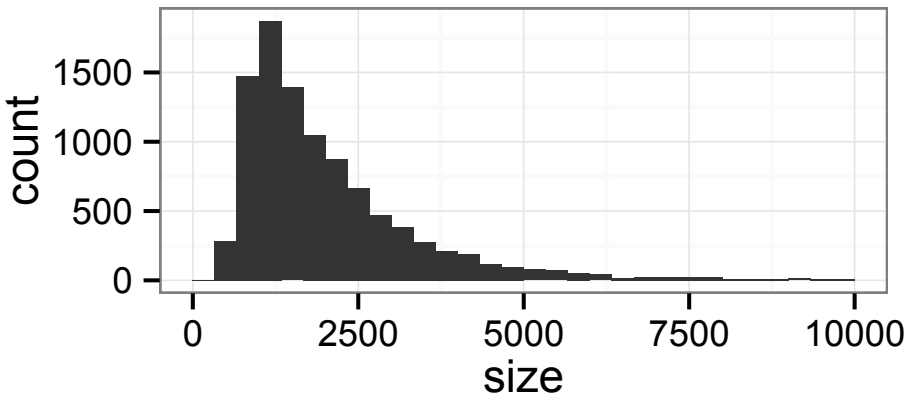


Figure 2.1: Size distribution of 9855 nonredundant loci not present in the human reference genome analyzed.

Section 3: Copy number variant calling methods

Segmentation of duplication and deletion loci: To detect duplications and deletions within genomes, we developed a new algorithm based on the concept of scale-space filtering(Ventura et al. 2011; Witkin 1984). Genome copy numbers were predicted in tiled contiguous windows at a resolution of 500 bp of unmasked sequence in each genome analyzed. We strived to segment these windowed estimates into diploid, duplicated, or deleted regions encompassing contiguous sets of windows. Additionally, we aimed to segment complex duplication architectures to inform the underlying SD structure (**Figure 3.1**).

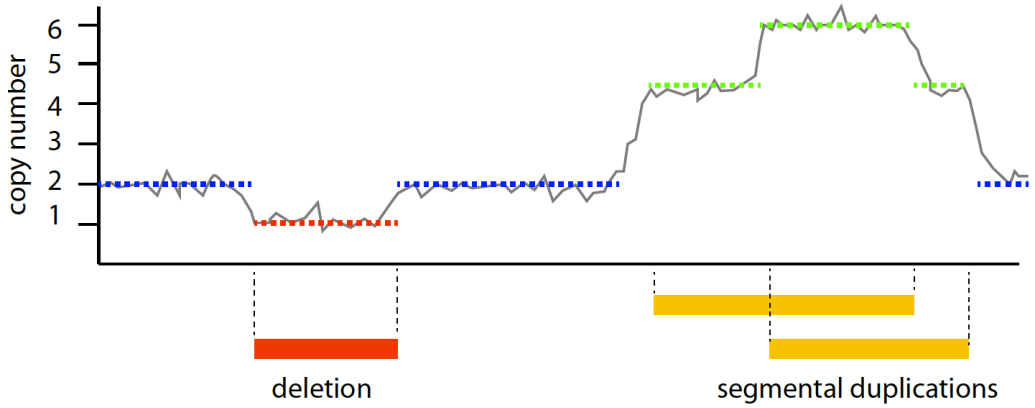


Figure 3.1: A schematic of the desired segmentation output. The gray line represents the predicted copy number of contiguous tiled windows. An ideal segmentation should distinguish deletion (red), diploid (blue), and duplication (green) segments. Also, complex duplication segments should be sub-segmented into constitutive duplications of differing copy number.

To construct this segmentation we first applied scale-space filtering over the windowed copy number estimations. Scale-space filtering is a multi-resolution technique that transforms a waveform signal *f(x)* into a set of waveforms *f(x,σ),* parameterized by the '*scale*' variable σ. The transformation is achieved by applying a Gaussian convolution over the waveform for multiple values of σ, the standard deviation of the Gaussian. The zero crossings of the second derivative of the transformed waveform, corresponding to peaks of the first derivative of the convolution, are the loci of inflection of the original curve, smoothed over a particular scale. Zero contours of the second derivative within scale-space are assumed to correspond to the same underlying event and events are assumed to originate from the *x* location where $\sigma \rightarrow 0$. Thus, for each scale-space σ_i, we traverse the scale-space zero contours from σ_i to σ_{min} and construct a set of contour intercepts corresponding to inflections on our original signal. Finally, we select a scale-space cutoff, below which all contour intercepts are discarded, and use the remaining intercepts to construct an initial set of segments. Finally, we hierarchically cluster these segments assigning each segment a value

equivalent to the median of the copy number of the windows it encompasses and greedily merging the pair of adjacent segments that shows the least difference in estimated copy number. This merging procedure is repeated until the minimum difference between adjacent segments reaches a designated threshold, in our case 0.5, corresponding to a copy number difference of 0.5 between adjacent segments.

Identifying deletions - methods and sensitivity: We constructed a set of deletions for each species by running our segmentation algorithm on each of the genomes analyzed individually, then merging the total set of events identified in all individuals into a species deletion set. In this way, the population of individuals is used to discover the total possible set of deletion events and these events can then be genotyped in all individuals. To test the accuracy of this method, we then compared the set of deletions called in gorillas to those recently identified in the Western lowland gorilla Kwan(Hormozdiari et al. 2009; Ventura et al. 2011) using an orthologous paired-end-read approach (Variation Hunter(Hormozdiari et al. 2009)) and validated by array comparative genomic hybridization (arrayCGH) **(Figure 3.2)**. Using segmentations based on 500 bp windows in the single individual Kwan, we were able to rediscover 97.6% (81/83) of deletions encompassing 3 kbp or more of unmasked sequence that were previously identified by paired-end analysis in this same individual and 89.6% of deletions encompassing 1 kbp or more of unmasked sequence (457/510). As smaller deletions were more challenging to discover, we leveraged the fact that 32 gorilla genomes were sequenced and compared the total set of gorilla deletions discovered by our algorithm to those identified by paired-end sequencing in Kwan. 95.3% (486/510) and 98.8% (159/161) of events encompassing greater than 1 kbp or greater than 2 kbp unmasked base pairs, respectively, were rediscovered in the total gorilla deletion set, demonstrating the power of leveraging a population of sequenced individuals.

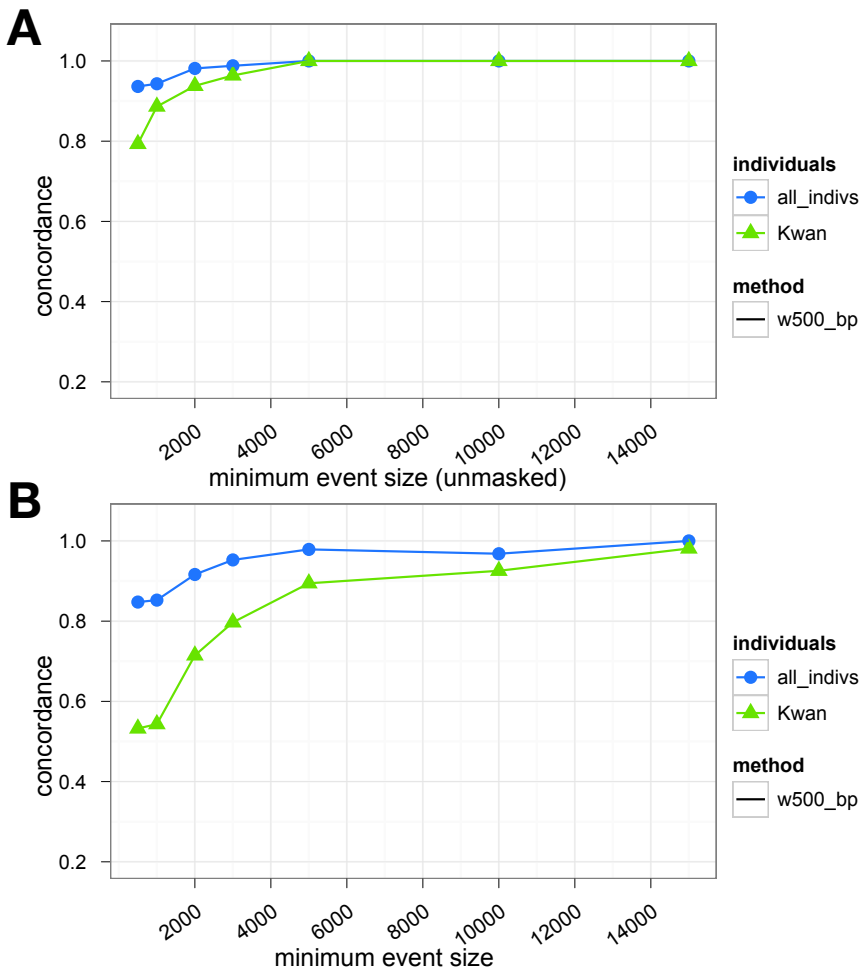


Figure 3.2: The concordance rate of deletion calls detected by the read-depth-based segmentation algorithm compared to those made by paired-end structural variation detection using Variation Hunter(Bailey et al. 2002; Hormozdiari et al. 2009)**. Concordance rates are shown for different size thresholds using the number of unmasked base pairs (A) underlying an event (i.e., the number of base pairs used in the segmentation) and the total event size (B). Concordance rates are plotted for 500 bp segmentations and for events discovered in just Kwan and events discovered across the set of all gorillas (n=32).**

Identifying duplication segmentations: For each species we constructed a consensus set of duplication segments by merging the individual segmentations of all individuals in a species with the purpose of leveraging information across multiple individuals to inform the true SD boundaries. Segmentations across multiple individuals are largely similar. Additionally, as demonstrated from the detection of deletion loci across multiple individuals above, duplicated loci discovered in one individual can then be genotyped across the population of individuals. Merged duplication segmentations were constructed by first considering all duplicated regions (duplicated regions may or may not encompass multiple segments; **Figure 3.1**) detected across any individual. As demonstrated for the deletions above, we have significant power to detect events at a threshold of 3 kbp with 96.4% of deletions identified from discovery

in a single individual. For each duplicated region, we assessed all individuals where the duplication was present and determined the median number of segments (*k)* identified in individuals in this duplicated region. Each edge between adjacent segments was also assigned a count equal to the total number of individuals in which this edge was identified. A consensus segmentation *k* segments long was then constructed by greedily selecting the *k+1* edges with the highest counts **(Figure 3.3)**.

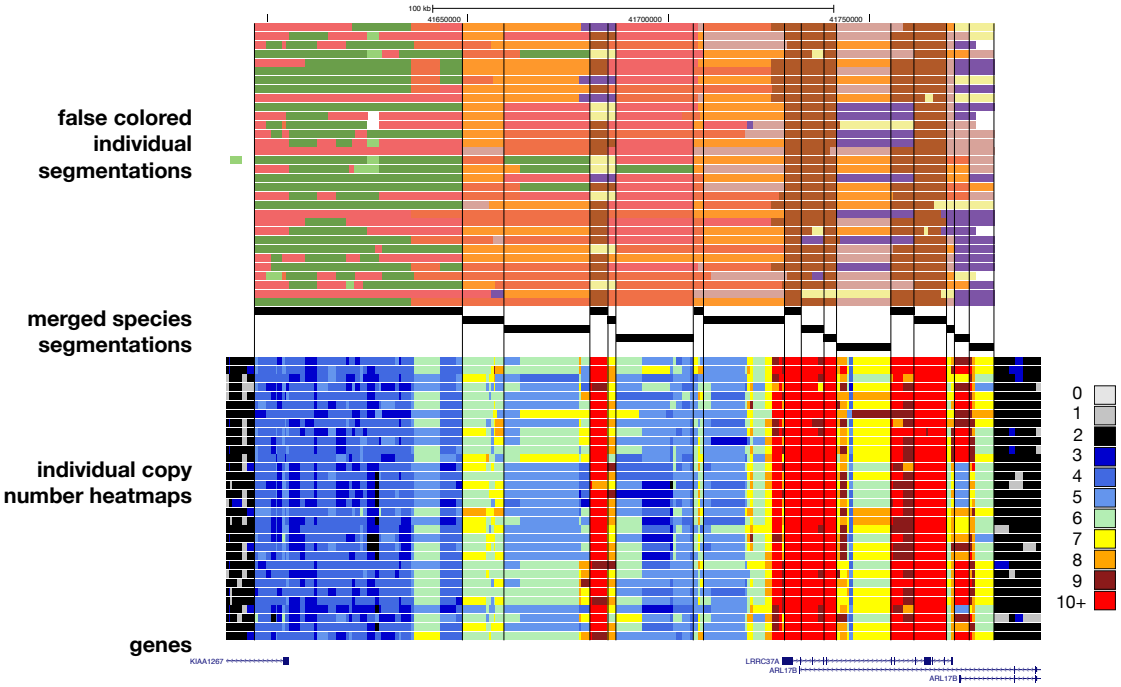


Figure 3.3: An example of a merged species segmentation overlaid over copy number heatmaps for 32 Western gorillas for a duplicated region on 17q21. The individual duplication segmentations (top, false colored) are merged (middle bars and black lines) to construct a more robust species segmentation from which copy numbers can be assessed and compared between individuals. Each window represents 500 bp of unmasked sequence scaled to genomic coordinates.

To test the effectiveness of our method, we compared the consensus duplication segmentation made across the 10 humans we analyzed to annotated SDs in the reference genome(Bailey et al. 2002) called from genome self-alignments (whole-genome alignment comparison, WGAC) of a single individual. We identified 172,817,129 (**non-copy number corrected**) duplicated bp among the 10 humans analyzed (excluding sex chromosomes), 139,560,079 bp of which were present in at least 5/10 individuals, compared to 130,447,077 duplicated bp identified in the human reference genome by WGAC. 95.4% of all SDs containing >3 kbp of unmasked sequence in the human reference genome were identified in our analysis of these 10 human individuals and 92.7% of all annotated duplications were identified in at least 5/10 of these individuals (**Figure 3.4**).

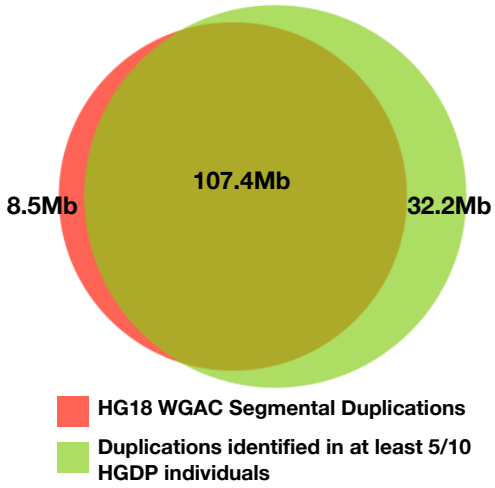


Figure 3.4: The overlap between duplicated base pairs identified from WGAC and identified by read-depth followed by segmentation in 10 diverse human individuals and present in at least 5/10 individuals containing at least 3 kbp of unmasked sequence.

We next assessed our segmentation by comparing the boundaries of known SDs, identified from self-alignments of the human genome(Bailey et al. 2002), to those identified by our segmentation algorithm. We, thus, took all segments overlapping known SDs and compared the edges of each segmentation to the set of known human SD edge boundaries. Our duplication segmentation edges, which were based on copy number estimates across 500 unmasked bp tiled windows, had a median distance of 367 unmasked bp to the nearest known SD edge **(Figures 3.5 and 3.6)** and a median total distance of 1373 total base pairs to the nearest known SD boundary. Since the distance between known duplication edges and our segmentation edges is at the resolution of our windowed copy number estimates, we conclude that our segmentation is accurately reflecting the underlying duplication structure of the genomes analyzed.

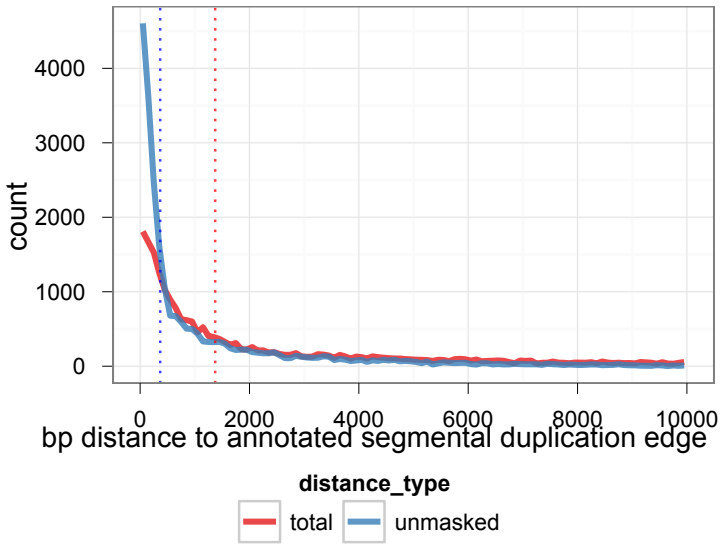


Figure 3.5: Histograms of the distance in total and unmasked base pairs from predicted duplication segmentation edges to known SD edges in the human genome. The median distance from a predicted duplication boundary to a known duplication boundary (dotted lines) is 367 bp of unmasked sequence or 1373 total bp. Human duplication segmentations were identified from duplications identified in 10 human genomes.

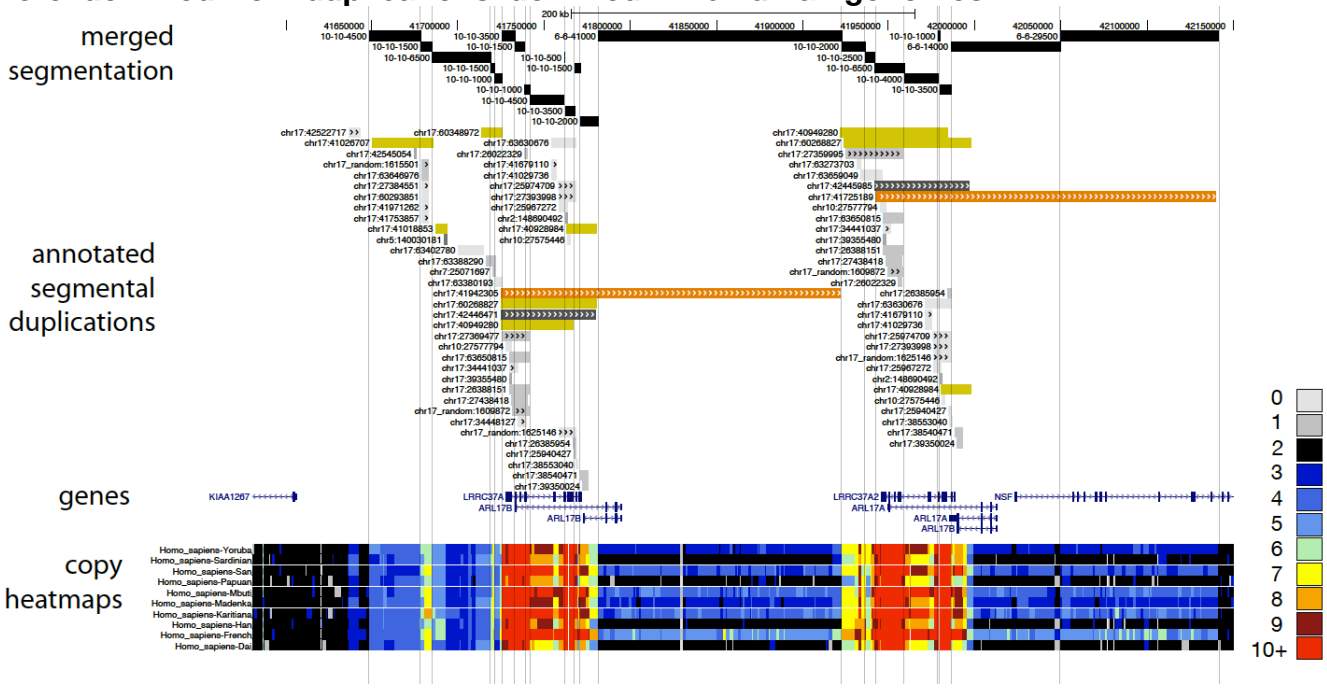


Figure 3.6: An example of a complex duplicated locus on 17q21 in human. Human-annotated SDs are shown overlaid with predicted merged human species segmentations and copy number heatmaps of the 10 individuals analyzed. Boundaries of predicted human SD edges correspond well with annotated SD boundaries.

Section 4: Validations

ArrayCGH validations: We designed three custom arrays to validate our call-sets: a duplication array, a deletion array, and an array targeted to sequence present in nonhuman primate reference genomes but not in the human reference. A total of 58 total hybridization experiments (29 test/reference, reference/test dye-swap pairs) were performed on these platforms using the human HapMap individual NA12878 as a reference against 18 different nonhuman primate individuals (4 gorillas, 2 bonobos, 4 orangutans, and 8 chimpanzees). Primate and human DNA were labeled with Cy3 or Cy5 dye (NimbleGen labeling kit). Labeled DNA were mixed in equal amounts and hybridized to either a NimbleGen 3 X 720K or Agilent 2 X 400K array for 40 hours at 42 degrees. Slides were then washed, scanned, and analyzed. Individual noisy and poorly performing probes were identified as those giving opposite signals in dye swap experiments and were discarded. Additionally, only loci with three or more probes were considered. Lineage-specific deletions were considered confirmed if a mean log2ratio of ≤-0.5 across a locus was seen

in all individuals in that lineage. Duplications were considered identically using a log2ratio cutoff of 0.25. Lineage-specific duplications along the human lineage were additionally considered confirmed if they intersected with known SDs identified by WGAC(Lichter et al. 1990; Bailey et al. 2002). The validation rate among all experiments ranged from 96.9% to 98.1% for fixed sites and was 85% for copy number polymorphisms (CNPs) demonstrating our call-set to be highly robust (**Table 4.1**).

Table 4.1: Validation results summary. In total 9034 loci were experimentally tested, of which 8641 were validated, resulting in an overall validation rate of 95.65%.

Method	Targeted variation	Experiments performed	Loci experimentally tested	Loci confirmed	Validation rate
arrayCGH	**Duplications**	14	3776	3660	96.93%
FISH	**Duplications**	104	104	102	98.08%
arrayCGH	**Deletions**	14	2503	2476	98.92%
arrayCGH	**Nonhuman reference sequence deletions (absent from human reference)**	30	1518	1490	98.16%
arrayCGH	**CNPs**	15	1520	1294	85.13%
	Total	**177**	**9421**	**9022**	95.76%

Fluorescent *in situ* hybridization (FISH) validations: To further assess the structure of duplications and deletions identified, we performed 109 FISH experiments on human and nonhuman great ape cell lines across 53 different loci. Metaphase and interphase nuclei from nonhuman primates (common chimpanzee, gorilla, bonobo, and orangutan) were obtained from lymphoblastoid or fibroblast cell lines; human metaphase spreads were derived from PHA-stimulated peripheral lymphocytes of normal donors by standard procedures. Briefly, cells were treated with colcemid (D1925; Sigma) and collected by centrifugation, incubated in a hypotonic solution (KCl 0.56%), and then washed three times and stored in a fixative solution (one part acetic acid, three parts methanol). DNA from human fosmid clones was extracted using a Plasmid Miniprep kit (Bio-Rad, Cat# 732-6100). FISH assays were performed as previously described by Lichter et al(Cheng et al. 2005; Lichter et al. 1990). Fosmids were directly labeled either with Cy3-dUTP or fluorescein-dCTP by nick-translation reactions. Three hundred nanograms of labeled probe were used for the FISH experiments. Hybridization was performed at 37°C in 2XSSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 μg of sonicated salmon sperm DNA in a volume of 10 μL. High stringency post-hybridization washings (three) were at 60°C in 0.1X SSC. Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a CCD camera. Fluorescence signals detected with Cy3 and fluorescein filters and chromosomes and nuclei images detected with DAPI filter were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

Of the 109 *in situ* experiments performed, we confirmed in 102 cases the predicted duplication and were able to resolve the duplication locus (**Table 4.2**). Five experiments were inconclusive showing only background signal resulting in an overall validation rate of 98% (102/104). Each of the two remaining duplications not resolved by FISH showed strong signals by arrayCGH, suggesting a FISH false-negative (**Table 4.2**). Overall, the FISH and arrayCGH validation rates results are highly concordant, resulting in an overall validation rate of ~97% (**Table 4.1**).

Table 4.2: Fluorescent *in situ* hybridization experiments performed on 44 loci. For duplicated loci the location of the duplication resolved by FISH is described

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	GGO mapping
1	chr1 144405804 144477396	gorilla	WIBR2-3559A01_G248P87580A1	chr1 144410826 144451556	chr1q	chr1 cen tandem duplicated
4	chr1 5004069 5226396	gorilla	WIBR2-1327G02_G248P83830D1	chr1 5121247 5159131	chr1p	chr1, chr1II
5	chr10 80749080 80787310	gorilla	WIBR2-3713M05_G248P801584G3	chr10 80748377 80787033	chrXq	chrXp, chrXq
15	chr16 27122308 27140590	gorilla	WIBR2-1200H11_G248P84685D6	chr16 27110478 27152058	chrXVIp	interchromosomal duplication
17	chr16 11678625 11783121	gorilla	WIBR2-2615L04_G248P82867F2	chr16 11705251 11742593	chrXVIp	chrXVI intrachromosomal duplication
20	chr16 16012223 16050785	gorillas	WIBR2-0620N08_G248P81143G4	chr16 16012224 16050785	chrXVIp	chrXVI intrachromosomal duplication
31	chr2 114161861 114212636	gorilla	WIBR2-3279N11_G248P89293G6	chr2 114166386 114208379	chr1lq	telomeric signals on most of chromosomes
34	chr4 1775908 1827411	gorilla	WIBR2-0947D18_G248P82039B9	chr4 1781550 1825464	chr1Vp	telomeric signals on 10 chromosomes
36	chr5 34020735 34084124	gorilla	WIBR2-2368K23_G248P83676F12	chr5 34031475 34075334	chrVp	chrV, chrX
37	chr5 77402005 77516591	gorilla	WIBR2-2627N05_G248P82853G3	chr5 77435394 77473881	chrVq	interchromosomal duplication
39	chr7 156435178 156458637	gorilla	WIBR2-0884C07_G248P8628B4	chr7 156425158 156469861	chrVIIq	chr1 telomeric signal, chrVII telomeric signal
44	chr9 123814932 123988455	gorilla	WIBR2-1375F20_G248P83787C10	chr9 123853669 123893315	chrIXq	telomeric signals on most of chromosomes
48	chr9 89429256 89626125	gorilla	WIBR2-0837A23_G248P82188A12	chr9 89491230 89526537	chrIXq	chr IX slightly tandem duplicated

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PPY mapping
3	chr1 15656749 15686696	orangutan	WIBR2-0436H17_G248P81629D9	chr1 15650840 15687775	chr1p	background
6	chr11 20117749 21118417	orangutan	WIBR2-2734D12_G248P8315B6	chr11 20372162 20411250	chrXIp	chr XI cen tandem duplicated
8	chr13 23270273 23349987	orangutan	WIBR2-0496F08_G248P801987C4	chr13 23309863 23347197	chrXIIq	chr XIII slightly tandem duplicated
25	chr17 55475334 55511074	orangutan	WIBR2-1710A05_G248P86704A3	chr17 55479535 55518239	chrXVIIq	chr XVII pcen, qcen
27	chr2 101380255 101457597	orangutan	WIBR2-3128H14_G248P89387D7	chr2 101397563 101439695	chr1lq	chr1p tandem duplicated
43	chr8 86282620 86524453	orangutan	WIBR2-0608A19_G248P80208A10	chr8 86341881 86385282	chrVIIIq	chrVIII pcen, VIIIq

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PPA mapping
7	chr11 70513191 70625827	bonobo	WIBR2-1799M24_G248P87294G12	chr11 70546621 70584617	chrXIq	chrXIq
24	chr16 66600825 66618362	bonobo	WIBR2-3683L04_G248P801087F2	chr16 66589367 66627782	chrXVIq	chrXVIp, slightly tandem duplicated

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping
11	chr14 91860435 92021289	chimpanzee	WIBR2-0903K12_G248P8634F6	chr14 91896413 91935318	chrIVq	chrXIVq, chrXIVp heterozygous, chr1lp parm
32	chr2 203030616 203226332	chimpanzee	WIBR2-1459K22_G248P84246F11	chr2 203101952 203142773	chr1lq	chr1lq, slightly tandem duplicated

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping	PPA mapping	GGO mapping
14	chr15 39927488 39955630	chimpanzee,bonobo,gorilla	WIBR2-2454C17_G248P83476B9	chr15 39911201 39954965	chrXVq	telomeric signals on most of	telomeric signals on 10 chromosomes	chrXVq, chr1p highly tandem duplicated
19	chr16 14871940 14897342	chimpanzee,bonobo,gorilla	WIBR2-3745N23_G248P802117G12	chr16 14870206 14907880	chrXVIp tandem	chrXVIp tandem duplicated	chrXVIp tandem duplicated	chrXVIp, deeply dupl intrachromally
23	chr16 65548821 65602343	chimpanzee,bonobo,gorilla	WIBR2-3791N09_G248P802013G5	chr16 65562046 65603553	chrXVIq	chrXVIq, telomeric signals on 4	chrXVI, chrXVII	chrXVI, chr1lq
38	chr6 65310956 65499167	chimpanzee,bonobo,gorilla	WIBR2-1439F04_G248P84319C2	chr6 65323861 65362941	chrVIq	chrVIcen tandem duplicated	chrVIpcen qcen tandem duplicated	chrVIcen tandem duplicated
38	"	chimpanzee,bonobo,gorilla	WIBR2-2633O09_G248P83412H5	chr6 65364960 65402557	chrVIq	chrVIq slightly tandem duplicated	chrVIq tandem duplicated	chrVIcenq, slightly tandem duplicated

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping	PPA mapping	
18	chr16 11874623 11900245	chimpanzee,bonobo	WIBR2-3233C17_G248P88748B9	chr16 11860389 11900704	chrXVIp	chrXVIp, duplicated	chrXVIp, duplicated	
22	chr16 28630283 28654551	chimpanzee,bonobo	WIBR2-0862B15_G248P8725A8 *see next region (RP11-1084a18 in Marques et al)	chr16 28611489 28653888	chrXVIp	dupl intrachrom	intrachromosomal duplication	
28	chr2 102213094 102395839	chimpanzee,bonobo			chr1lq	background	intrachromosomal duplication	
29	chr2 102294394 102334929	chimpanzee,bonobo	ABC8_40986800_H8	chr2 102320495 102354897	chrII, chrXIV	chrIIp, chr1lq	chrIIp, chr1lq	close to pericentromere
29	"	chimpanzee,bonobo	ABC8_43206800_N22	chr2 102265757 102307366	chr1lq	chr1lq tandem duplicated	intrachromosomal duplication	
33	chr3 34948 74373	chimpanzee,bonobo	WIBR2-0615D22_G248P80579B11	chr3 36462 74164	chrIIIp	terminal signals	intrachromosomal duplication	
35	chr5 23540210 23565056	chimpanzee,bonobo	WIBR2-0524E16_G248P8246C8	chr5 23534183 23569271	chrVp	background	telomeric signals	
42	chr8 16167128 16219170	chimpanzee,bonobo	WIBR2-0997H18_G248P84619D9	chr8 16178748 16215490	chrVIIIp	interchromosomal duplication	interchromosomal duplication	
46	chr9 675136 771511	chimpanzee,bonobo	WIBR2-1748N11_G248P86645G6	chr9 704536 746797	chrIXp	telomeric signals	telomeric signals	
50	chr16:88627470-88628110	chimpanzee,bonobo	WIBR2-1907G21_G248P86940D11	chr16 88627472 88671488	chrXVIq	telomeric signals	telomeric signals on most of chromosomes	

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping	PPA mapping	PPY mapping	GGO mapping
2	chr1 146,613,026 146,700,321	tan,human	ABC12_7917949_E10	chr1 146563626 146604053	tandem duplicated,	chr1qcen tandem duplicated	chr1 slightly tandem duplicated	chr1	chr1cen tandem duplicated
2	"	tan,human	WIBR2-2675D14_G248P88911B7	chr1 146633091 146677498	chr1qcen tandem	chr1cen tandem duplicated, chr1q	chr1 tandem dupl, chrXcen	1cen, chrX cen	chr1cen tandem duplicated, chr1q
9	chr13 56600521 56648570	tan,human	ABC7_43067700_J1	chr13 56613302 56644175	tandem duplicated	chrXIIIq, chrVIIq	chrXIII slightly tandem dupl	chrXIII slightly tandem dupl	chrXIII slightly tandem dupl
16	chr15 80438418 80451290	(chimpanzee,bonobo,gorilla,orang	ABC8_43204800_I6	chr15 80427561 80464299	chrXV, tandem duplicat	chrXV, tandem duplicated	chrXV, tandem duplicated	chrXV, tandem duplicated	chrXV, tandem duplicated
21	chr16 18349961 18377021	tan,human	WIBR2-2468F23_G248P83589C12	chr16 18343163 18375601	duplicated intrachrom	(both tandem and interspersed) and	(both tandem and interspersed) and	duplicated	chrXVI highly dupl intrachrom (both tandem and interspersed) and interchrom, telomeric signals on chrXVII
26	chr19 60959888 60979394	tan,human chimpanzee,bonobo,gorilla,orangu	WIBR2-1349F16_G248P83731C8	chr19 60950492 60988308	tandem duplicated chrIXpcen, chrIXqcen tandem duplcated,	background	chrXIX slightly tandem dupl	chrXIX, slightly tandem duplicated	chrXIX, slightly tandem duplicated
47	chr9 67516581 67560113	tan,human chimpanzee,bonobo,gorilla,orangu	ABC8_2620740_C8	chr9 67524438 67559082	chrIIcen	chrIXcen, chrIIIcen, chrIp	chrIXcen, chrIIIcen, chrIp	chrIXcen, chrXVcen	chrIXcen, chrXVcen
49	primate chr4 8927830 9032272	tan,human	WIBR2-3711P11_G248P802513H6	chr4 8934169 8977279	chrIV slightly tandem duplicated	chrIV, chrVI	chrIV, chrVI, both highly tandem dplicated	chrIV tandem duplicated, heterozygously	chrIV, chrVIII

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping	PPA mapping	PPY mapping	GGO mapping
10	chr14 44862782 45355696	Gorilla deletion	WIBR2-1696I03_G248P87976E2	chr14 44774835 44820642	chrXIVq, IIIp	chrXIVq, IIIp	chrXIVq, IIIp	chrXIVq, IIIp	chrXIVq, IIIp
10	"	Gorilla deletion	WIBR2-3167E03_G248P8048C2 *see next region (RP11-235j3 in Marques et al)	chr14 44869895 44915861	chrXIVq	chrXIVq telomeric signals on most of	chrXIVq telomeric signals on most of	chrXIVq	deletion, no signal
13	chr15 39835104 39998952	chimp,bonobo			chrXV	chromosomes	chromosomes	chrXV, slightly tandem duplicated	chrXV, chrIp telomeric signal tandem duplicated
40	chr7 74997169 75097487		WIBR2-1048D01_G248P82429B1	chr7 75025034 75066381	chrVII	chr VII, tandem duplication	chr VII, tandem duplication	chrVII	chrVII
41	chr7 78814148 78979756	chimp, bonobo	WIBR2-2361K03_G248P83384F2	chr7 78923255 78960262	chrVII	chr VII, tandem duplication	chr VII, tandem duplication	background	chrVII

Slide nr	location (hg18)	Species and notes on FISH type	probe(s)	Probe coordinates	HSA mapping	PTR mapping	PPA mapping	GGO mapping	PPY mapping
12	chr15 20050550 20087802	gorilla,chimpanzee	WIBR2-2094M06_G248P87722G3	chr15 20049675 20089158	chrXV, small interchromosomal duplications	chrXV	background (faint, ambiguous signal)	chrXV cen tandem duplication, chrIIp cen, chrIIqcen	background (faint, ambiguous signal)
30	chr2 113911738 113970250	chimpanzee,bonobo,human	WIBR2-2982J16_G248P88319E8	chr2 113914653 113950832	HSA mapping chrIIp, chrIXpcenq	PTR mapping faint telomeric signals	PPA mapping telomeric signals		
45	chr9 38375755 38415229	gorilla, orangutan	WIBR2-0945A06_G248P89854A3	chr9 38379588 38417406	HSA mapping chrIX	PPY mapping chrIX, slightly tandem duplicated	GGO mapping chrIX cen, chrIIp cen, chrIIq cen		

Section 5: Lineage-specific great ape segmental duplications and deletions

Lineage-specific SDs: Lineage-specific duplication loci were assigned by assessing all duplicated segments and labeling them as lineage-specific SDs if >50% of individuals in a particular lineage were at least copy number 4 while <10% of individuals in all other lineages showed any signature of duplication (<10% ≥copy 3). All segments were assigned a duplication state. The same approach was taken to identify expansions, requiring all individuals in a particular lineage to have a higher copy number than all other individuals. The total number of duplicated base pairs in a specific lineage was thus calculated by summing the duplication content of all duplicated segments corrected for the copy number of SDs present in the human genome (this process has previously been termed *copy number correction*(Cheng et al. 2005; Marques-Bonet et al. 2009; Ventura et al. 2011)) and excluding the ancestral locus of the duplication (**Table 5.1**). We note that in previous analyses of lineage-specific SDs we have not excluded the ancestral locus from total counts(Meyer et al. 2012; Cheng et al. 2005; Marques-Bonet et al. 2009; Ventura et al. 2011). However, to directly compare different rates in the accumulation of genetic variation, excluding the ancestral locus is necessary. Duplicated base pair counts were also computed over the nonhuman primate reference assembly sequence that was not present in the human reference genome. The overall contribution of this sequence to the total number of duplicated base pairs was minimal however (~2%).

Table 5.1: Counts of the number of events and total base pairs specifically duplicated in different lineages. Base-pair counts are corrected for the SD content of the human reference genome and to exclude the ancestral duplication locus.

lineage	copy corrected duplicated base pairs	number of sites
Gbeg-Ggod-Ggog-Hde-Hsa-Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	184296495	7118
Gbeg-Ggod-Ggog	29413779	278
Pab-Ppy	27854532	885
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	24491334	1172
Ppa-Ptre-Ptrs-Ptrt-Ptrv	6050953	121
Hde-Hsa	5715010	282
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	4227312	269
Ppa	860900	67
Ptre-Ptrs-Ptrt-Ptrv	802219	30
Ppy	487342	48
Gbeg	446285	89
Pab	361732	38
Ggod-Ggog	339371	35
Hsa	280445	19
Hde	142093	6
Ggod	111722	31
Ptrv	31250	15

Ptrt	20823	10
Ptrs	13994	2
Ptre-Ptrv	13830	2
Ptrs-Ptrt	1790	1
Ptre	1302	1
Ggog	5600	2
BONOBO-CHIMP-EASTERNGORILLA-WESTERNGORILLA	25555717	103
EASTERNGORILLA-HUMAN-WESTERNGORILLA	2646209	204
EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	2002046	78
EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	1776400	211
BONOBO-CHIMP-HUMAN-ORANGUTAN	1492574	146
HUMAN-ORANGUTAN	1015592	86
CHIMP-EASTERNGORILLA-HUMAN-WESTERNGORILLA	991242	64
CHIMP-EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	705223	55
BONOBO-CHIMP-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	642072	27
BONOBO-CHIMP-ORANGUTAN	435420	17
Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	426930	6
Hde-Hsa-Pab-Ppy	420272	9
HUMAN-WESTERNGORILLA	243475	65
Gbeg-Ggod-Ggog-Pab-Ppy	172364	17
BONOBO-EASTERNGORILLA-HUMAN-WESTERNGORILLA	148340	16
CHIMP-HUMAN	140254	30
BONOBO-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	122483	8
HUMAN-ORANGUTAN-WESTERNGORILLA	118243	53
ORANGUTAN-WESTERNGORILLA	111772	19
CHIMP-EASTERNGORILLA-WESTERNGORILLA	97792	3
BONOBO-CHIMP-WESTERNGORILLA	93547	23
BONOBO-HUMAN	46010	12
CHIMP-EASTERNGORILLA	34776	1
CHIMP-HUMAN-ORANGUTAN	27975	10
BONOBO-ORANGUTAN	23010	5
BONOBO-EASTERNGORILLA-WESTERNGORILLA	18598	1
Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	13736	5
BONOBO-ORANGUTAN-WESTERNGORILLA	13600	2
BONOBO-EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	10933	7
CHIMP-HUMAN-WESTERNGORILLA	8363	4
CHIMP-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	7058	2
EASTERNGORILLA-ORANGUTAN	6920	2
BONOBO-CHIMP-ORANGUTAN-WESTERNGORILLA	5015	2
BONOBO-CHIMP-EASTERNGORILLA	4623	2
Gbeg-Ggod-Ggog-Hde-Hsa	4616	2
CHIMP-ORANGUTAN	2466	1
BONOBO-CHIMP-HUMAN-ORANGUTAN-WESTERNGORILLA	2054	3
BONOBO-CHIMP-EASTERNGORILLA-HUMAN-ORANGUTAN	1924	3
BONOBO-HUMAN-ORANGUTAN	1778	1
BONOBO-CHIMP-HUMAN-WESTERNGORILLA	1303	2
EASTERNGORILLA-HUMAN-ORANGUTAN	931	5
BONOBO-HUMAN-WESTERNGORILLA	519	1
BONOBO-HUMAN-ORANGUTAN-WESTERNGORILLA	507	2
EASTERNGORILLA-HUMAN	301	1
BONOBO-CHIMP-EASTERNGORILLA-HUMAN	145	1

Resolving duplications inconsistent with the phylogeny: Some of the duplication events identified were inconsistent with the species' phylogeny and, thus, could not simply be assigned to any branch. For example, ~25.6 Mbp of sequence were identified as duplicated in chimpanzees, bonobos, and gorillas, but not in the human lineage. The most parsimonious explanation for this result in many cases is that this duplicated sequence was present in the ancestor of African great apes and was deleted in the human lineage. This sequence was subsequently added to correct lineage using the most parsimonious explanation.

Duplication rates: The rate of duplication was directly compared to the rate of single nucleotide substitution from single nucleotide polymorphism (SNP) data calculated for 86 of the same individuals assessed for duplication content (see companion Prado/Sudmant 2012; **Table 5.2**). Briefly, all individuals were aligned to the HG18 reference genome using BWA with relaxed edit distance parameters to account for the increased divergence of these species. SNPs were called using the GATK pipeline on a population level and were filtered for depth-of-coverage, proximity to an indel, and allele balance. Genetic distances between all species were then calculated by computing the genetic distances between all individuals and taking the mean pairwise divergence between individuals of different populations. A tree was constructed from this distance matrix by neighbor joining. The Denisova individual was placed as a sister group to the human genome splitting at 12.35% the branch length of the human-chimpanzee common ancestor with a branch length 1.2% shorter than human as reported in Meyer *et al* 2012(Sun et al. 2012; Meyer et al. 2012; Langergraber et al. 2012). Divergence times were calculated using a mutation rate scaled to a human-chimpanzee divergence time of 6.5 million years ago, (8.79e-10 substitutions/year per base **Table 5.2**). These speciation times match well with the literature(Sun et al. 2012; Langergraber et al. 2012). The rate of duplication along each branch was calculated by dividing the per-base-pair duplication rate (total copy number corrected duplicated bp/2.867e9) by the per-base-pair substitution rate and additionally by the estimated divergence times in millions of years (**Table 5.3**).

As we previously reported(1000 Genomes Project Consortium et al. 2012; Hach et al. 2010; Marques-Bonet et al. 2009; Ventura et al. 2011), we find a burst of duplication activity in the ancestor of African great apes (**Figure 5.1**) corresponding to a duplication rate along the African great ape lineage approximately 2.6-fold the rate of substitution and ~7-fold higher than the rate along the branch leading to the human lineage. This increased rate of duplication continued along the gorilla lineage and the chimpanzee-human ancestral lineage.

Table 5.3: Lineage-specific per-base-pair duplication and deletion rates as function of the number of substitutions along a lineage and the time in millions of years (MY)

lineage	copy corrected duplicated basepairs	deleted basepairs	duplicated bp / substituted bp	deleted bp / substituted bp	duplicated bp / million years (Mbp)	deleted bp / million years (Mbp)	duplicated sites	deleted sites	duplicated sites/MY	deleted sites/MY	deleted bp >5kb	deleted sites >5kb	deleted sites >5kb / million years
Ppy	487342	411190	0.1	0.09	0.27	0.23	48	36	26.67	20	352421	19	10.56
Pab	361732	210551	0.08	0.05	0.2	0.12	38	10	21.11	5.56	198160	6	3.33
Pab-Ppy	27854532	11292486	0.76	0.31	1.92	0.78	885	1111	61.03	76.62	9469243	562	38.76
Hde	142093	931589	0.07	0.49	0.18	1.16	6	113	7.5	141.25	755770	58	72.5
Hsa	280445	52505	0.13	0.02	0.35	0.07	19	12	23.75	15	25144	2	2.5
Hde-Hsa	5715010	4303500	0.38	0.29	0.95	0.72	282	919	47	153.17	2193258	224	37.33
Ppa	860900	1644331	0.17	0.32	0.43	0.82	67	149	33.5	74.5	1456279	86	43
Ptrv	31250	283012	0.01	0.11	0.03	0.28	15	34	15	34	228196	17	17
Ptre	1302	35715	0	0.02	0	0.04	1	7	1.11	7.78	21178	2	2.22
Ptre-Ptrv	13830	21208	0.04	0.06	0.14	0.21	2	4	20	40	12213	1	10
Ptrs	13994	10633	0.01	0	0.01	0.01	2	3	2	3	0	0	0
Ptrt	20823	12655	0.01	0.01	0.02	0.01	10	5	11.11	5.56	0	0	0
Ptrs-Ptrt	1790	8730	0.01	0.05	0.02	0.09	1	1	10	10	8730	1	10
Ptre-Ptrs-Ptrt-Ptrv	802219	451846	0.35	0.19	0.89	0.5	30	64	33.33	71.11	332534	30	33.33
Ppa-Ptre-Ptrs-Ptrt-Ptrv	6050953	7943919	0.54	0.71	1.38	1.81	121	665	27.5	151.14	7173626	428	97.27
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	4227312	1232172	1.17	0.34	3.02	0.88	269	213	192.14	152.14	806578	66	47.14
Gbeg	446285	533317	0.15	0.18	0.37	0.44	89	72	74.17	60	422322	37	30.83
Ggod	111722	1194717	0.05	0.53	0.12	1.33	31	146	34.44	162.22	956540	74	82.22
Ggog	5600	0	0	0	0.01	0	2	0	2.22	0	0	0	0
Ggod-Ggog	339371	178376	0.7	0.37	1.7	0.89	35	9	175	45	171072	7	35
Gbeg-Ggod-Ggog	29413779	7701191	1.76	0.46	4.46	1.17	278	770	42.12	116.67	6665912	460	69.7
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	55560468	8729420	2.61	0.41	6.61	1.04	1744	1185	207.62	141.07	6281670	391	46.55
Gbeg-Ggod-Ggog-Hde-Hsa-Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	192822491	165201851	0.07	0.06	192.82	165.2	7863	316	7863	316	164992705	244	244

Table 5.2: Divergence estimates from Prado and Sudmant *et al* (under review). Divergence times in years are estimated using a mutation rate of $8.79X10^{-10}$ bp/year, calibrated to a 6.5 million human-chimpanzee divergence time.

lineage	branch distance to leaf	branch length	time to leaf (MYA)	branch length (million years)
Gbeg-Ggod-Ggog-Hde-Hsa-Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	0.014354	-	16.3	-
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.006928	0.007422	7.9	8.4
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.005713	0.00126	6.5	1.4
Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.001752	0.00391	2	4.4
Pab-Ppy	0.001585	0.012788	1.8	14.5
Gbeg-Ggod-Ggog	0.000997	0.00584	1.1	6.6
Ptre-Ptrs-Ptrt-Ptrv	0.000937	0.00081	1.1	0.9
Ptrs-Ptrt	0.00085	0.00006	1	0.1
Ggod-Ggog	0.00081	0.00017	0.9	0.2
Ptre-Ptrv	0.000835	0.00013	0.9	0.1
Hde-Hsa	0.000737	0.005233	0.8	6
Ppy	0	0.00162	0	1.8
Pab	0	0.00155	0	1.8
Gbeg	0	0.00103	0	1.2
Hde	0	0.000666	0	0.8
Hsa	0	0.000737	0	0.8
Ppa	0	0.00177	0	2
Ggod	0	0.00079	0	0.9
Ggog	0	0.00083	0	0.9
Ptrv	0	0.00086	0	1
Ptre	0	0.00081	0	0.9
Ptrs	0	0.00087	0	1
Ptrt	0	0.00083	0	0.9

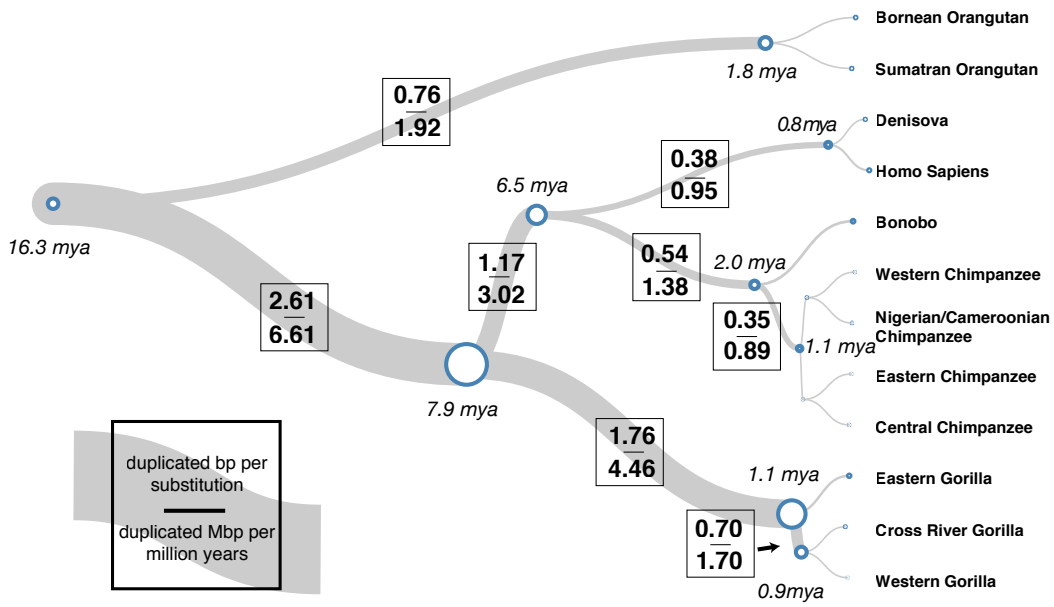


Figure 5.1: A species/subspecies tree with branch lengths proportional to the number of substitutions and branch widths scaled to the base-pair duplication rate per-nucleotide substitution and labeled for nonterminal branches. A burst of duplications in the ancestor of African great apes is apparent with duplication rates along the African great ape lineage of ~7-fold the rate along the human lineage per million years.

Lineage-specific deletions: Lineage-specific deletions were assigned as above for duplications, including sequence not present in the human reference genome (see **Section 2**). We required at least 80% of individuals to have a complete deletion signature in a particular lineage and all other lineages to show evidence of at least a single copy of the segment in 100% of all individuals assayed to ensure we were capturing fixed lineage-specific deletion events (**Figure 5.2**, **Table 5.4**). In contrast to the rate of duplication, we find the rate of deletion is largely clocklike with the major exception of the chimpanzee ancestral branch (**Table 5.3, Figure 5.3**), on which we observe a ~2-fold acceleration in the rate of chimpanzee deletions. This excess in deletion events correlates with a predicted severe reduction in effective population size, possibly the result of an extreme bottleneck in the chimpanzee-bonobo ancestor (**Figure 5.4,** Prado/Sudmant under review).

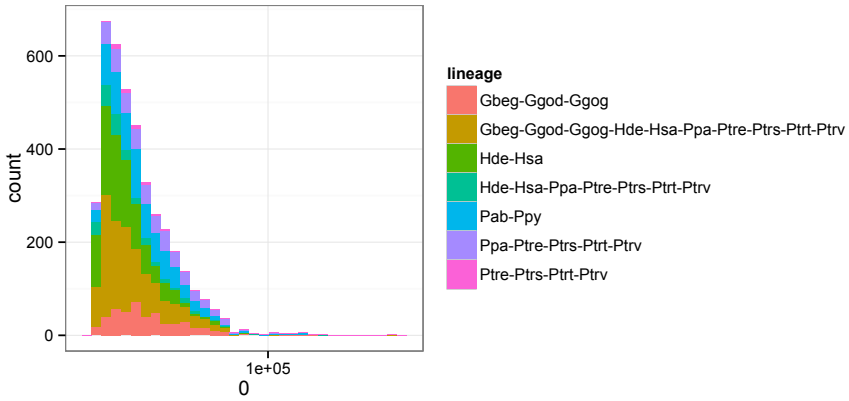


Figure 5.2: Size distribution (in log space) of lineage-specific duplications and deletions identified in great apes.

Table 5.4: Counts of all lineage-specific deletions identified in great apes and the number of sites identified.

species	total_bp	total_events	total_bp_gt_5kb	total_events_gt_5kb
Ptrs	7887	2	0	0
Ptrt	10655	4	0	0
Ptrv	276145	32	228196	17
Ggod-Ggog	174258	8	171072	7
Ppy	411190	36	352421	19
Ppa-Ptre-Ptrs-Ptrt-Ptrv	6495864	575	5727585	339
Ptre	35715	7	21178	2
Hde-Hsa	32339	2	32339	2
Hde	931589	113	755770	58
Ppa-Ptre-Ptrs-Ptrt-Ptrv-Pab-Ppy	373174	13	354786	8
Gbeg-Ggod-Ggog-Pab-Ppy	356638	31	330272	23
Ptrs-Ptrt	8730	1	8730	1
Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	1072867	76	1008570	57
Ppa	1644331	149	1456279	86
Ptre-Ptrv	21208	4	12213	1
Pab-Ppy	10562674	1067	8739431	518
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv-Pab-Ppy	165022599	264	164936373	237
Gbeg-Ggod-Ggog	6271686	663	5236407	353
Gbeg	527951	70	422322	37
Ptre-Ptrs-Ptrt-Ptrv	441826	62	326028	29
Ggod	1194717	146	956540	74
Hsa	20353	2	17909	1
Pab	210551	10	198160	6

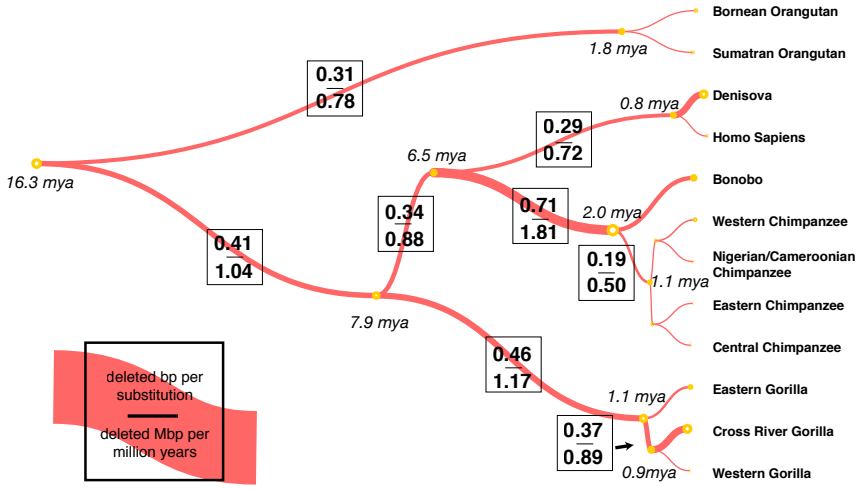


Figure 5.3: A species/subspecies tree with branch lengths proportional to the number of substitutions and branch widths scaled to the base-pair deletion rate per nucleotide substitution and labeled for nonterminal branches. The rate of deletion is largely clocklike with respect to the substitution rate; however, we find a ~2-fold increase in the rate of deletion in the chimpanzee ancestor.

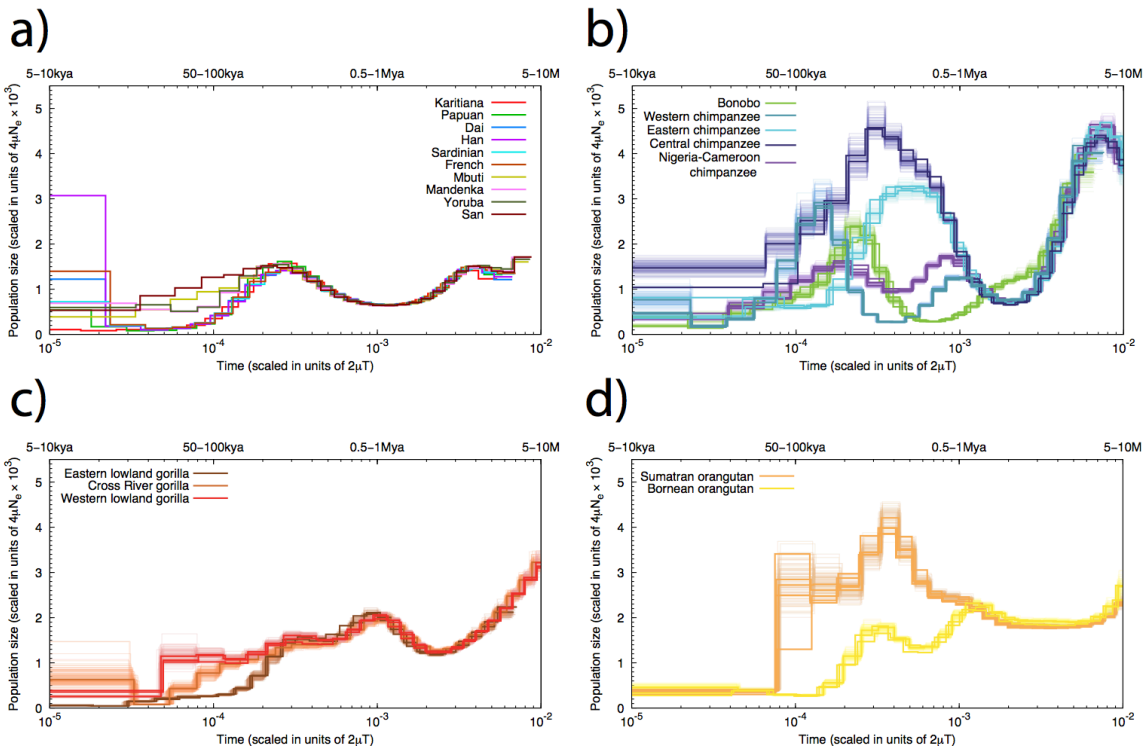


Figure 5.4: Pairwise Sequentially Markovian Coalescent (PSMC) profiles of effective population size (by scaled mutation rate) as a function of time (measured by pairwise sequence divergence, lower x-axis, and in years assuming a mutation rate ranging 10^{-9} – $5x10^{-10}$, upper x-axis) from Prado and Sudmant (under review). In panels b,c,d thin lines indicate 100 bootstrapped replicates to indicate confidence. A severe crash in the effective population size of the chimpanzee-bonobo ancestor is observed

just prior to the separation of the chimpanzee and bonobo species. Steep declines are also observed in Sumatran orangutans, bonobos, and Western chimpanzees.

In concert, these analyses demonstrate how two diametric forms of structural variation have differently shaped the genomes of great apes (**Figure 5.5**). While a burst of duplications characterized the African great ape ancestral genome, this rate subsequently rapidly decreased in the chimpanzee-human ancestral lineage and decreased, at a slower rate, in the gorilla lineage. The rate of deletion, in contrast, has remained largely consistent with the exception of an increase in the chimpanzee ancestor. This striking excess of deletions coincides with a drop in the predicted effective population size, suggesting a severe bottleneck may have resulted in the collapse of this population and, thus, a surfeit of deletions accumulating. While the overall number of base pairs affected by duplication far exceeds that affected by deletion, the relative contribution of duplicated base pairs to deleted base pairs is markedly diverse among different lineages. Duplicated base pairs outnumber deleted base pairs by a factor 6.3 in the African great ape ancestor. In the human ancestor and the chimpanzee ancestor, these ratios are 1.3 and 0.76, respectively, exhibiting a markedly higher relative impact of deletions in these lineages.

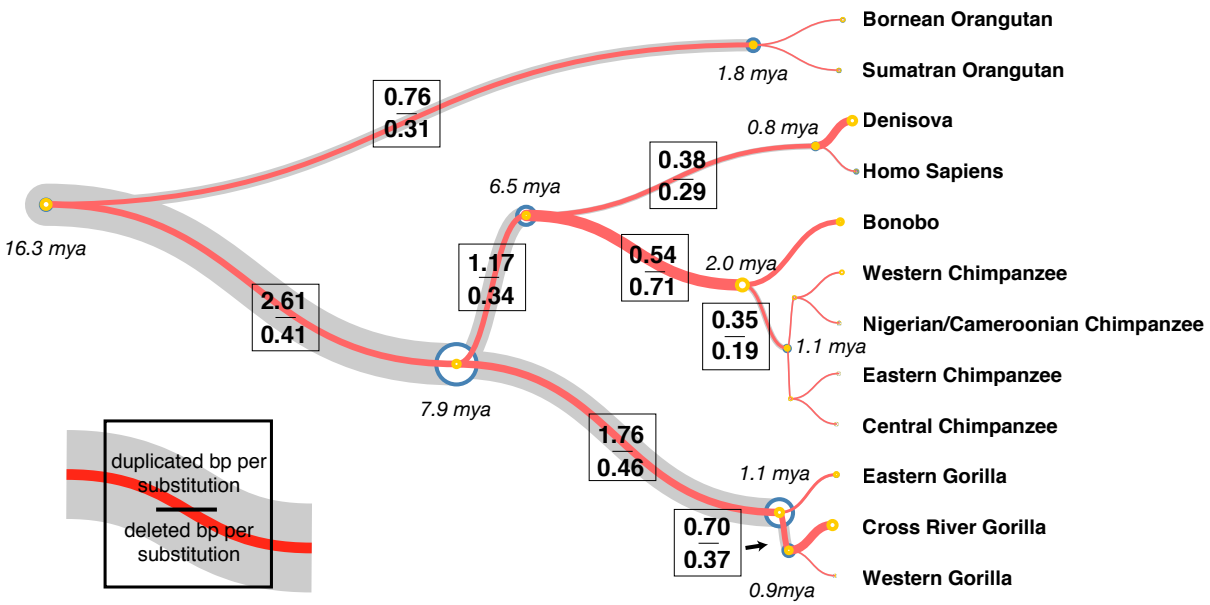


Figure 5.5: The superimposed duplication and deletion rate trees with all nonterminal branches labeled with the duplication and deletion rates per substitution. In the *Pan* lineage the rate of deletion has dramatically increased with respect to the rate of duplication.

Section 6: Lineage-specific deletions and diversity of sequence not represented in the human reference genome

Evolution and diversity of human reference deletions: Of the 9855 loci present in nonhuman primate reference genomes but not represented in the human reference genome, 5777 loci (13.54 Mbp) corresponded to lineage-specific fixed deletions and were assigned to their respective branches (see **Section 5**, **Table 6.1**, **Figure 6.1**). In addition to these loci identified as fixed deletions, 52 loci encompassing 179 kbp of sequence were not present in any of the individuals assessed, suggesting that these loci largely represent mis-assemblies in addition to rapidly diverging sequence and copy number polymorphic loci.

Table 6.1: Summary of sequence absent from the human reference genome.
***Sequence identified to be segregating at >5% frequency in 620 diverse humans.**

Lineage	Fixed deletions bp (loci)	Duplicated copy corrected bp (loci)	Copy number polymorphic bp (loci)
Human	3588396 (917)	168165 (2)	695007 (324)*
Human-Pan	1232172 (213)	2529 (1)	-
Human-Pan-Gorilla	8729420 (1185)	11850 (3)	-
Human-Pan-Gorilla-Orangutan	179252 (52)	307249 (46)	-



Figure 6.1: A heatmap of sequence absent from the reference genome. Each column represents a locus; blocks are colored to represent per-population frequency estimates of the presence or absence of each particular sequence. The cladogram to the left represents hierarchical clustering of the populations surveyed and recapitulates all population relationships.

We next strived to identify which of these loci not present in the human reference genome could be found segregating in the human population. Accordingly, we mapped short-read shotgun sequence from 620 diverse humans from 13 different populations sequenced as part of the 1000 Genomes Project(Siepel et al. 2005; 1000 Genomes Project Consortium et al. 2012) to loci not present in the human reference genome and determined the approximate frequency of each of these regions in individual populations based on the *presence/absence* of signal. In total, we identified 316 loci corresponding 683 kbp of sequence segregating in at >5% frequency in at least one population (**Table 6.1**). We note that these are not strictly allele frequencies but frequencies of the presence/absence of any particular sequence, as accurate copy number

genotyping of these sequences is problematic due to their divergence, small size, and the low coverage of these samples. Hierarchical clustering of the 13 populations based on these frequency estimates completely recapitulates all of the inter-population relationships (**Figure 6.2**), including those between admixed populations, and places the African populations as an outgroup to the European, Asian and North American populations as is expected for an out-of-Africa origin of Homo sapiens.

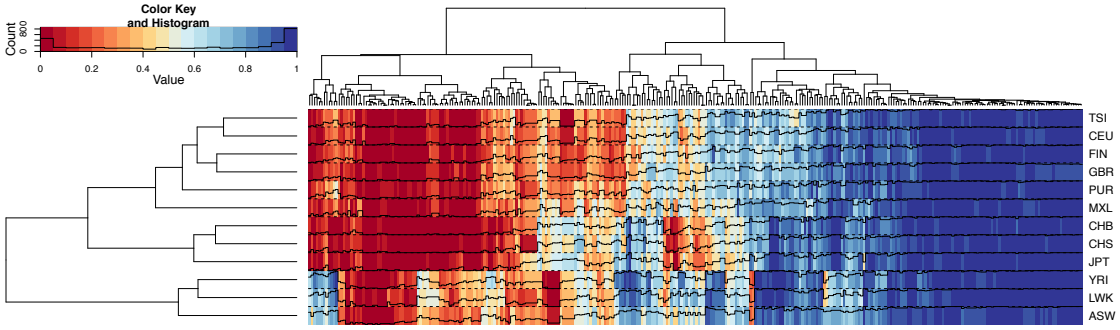


Figure 6.2: A heatmap of sequences genotyped for presence/absence in 620 diverse humans from the 1000 Genomes Project. Columns consist of 324 loci with a frequency >5% in at least one of the populations assessed. Blocks are colored by the per-population frequency and both rows and columns are clustered by hierarchical clustering recapitulating the inter-population relationships.

We calculated a modified Fst (indicated here as Fst*) statistic for each of these loci using the frequency of the presence of a sequence as a surrogate for allele frequency (**Figure 6.3**). Ranking events by these Fst* values, we identified a number of sites of extreme population differentiation, including 53.8 kbp of continental-population-specific sequence absent from all other continental populations groups (52.5 kbp African- and 1.4 kbp European-specific sequence segregating at least 1% frequency) in addition to 77.4 kbp of sequence segregating exclusively in a single continental population and 152.9 kbp of sequence segregating in a single subpopulation at a frequency of at least 1% while fixed in all other populations.

Table 6.1: Population-specific segregating copy number variant loci not found in the human reference genome. (CNVs=copy number variations)

Population	YRI	CHB	ASW	TSI	MXL	LWK	CEU	CHS	GBR	PUR	FIN	JPT	CLM
# of CNVs	5	2	1	4	3	2	2	2	14	11	8	1	5
Total bp sum	44978	4900	871	4319	4895	2002	5494	10092	31903	18243	16614	2744	5798

Functional impact of deleted sequence absent from the human reference: Compared to the human reference genome, sequences absent from the human reference yet present in other nonhuman primate reference genomes are less

likely to be annotated for functional elements. We, thus, performed a number of analyses to assess the functional relevance of sequence deleted throughout the great ape lineage and absent in the human reference. We first assessed the conservation of these elements using Phastcons highly conserved elements (HCEs(Siepel et al. 2005), UCSC genome browser). Among the 13.54 Mbp of human-deleted sequence, we identify 180,522 bp of conserved sequence (2219 total elements) with 9910/2385 human deleted loci (38.1%) encompassing an element. As HCEs encompass 3-8% of the human reference genome(Brawand et al. 2011; Siepel et al. 2005), this represents a depletion of conserved loci in human-deleted loci. We next assessed RefSeq gene and mRNA annotations to identify deleted coding elements. We found three complete or partially deleted genes (**Table 6.2**).

Table 6.2: Annotated genic elements absent from the reference genome and deleted in 620 diverse humans.

Locus	Gene (approved name)	Lineage lost	Description
chr19:56319337-56321545	*SIGLEC13*	Human	Complete gene deletion – previously identified in Want *et al,* 2012. This gene is expressed in chimpanzee monocytes which are responsible for innate immunity to bacteria.
chr19:7839944-7842326	*CD209L2 (CLECM4)*	Human	3 exon deletion of the gene encoding CD209 antigen like protein E. CD209 genes encode for C-type lectins which recognize bacteria, mycobacteria, mycobacteria and, viruses and protozoa. CD209L2 has been previously identified as a human-specific deletion. Oritz *et al,* 2008.
chr1_random:2270147-2276474	*LOC100171780*	Gorilla-Chimp-Human	Uncharacterized protein

As nonhuman primate gene annotations are potentially incomplete, we next assessed RNAseq data from six tissues each in gorillas, humans, chimps, bonobos, and orangutans (55 experiments assessed total) generated in Brawand *et al*(Brawand et al. 2011) (**Table 6.3)**.

Table 6.3: Summary table of 55 RNAseq experiments from Brawand *et al* (Langmead and Salzberg 2012; Brawand et al. 2011; Trapnell et al. 2012) **across all great ape lineages mapped with Bowtie 2.0 to each respective primate reference genome. cb=cerebellum, ts=testis, kd=kidney, ht=heart, br=brain, lv=liver**

ID	Species	Tissue	Sex	Reads Placed (Primate Reference)
GSM752656	ggo	cb	M	11108268
GSM752663	ggo	ts	M	12161822
GSM752659	ggo	kd	F	12879458
GSM752658	ggo	ht	M	15376965
GSM752655	ggo	cb	F	15966074
GSM752657	ggo	ht	F	16347127
GSM752660	ggo	kd	M	19509289
GSM752653	ggo	br	F	20146569
GSM752661	ggo	lv	F	22299848
GSM752662	ggo	lv	M	25664046
GSM752696	hsa	br	M	2145713
GSM752707	hsa	ts	M	3933518
GSM752692	hsa	br	M	9305888
GSM752703	hsa	kd	M	12157025
GSM752701	hsa	ht	M	12557133
GSM752694	hsa	br	M	13052578
GSM752699	hsa	ht	F	13115883
GSM752700	hsa	ht	M	14363375
GSM752702	hsa	kd	F	14780007
GSM752704	hsa	kd	M	15587456
GSM752691	hsa	br	F	15713369
GSM752706	hsa	lv	M	15724554
GSM752697	hsa	cb	F	18974942
GSM752708	hsa	ts	M	20019814
GSM752698	hsa	cb	M	23395954
GSM752705	hsa	lv	M	24466576
GSM752690	ppa	ts	M	7470392
GSM752689	ppa	lv	M	10507559
GSM752685	ppa	ht	M	13320503
GSM752680	ppa	br	F	14135815
GSM752684	ppa	ht	F	14282705
GSM752687	ppa	kd	M	15046031
GSM752686	ppa	kd	F	16010719
GSM752682	ppa	cb	F	16159020
GSM752683	ppa	cb	M	16411415
GSM752688	ppa	lv	F	18219997

GSM752681	ppa	br	M	22253609
GSM752646	ppy	cb	F	11143394
GSM752648	ppy	ht	M	13409196
GSM752649	ppy	kd	F	14743211
GSM752651	ppy	lv	F	15581764
GSM752650	ppy	kd	M	16265868
GSM752644	ppy	br	F	18775829
GSM752647	ppy	ht	F	19782162
GSM752652	ppy	lv	M	22657746
GSM752677	ptr	lv	M	8012561
GSM752664	ptr	br	F	8738268
GSM752671	ptr	cb	M	9975348
GSM752678	ptr	ts	M	11938608
GSM752674	ptr	kd	F	17641193
GSM752670	ptr	cb	F	17948479
GSM752672	ptr	ht	F	18475125
GSM752676	ptr	lv	F	19954292
GSM752673	ptr	ht	M	20050737
GSM752675	ptr	kd	M	24779565

Reads were mapped to each of their respective species' reference genomes separately with the Bowtie 2 aligner using TopHat(Trapnell et al. 2012; Langmead and Salzberg 2012). Transcripts were then assembled using the Cufflinks package(Trapnell et al. 2012) (**Table 6.4**). Both TopHat and Cufflinks were run with default parameters, thus, only transcripts supported by at least 10 RNAseq fragments were reported. Additionally we required transcripts to show intron-exon structure to minimize false positives. Among the identified transcripts, we rediscovered 2/3 annotated loci described above. No evidence of the CD209L2 annotated transcript was found. In macaque CD209L2 is expressed in the liver, spleen, lymph nodes, heart, and skin; however, its paralogs are specifically expressed in dendritic cells and lymph nodes, suggesting that the transcript may simply be absent from the tissues assessed here. We identified 115 deletions along the human lineage encompassing expressed transcripts. Of the 115 regions containing expressed transcripts, 57 (49.6%) additionally contained conserved elements, representing a ~6-fold enrichment for conserved elements in regions containing an expressed transcript compared to genomic background.

Table 6.4: Fixed deletions absent from the human reference genome containing *de novo* assembled transcripts.

Species Reference	Regions containing expressed transcripts (previously annotated transcripts identified)	Regions containing an expressed transcript and conserved elements
ponAbe2	83 (2)	38
gorGor3	1	0
panTro3	57 (1)	19

Great ape gene deletions: We assessed lineage-specific gene deletions throughout the great ape lineage for underlying gene content. For deletions identified from mappings to the human reference genome, RefSeq gene models were used. As nonhuman primate reference genomes are not as well annotated, for assessing the genic content of deletions identified among these references, we use *de novo* assembled transcripts that contain an open reading frame (ORF). ORFs were identified within assembled transcripts using the TransDecoder package (http://transdecoder.sourceforge.net/). In total, we identified 340 genes in which at least one exon was specifically deleted in the

great ape lineage (**Table 6.5**). 29 gene deletions were present in multiple lineages and inconsistent with the overall species tree hence representing cases of either homoplasy or incomplete lineage sorting.

Table 6.5: Lineage-specific genic deletion counts in the great ape lineage. Gene counts represent the total number of fixed lineage-specific genes containing at least a single exon deletion. Rates are calculated per million years (MY) along each branch. A surfeit of gene deletions is observed along the chimpanzee-bonobo lineage.

lineage	genes	exons	genes >50% deleted	gene deletions/MY
Gbeg	1	1	0	0.833333
Gbeg-Ggod-Ggog	35	128	21	5.30303
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	40	67	32	4.761905
Ggod	41	120	11	45.555556
Ggod-Ggog	2	3	0	10
Hde	11	24	6	13.75
Hde-Hsa	37	61	20	6.166667
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	6	6	4	4.285714
Pab	7	7	5	3.888889
Pab-Ppy	90	149	64	6.206897
Ppa	7	18	5	3.5
Ppa-Ptre-Ptrs-Ptrt-Ptrv	57	144	34	12.954545
Ppy	4	5	4	2.222222
Ptre	1	1	1	1.111111
Ptre-Ptrs-Ptrt-Ptrv	1	1	0	1.111111
Ptre-Ptrv	0	0	0	0
Ptrs	1	1	0	1
Ptrs-Ptrt	0	0	0	0
ils_Gbeg-Ggod-Ggog-Hde-Hsa	3	3	2	-
ils_Gbeg-Ggod-Ggog-Pab-Ppy	7	11	6	-
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	20	55	13	-

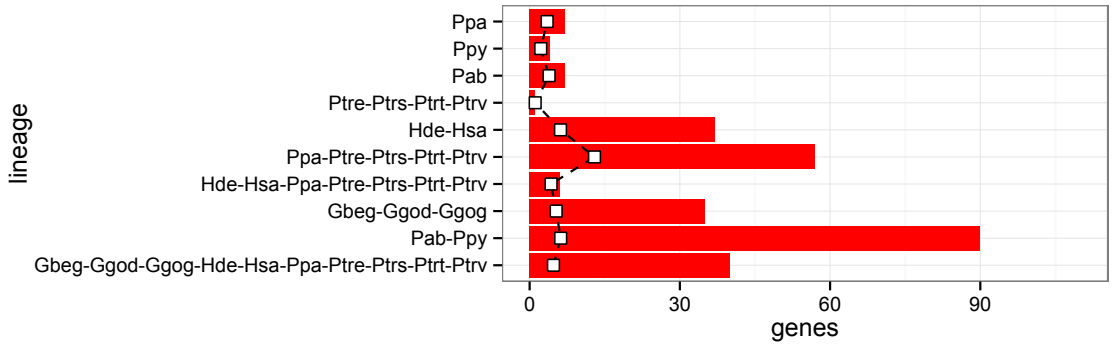


Figure 6.3: A histogram of the number of genes with exons deleted per lineage is plotted in red. White boxes connected by black dotted lines represent the number of gene deletions per million years in each lineage. A 3-fold acceleration in the number of gene deletions is present in the chimpanzee-bonobo lineage.

A list of deleted genes is provided in supplementary **Table 6.6.**

Table 6.6: List of genes with exon encompassing deletions in the great ape lineage. Genes identified along the human lineage are open reading frames identified from assembly of RNAseq data (CUFF prefix).

lineage	gene	gene_coordinate	reference genome	deleted exons	total exons	gene length	genic bp deleted
Ppa	OR11A1	chr6:29501259-29503488	hg18	1	1	2229	155
Ppa	MIR548AJ2	chr4:67035636-67150423	hg18	2	4	114787	2197
Ppa	ANO6	chr12:43896036-44120454	hg18	1	20	224418	2336
Ppa	ACSM1	chr16:20542059-20610079	hg18	4	13	68020	22898
Ppa	IGFL1	chr19:51424848-51426340	hg18	4	4	1492	1492
Ppa	D21S2088E	chr21:23655296-23679027	hg18	3	5	23731	20118
Ppa	SAMD9	chr7:92566761-92585272	hg18	3	3	18511	18511
Ggod	SLMO1	chr18:12397894-12422236	hg18	1	7	24342	2748
Ggod	PRTN3	chr19:791984-799175	hg18	1	5	7191	219
Ggod	MOB1A	chr2:74236718-74259503	hg18	1	6	22785	528
Ggod	VPS37B	chr12:121915827-121946665	hg18	2	4	30838	3463
Ggod	PRMT7	chr16:66902377-66948670	hg18	2	19	46293	1652
Ggod	JOSD2	chr19:55701070-55706157	hg18	5	5	5087	5087
Ggod	TMEM8A	chr16:360776-371951	hg18	13	13	11175	11175
Ggod	CCDC90A	chr6:13894759-13922771	hg18	1	9	28012	1763
Ggod	SKOR1	chr15:65899095-65913228	hg18	9	15	14133	10662
Ggod	LRRC4B	chr19:55711961-55763114	hg18	1	3	51153	6235
Ggod	RYR2	chr1:235272324-236063911	hg18	1	105	791587	3914
Ggod	DAGLB	chr7:6415271-6454168	hg18	1	15	38897	288
Ggod	GRIN1	chr9:139153429-139183035	hg18	1	21	29606	1165
Ggod	AZU1	chr19:778830-783017	hg18	5	5	4187	4187
Ggod	SLC30A4	chr15:43561969-43602294	hg18	2	8	40325	498
Ggod	CACNA1G	chr17:45993427-46059831	hg18	7	38	66404	7074
Ggod	EFNA2	chr19:1237152-1252429	hg18	1	4	15277	1308
Ggod	DCAF15	chr19:13924318-13933256	hg18	3	13	8938	2150
Ggod	ASPDH	chr19:55706668-55709759	hg18	8	8	3091	2286
Ggod	CARD14	chr17:75758385-75797725	hg18	2	23	39340	51
Ggod	GUCY2E	chr11:76068857-76110481	hg18	3	21	41624	3145
Ggod	LGALS3BP	chr17:74478929-74487656	hg18	3	6	8727	4521
Ggod	BRAT1	chr7:2543969-2561918	hg18	1	14	17949	939
Ggod	DOT1L	chr19:2115147-2183577	hg18	16	28	68430	18076
Ggod	E2F6	chr2:11501951-11523748	hg18	1	9	21797	1480
Ggod	EGLN2	chr19:45996887-46006186	hg18	1	7	9299	3271
Ggod	GAK	chr4:833064-916174	hg18	7	28	83110	17294
Ggod	CAMSAP3	chr19:7566787-7589196	hg18	1	19	22409	1378
Ggod	PODNL1	chr19:13902999-13925204	hg18	1	8	22205	14622
Ggod	INPP5F	chr10:121475548-121578652	hg18	1	21	103104	285
Ggod	EXD3	chr9:139321168-139437535	hg18	2	22	116367	22615
Ggod	AGPAT3	chr21:44109543-44231903	hg18	1	11	122360	3230
Ggod	TTC40	chr10:134471885-134606079	hg18	2	58	134194	2358
Ggod	IQCE	chr7:2565157-2620894	hg18	1	22	55737	6389
Ggod	PTX4	chr16:1475940-1478469	hg18	3	3	2529	2529
Ggod	SLC7A6OS	chr16:66892018-66902369	hg18	2	5	10351	1386
Ggod	ZNF843	chr16:31354385-31361849	hg18	1	2	7464	322
Ggod	ABO	chr9:135120383-135140451	hg18	2	7	20068	2562
Ggod	RAB4B-EGLN2	chr19:45975963-46006186	hg18	1	12	30223	3271
Ggod	LOC100134368	chr16:372241-382961	hg18	2	3	10720	2245
Ggod	OTUB1	chr11:63509900-63522468	hg18	1	6	12568	1670
Ggod-Ggog	PWRN1	chr15:22354396-22384019	hg18	2	9	29623	9696
Ggod-Ggog	CTTN	chr11:69922259-69960338	hg18	1	18	38079	1146
Ptrv	PDZK1	chr1:144439022-144475563	hg18	1	10	36541	4259
Gbeg-Ggod-Ggog	COL6A4P1	chr3:15181872-15222470	hg18	4	5	40598	30621
Gbeg-Ggod-Ggog	DNAH14	chr1:223183978-223653619	hg18	54	84	469641	289949
Gbeg-Ggod-Ggog	CST11	chr20:23379040-23381482	hg18	3	3	2442	2442
Gbeg-Ggod-Ggog	PIBF1	chr13:72254230-72488592	hg18	1	18	234362	3300
Gbeg-Ggod-Ggog	KLK2	chr19:56068500-56075635	hg18	4	5	7135	4786
Gbeg-Ggod-Ggog	LINC00467	chr1:209622719-209672500	hg18	2	6	49781	12602
Gbeg-Ggod-Ggog	GGPS1	chr1:233558375-233574467	hg18	1	5	16092	1376
Gbeg-Ggod-Ggog	BTNL3	chr5:180348450-180366333	hg18	4	8	17883	13946
Gbeg-Ggod-Ggog	CCL26	chr7:75236777-75257000	hg18	3	4	20223	19842
Gbeg-Ggod-Ggog	OR4F15	chr15:100175912-100176851	hg18	1	1	939	939

Gbeg-Ggod-Ggog	ADAM2	chr8:39720411-39814936	hg18	2	21	94525	15908
Gbeg-Ggod-Ggog	EID2B	chr19:44713469-44715334	hg18	1	1	1865	1386
Gbeg-Ggod-Ggog	JAKMIP3	chr10:133768302-133848303	hg18	1	24	80001	450
Gbeg-Ggod-Ggog	FAM75E1	chr9:89687591-89693634	hg18	4	4	6043	6043
Gbeg-Ggod-Ggog	GPX6	chr6:28579051-28591549	hg18	1	5	12498	1121
Gbeg-Ggod-Ggog	LOC100294145	chr6:32969930-32979513	hg18	1	2	9583	7010
Gbeg-Ggod-Ggog	ZNF766	chr19:57464635-57487788	hg18	1	4	23153	7483
Gbeg-Ggod-Ggog	MS4A13	chr11:60039461-60066767	hg18	4	7	27306	20083
Gbeg-Ggod-Ggog	CSDAP1	chr16:31486588-31488346	hg18	1	1	1758	1758
Gbeg-Ggod-Ggog	OR6B3	chr2:240633166-240634162	hg18	1	1	996	204
Gbeg-Ggod-Ggog	TAKR	chr10:4903858-4948465	hg18	4	11	44607	8008
Gbeg-Ggod-Ggog	SPINK5	chr5:147423727-147497118	hg18	2	34	73391	3599
Gbeg-Ggod-Ggog	LOC147093	chr17:36959383-36964273	hg18	1	3	4890	2588
Gbeg-Ggod-Ggog	DEFB114	chr6:50035963-50039777	hg18	2	2	3814	3814
Gbeg-Ggod-Ggog	OR2Z1	chr19:8702390-8703335	hg18	1	1	945	945
Gbeg-Ggod-Ggog	SPATA21	chr1:16597724-16636506	hg18	1	13	38782	2665
Gbeg-Ggod-Ggog	OR2G6	chr1:246751570-246752521	hg18	1	1	951	951
Gbeg-Ggod-Ggog	PRR21	chr2:240629902-240631072	hg18	1	1	1170	1170
Gbeg-Ggod-Ggog	KLK3	chr19:56049982-56055832	hg18	1	5	5850	1099
Gbeg-Ggod-Ggog	OLAH	chr10:15125900-15155857	hg18	7	9	29957	23326
Gbeg-Ggod-Ggog	RPS7P5	chr1:238237446-238243183	hg18	1	2	5737	4430
Gbeg-Ggod-Ggog	ADAM5P	chr8:39291338-39379532	hg18	2	15	88194	3346
Gbeg-Ggod-Ggog	SELV	chr19:44697592-44703166	hg18	5	6	5574	4405
Gbeg-Ggod-Ggog	GML	chr8:143913218-143925264	hg18	4	4	12046	12046
Gbeg-Ggod-Ggog	KDM4C	chr9:6710862-7165648	hg18	1	18	454786	6811
ils_Gbeg-Ggod-Ggog-Pab-Ppy	LOC348120	chr15:19410425-19463292	hg18	1	5	52867	8585
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR4F3__OR4F16__OR4F29	chr1:610958-611897	hg18	1	1	939	939
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR4F21	chr8:106085-107024	hg18	1	1	939	939
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR5D16	chr11:55362803-55363790	hg18	1	1	987	987
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR4F16__OR4F3__OR4F29	chr5:180726893-180727832	hg18	1	1	939	939
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR4F13P	chr15:100199844-100208050	hg18	5	5	8206	8206
ils_Gbeg-Ggod-Ggog-Pab-Ppy	OR4F16__OR4F29__OR4F3	chr1:357521-358460	hg18	1	1	939	939
Pab	OR9G4	chr11:56266879-56267863	hg18	1	1	984	984
Pab	OR9G9__OR9G1	chr11:56224439-56225357	hg18	1	1	918	918
Pab	OR5AR1	chr11:56187737-56188670	hg18	1	1	933	933
Pab	OR8U8	chr11:55899675-56265136	hg18	1	3	365461	48285
Pab	OR5AP2	chr11:56165540-56166491	hg18	1	1	951	951
Pab	DGCR5	chr22:17338010-17362142	hg18	1	4	24132	1280
Pab	OR5M1	chr11:56136606-56137554	hg18	1	1	948	948
Ppa-Ptre-Ptrs-Ptrt-Ptrv	MIR4267	chr2:110184826-110184908	hg18	1	1	82	82
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR4M2	chr15:19869841-19870925	hg18	1	1	1084	1084
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ADAM3A	chr8:39427720-39499627	hg18	6	22	71907	19443
Ppa-Ptre-Ptrs-Ptrt-Ptrv	IL36B	chr2:113496138-113526911	hg18	6	7	30773	14398
Ppa-Ptre-Ptrs-Ptrt-Ptrv	CMPK1	chr1:47572055-47617098	hg18	1	7	45043	6354
Ppa-Ptre-Ptrs-Ptrt-Ptrv	IL37	chr2:113387018-113392929	hg18	6	6	5911	3689
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C3	chr12:54011751-54012687	hg18	1	1	936	936
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ADAMTS20	chr12:42034278-42231991	hg18	2	39	197713	2240
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR10A4	chr11:6854431-6855426	hg18	1	1	995	995
Ppa-Ptre-Ptrs-Ptrt-Ptrv	KRTAP19-6	chr21:30835724-30836052	hg18	1	1	328	328
Ppa-Ptre-Ptrs-Ptrt-Ptrv	C17orf57	chr17:42756325-42873676	hg18	9	25	117351	44505
Ppa-Ptre-Ptrs-Ptrt-Ptrv	KRTAP19-7	chr21:30855287-30855479	hg18	1	1	192	192

Ppa-Ptre-Ptrs-Ptrt-Ptrv	C12orf54	chr12:47162552-47176564	hg18	9	9	14012	13096
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ZNF248	chr10:38105459-38187018	hg18	1	7	81559	863
Ppa-Ptre-Ptrs-Ptrt-Ptrv	C11orf65	chr11:107758936-107843468	hg18	1	9	84532	4234
Ppa-Ptre-Ptrs-Ptrt-Ptrv	SULT1C2P1	chr2:108305125-108336686	hg18	4	4	31561	31561
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR4N4	chr15:19883836-19885179	hg18	1	1	1343	146
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR10H1	chr19:15778816-15779936	hg18	1	1	1120	1120
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC100130954	chr9:137606591-137618779	hg18	3	6	12188	9329
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC286177	chr8:58336338-58341724	hg18	2	3	5386	623
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ACSM1	chr16:20542059-20610079	hg18	1	13	68020	5813
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC100289473	chr20:1702010-1708392	hg18	1	4	6382	317
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR4Q3	chr14:19285426-19286368	hg18	1	1	942	942
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C70	chr12:54149250-54150189	hg18	1	1	939	939
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC727924	chr15:19779395-19872452	hg18	7	11	93057	39001
Ppa-Ptre-Ptrs-Ptrt-Ptrv	AFG3L1P	chr16:88566488-88594696	hg18	2	14	28208	2887
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR4M1	chr14:19318321-19319263	hg18	1	1	942	942
Ppa-Ptre-Ptrs-Ptrt-Ptrv	SULT1C3	chr2:108230082-108248239	hg18	3	7	18157	6526
Ppa-Ptre-Ptrs-Ptrt-Ptrv	CYP2C18	chr10:96433240-96485937	hg18	7	9	52697	42888
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LINC00535	chr8:94427870-94781837	hg18	2	11	353967	12616
Ppa-Ptre-Ptrs-Ptrt-Ptrv	CEACAM21	chr19:46774370-46785037	hg18	7	7	10667	10667
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LGALS17A	chr19:44861853-44868853	hg18	8	8	7000	7000
Ppa-Ptre-Ptrs-Ptrt-Ptrv	APOL4	chr22:34915121-34930825	hg18	3	6	15704	10315
Ppa-Ptre-Ptrs-Ptrt-Ptrv	PRSS33	chr16:2773954-2776709	hg18	6	6	2755	2755
Ppa-Ptre-Ptrs-Ptrt-Ptrv	APOL1	chr22:34979062-34993523	hg18	1	7	14461	4080
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR10H5	chr19:15765858-15766806	hg18	1	1	948	948
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ZNF25	chr10:38278800-38305459	hg18	1	6	26659	12554
Ppa-Ptre-Ptrs-Ptrt-Ptrv	CARD18	chr11:104513657-104515671	hg18	3	3	2014	2014
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC100506714	chr22:43908302-43938326	hg18	2	4	30024	10923
Ppa-Ptre-Ptrs-Ptrt-Ptrv	ZNF847P	chr1:225951355-225952031	hg18	2	2	676	676
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR2T10	chr1:246822753-246823692	hg18	1	1	939	939
Ppa-Ptre-Ptrs-Ptrt-Ptrv	CRTAP	chr3:33130453-33164269	hg18	1	7	33816	1732
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC399708	chr10:2332512-2347268	hg18	1	3	14756	3835
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC221442	chr6:41176750-41216551	hg18	3	23	39801	11967
Ppa-Ptre-Ptrs-Ptrt-Ptrv	BTNL2	chr6:32470490-32482878	hg18	2	6	12388	4103
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C75	chr12:54045161-54046100	hg18	1	1	939	939
Ppa-Ptre-Ptrs-Ptrt-Ptrv	TRPC2	chr11:3604265-3615365	hg18	3	7	11100	7951

Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C2	chr12:54132264-54133203	hg18	1	1	939	939
Ppa-Ptre-Ptrs-Ptrt-Ptrv	DEPDC1B	chr5:59928495-60031750	hg18	1	11	103255	9326
Ppa-Ptre-Ptrs-Ptrt-Ptrv	HCG23	chr6:32466264-32469446	hg18	1	3	3182	441
Ppa-Ptre-Ptrs-Ptrt-Ptrv	PRSS22	chr16:2842728-2848172	hg18	1	6	5444	1037
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C65	chr12:54080579-54081518	hg18	1	1	939	939
Ppa-Ptre-Ptrs-Ptrt-Ptrv	DKFZp434J0226	chr19:51405338-51409934	hg18	1	6	4596	935
Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC646214	chr15:19185890-19194116	hg18	1	1	8226	1040
Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR2D2	chr11:6869380-6870307	hg18	1	1	927	927
Ppa-Ptre-Ptrs-Ptrt-Ptrv	B3GNTL1	chr17:78494955-78602975	hg18	4	13	108020	14227
Ppa-Ptre-Ptrs-Ptrt-Ptrv	DEFB128	chr20:116526-118264	hg18	2	2	1738	1738
Hde	HULC	chr6:8597440-8599080	hg18	1	2	1640	164
Hde	GHR	chr5:42459633-42757737	hg18	1	18	298104	3200
Hde	LCE3B	chr1:150852910-150853198	hg18	1	1	288	288
Hde	CHRNA4	chr20:61445105-61463192	hg18	6	6	18087	18087
Hde	LOC729506	chr5:8386595-8510677	hg18	2	6	124082	3523
Hde	LCE3C	chr1:150839761-150840186	hg18	1	1	425	425
Hde	CYP3A43	chr7:99263571-99302109	hg18	2	13	38538	2376
Hde	LOC391322	chr22:22703116-22704043	hg18	2	2	927	927
Hde	GSTT1	chr22:22706138-22714284	hg18	5	5	8146	8146
Hde	GSTTP2	chr22:22715937-22731899	hg18	1	5	15962	4873
Hde	GUCY2GP	chr10:114057925-114106343	hg18	2	19	48418	3461
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	PNLIPRP1	chr10:118340479-118358676	hg18	3	13	18197	7393
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC399815	chr10:124629138-124648220	hg18	1	6	19082	1286
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC201651	chr3:152970933-152985372	hg18	1	3	14439	2595
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CLC	chr19:44913732-44920509	hg18	4	4	6777	6777
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	TRIM61	chr4:166095047-166118268	hg18	2	5	23221	14070
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	KRTAP13-3	chr21:30719581-30720101	hg18	1	1	520	520
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	KRTAP13-4	chr21:30724464-30724947	hg18	1	1	483	483
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR2A7	chr7:143586721-143587654	hg18	1	1	933	933
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC100128593	chr9:138760433-138764184	hg18	3	3	3751	3243
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ADAM5P	chr8:39291338-39379532	hg18	2	15	88194	8332
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LOC728377	chr7:143587021-143614029	hg18	1	13	27008	7642
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	FAM218A	chr4:166097549-166099723	hg18	1	1	2174	2174
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR2A4	chr6:132063301-132064234	hg18	1	1	933	933

ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LCN6	chr9:138758290-138762801	hg18	7	7	4511	4511
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	TAG	chr5:12627968-12858295	hg18	1	4	230327	1864
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR6C76	chr12:54106304-54107243	hg18	1	1	939	939
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ADAM3A	chr8:39427720-39499627	hg18	16	22	71907	52464
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR7G2	chr19:9073944-9074982	hg18	1	1	1038	21
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	LCN10	chr9:138752439-138757232	hg18	6	6	4793	4793
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	OR1M1	chr19:9064920-9065862	hg18	1	1	942	942
Pab-Ppy	OR5D14	chr11:55319607-55320552	hg18	1	1	945	945
Pab-Ppy	OR52E2	chr11:5036455-5037433	hg18	1	1	978	978
Pab-Ppy	OR8H3	chr11:55646424-55647363	hg18	1	1	939	939
Pab-Ppy	CD163L1	chr12:7398823-7488016	hg18	1	20	89193	10192
Pab-Ppy	OR4C11	chr11:55127405-55128450	hg18	1	1	1045	1045
Pab-Ppy	OR5T3	chr11:55776251-55777274	hg18	1	1	1023	1023
Pab-Ppy	OR4P4	chr11:55162409-55163348	hg18	1	1	939	939
Pab-Ppy	ANKRD30B	chr18:14738238-14842737	hg18	15	36	104499	40220
Pab-Ppy	LOC727924	chr15:19779395-19872452	hg18	1	11	93057	11494
Pab-Ppy	OR4C12	chr11:49959584-49960647	hg18	1	1	1063	1063
Pab-Ppy	COL18A1	chr21:45649524-45758062	hg18	1	44	108538	8833
Pab-Ppy	LOC283688	chr15:78342464-78421202	hg18	1	3	78738	1625
Pab-Ppy	OR56A3	chr11:5925152-5926100	hg18	1	1	948	948
Pab-Ppy	LOC348120	chr15:19410425-19463292	hg18	1	5	52867	8148
Pab-Ppy	OR2G2	chr1:245818284-245819238	hg18	1	1	954	954
Pab-Ppy	PCM1	chr8:17824645-17931737	hg18	1	39	107092	3222
Pab-Ppy	MUC4	chr3:196959308-197023241	hg18	2	25	63933	29657
Pab-Ppy	OR8G5	chr11:123639932-123640973	hg18	1	1	1041	10
Pab-Ppy	GIMAP4	chr7:149895390-149901974	hg18	1	3	6584	120
Pab-Ppy	NUP210P1	chr3:127863613-127873472	hg18	2	5	9859	1730
Pab-Ppy	OR5T2	chr11:55756157-55757237	hg18	1	1	1080	1080
Pab-Ppy	KRTAP2-4	chr17:36474893-36475657	hg18	1	1	764	764
Pab-Ppy	OR8U8	chr11:55899675-56265136	hg18	2	3	365461	201070
Pab-Ppy	OR5T1	chr11:55799690-55800671	hg18	1	1	981	981
Pab-Ppy	OR52N2	chr11:5798141-5799107	hg18	1	1	966	450
Pab-Ppy	DDC	chr7:50493627-50600648	hg18	1	17	107021	618
Pab-Ppy	OR8I2	chr11:55617359-55618292	hg18	1	1	933	933
Pab-Ppy	OR5M11	chr11:56066391-56067309	hg18	1	1	918	918
Pab-Ppy	OR4C46	chr11:51371857-51372787	hg18	1	1	930	930
Pab-Ppy	OR5M3	chr11:55993539-55994590	hg18	1	1	1051	1051
Pab-Ppy	OR4S2	chr11:55174955-55175891	hg18	1	1	936	936
Pab-Ppy	ABCA17P	chr16:2330923-2416701	hg18	3	17	85778	5640
Pab-Ppy	MGC2752	chr19:63778577-63787574	hg18	1	2	8997	4051
Pab-Ppy	SNX29	chr16:11978102-12575647	hg18	1	21	597545	28352
Pab-Ppy	SAA3P	chr11:18090594-18094255	hg18	3	3	3661	3661
Pab-Ppy	OR5AS1	chr11:55554470-55555445	hg18	1	1	975	975
Pab-Ppy	CNGA1	chr4:47632750-47709718	hg18	2	12	76968	27934
Pab-Ppy	FAM66B	chr8:7146542-7200286	hg18	2	7	53744	7457
Pab-Ppy	C11orf40	chr11:4549228-4555626	hg18	3	4	6398	1909
Pab-Ppy	OR2G3	chr1:245835510-245836440	hg18	1	1	930	930
Pab-Ppy	OR5M9	chr11:55986520-55987453	hg18	1	1	933	933
Pab-Ppy	OR8H2	chr11:55629094-55630033	hg18	1	1	939	939
Pab-Ppy	OR4C15	chr11:55078358-55079471	hg18	1	1	1113	1084
Pab-Ppy	FLJ26245	chr16:34838423-34848496	hg18	3	3	10073	8880
Pab-Ppy	OR8J3	chr11:55660822-55661770	hg18	1	1	948	948
Pab-Ppy	OR8G1	chr11:123625632-123640966	hg18	1	2	15334	865
Pab-Ppy	OR5D13	chr11:55297489-55298434	hg18	1	1	945	945
Pab-Ppy	OR5J2	chr11:55700669-55701608	hg18	1	1	939	939

Pab-Ppy	OR5W2	chr11:55437701-55438634	hg18	1	1	933	933
Pab-Ppy	OR52J3	chr11:5024331-5025267	hg18	1	1	936	936
Pab-Ppy	OR5D18	chr11:55343681-55344623	hg18	1	1	942	942
Pab-Ppy	OR52E4	chr11:5862098-5863037	hg18	1	1	939	939
Pab-Ppy	OR4A15	chr11:54891935-54892970	hg18	1	1	1035	1035
Pab-Ppy	OR52E6	chr11:5818761-5819703	hg18	1	1	942	942
Pab-Ppy	ANKRD30A	chr10:37454790-37561501	hg18	29	36	106711	89302
Pab-Ppy	OR2AE1	chr7:99311620-99312592	hg18	1	1	972	972
Pab-Ppy	OR10AG1	chr11:55491609-55492515	hg18	1	1	906	906
Pab-Ppy	LOC283553	chr14:50869860-50902025	hg18	1	5	32165	2114
Pab-Ppy	KRTAP5-11	chr11:70970548-70971569	hg18	1	1	1021	1021
Pab-Ppy	UBAC1	chr9:137964635-137993047	hg18	1	10	28412	5548
Pab-Ppy	CECR7	chr22:15897459-15919682	hg18	1	4	22223	2172
Pab-Ppy	RN5S411	chr16:34838394-34838508	hg18	1	1	114	114
Pab-Ppy	LOC100507244	chr9:71957869-71980624	hg18	1	4	22755	3642
Pab-Ppy	MRGPRX3	chr11:18099077-18116603	hg18	1	3	17526	11419
Pab-Ppy	TRIM4	chr7:99325965-99355159	hg18	1	7	29194	668
Pab-Ppy	OR4B1	chr11:48194937-48195867	hg18	1	1	930	930
Pab-Ppy	OR5P2	chr11:7774096-7775065	hg18	1	1	969	969
Pab-Ppy	GCNT3	chr15:57691273-57699502	hg18	1	3	8229	2400
Pab-Ppy	OR4A16	chr11:54867252-54868239	hg18	1	1	987	987
Pab-Ppy	FLJ12334	chr2:19932095-19948289	hg18	1	10	16194	2868
Pab-Ppy	TCL6	chr14:95187267-95209542	hg18	1	8	22275	4983
Pab-Ppy	OR8U1	chr11:55899675-55900605	hg18	1	1	930	930
Pab-Ppy	AMDHD1	chr12:94861201-94886501	hg18	1	9	25300	7526
Pab-Ppy	SWT1	chr1:183392814-183527536	hg18	1	19	134722	2721
Pab-Ppy	OR4A5	chr11:51267953-51269024	hg18	1	1	1071	1071
Pab-Ppy	OR8J1	chr11:55884266-55885340	hg18	1	1	1074	1074
Pab-Ppy	OR5M8	chr11:56014486-56015422	hg18	1	1	936	936
Pab-Ppy	OR8H1	chr11:55814178-55815114	hg18	1	1	936	936
Pab-Ppy	OR5I1	chr11:55459507-55460452	hg18	1	1	945	945
Pab-Ppy	OR52E8	chr11:5834554-5835508	hg18	1	1	954	954
Pab-Ppy	OR5R1	chr11:55941309-55942284	hg18	1	1	975	975
Pab-Ppy	OR52B4	chr11:4345068-4346192	hg18	1	1	1124	1124
Pab-Ppy	OR8K3	chr11:55842358-55843297	hg18	1	1	939	939
Pab-Ppy	OR8K1	chr11:55870090-55871050	hg18	1	1	960	960
Pab-Ppy	OR5F1	chr11:55517732-55518677	hg18	1	1	945	945
Pab-Ppy	CLECL1	chr12:9759722-9777162	hg18	4	4	17440	10808
Pab-Ppy	OR8K5	chr11:55683445-55684369	hg18	1	1	924	924
Pab-Ppy	OR4C16	chr11:55096179-55097112	hg18	1	1	933	933
Pab-Ppy	OR7E5P	chr11:55502754-55510457	hg18	2	4	7703	2347
Pab-Ppy	OR5L1	chr11:55335518-55336454	hg18	1	1	936	936
Ppy	OR4C3	chr11:48303068-48304058	hg18	1	1	990	990
Ppy	OR4C13	chr11:49930518-49931547	hg18	1	1	1029	1029
Ppy	OR4A47	chr11:48466920-48467850	hg18	1	1	930	634
Ppy	OR4C45	chr11:48323475-48330575	hg18	2	2	7100	7100
Ptre	MIR3679	chr2:134601165-134601233	hg18	1	1	68	68
Gbeg	DYNC2H1	chr11:102485369-102855801	hg18	1	90	370432	3263
Ptre-Ptrs-Ptrt-Ptrv	BCRP2	chr22:19787304-19806575	hg18	1	6	19271	12201
Hde-Hsa	CUFF.10122.4__CUFF.10122.3__CUFF.10122.1__CUFF.10122.7__CUFF.10122.5__CUFF.10122.2__CUFF.10122.6	chr10:133269290-133306499	panTro3	1	13	37209	914
Hde-Hsa	CUFF.17274.1	chr12:121725492-121726553	panTro3	1	1	1061	1061
Hde-Hsa	CUFF.25325.2	chr16:21145204-21181787	panTro3	3	5	36583	3349
Hde-Hsa	CUFF.11622.1	chr11:58411158-58441246	panTro3	2	5	30088	5147
Hde-Hsa	CUFF.17260.1	chr12:121714673-121716537	panTro3	1	1	1864	1327
Hde-Hsa	CUFF.37897.1	chr2a:16396704-16432475	ponAbe2	1	6	35771	819
Hde-Hsa	CUFF.35318.1	chr19:43024841-43025176	panTro3	1	1	335	335
Hde-Hsa	CUFF.44627.1	chr2A:108588904-108628721	panTro3	1	8	39817	844
Hde-Hsa	CUFF.45567.1	chr2B:171197850-171216575	panTro3	1	4	18725	1249

Hde-Hsa	CUFF.34.2__CUFF.34.3__CUFF.34.1	chr1:282764-292085	ponAbe2	9	9	9321	9211
Hde-Hsa	CUFF.31789.5__CUFF.31789.4	chr19:49725476-49755401	ponAbe2	1	8	29925	524
Hde-Hsa	CUFF.58246.1__CUFF.58246.3__CUFF.58246.2	chr7:157214104-157216031	ponAbe2	2	2	1927	1927
Hde-Hsa	CUFF.32093.2__CUFF.32093.1__CUFF.32093.3	chr19:53072028-53077831	ponAbe2	3	6	5803	2575
Hde-Hsa	CUFF.11658.1__CUFF.11658.2	chr11:58457432-58510433	panTro3	1	6	53001	2207
Hde-Hsa	CUFF.58469.1	chr6:28457142-28462847	panTro3	1	2	5705	328
Hde-Hsa	CUFF.42529.8__CUFF.42529.7__CUFF.42529.6	chr2A:4996171-5029725	panTro3	1	7	33554	385
Hde-Hsa	CUFF.35324.1	chr19:43029614-43030908	panTro3	1	1	1294	672
Hde-Hsa	CUFF.35320.1	chr19:43025383-43026235	panTro3	1	1	852	852
Hde-Hsa	CUFF.10157.1	chr10:133283836-133287548	panTro3	3	4	3712	3624
Hde-Hsa	CUFF.64217.2__CUFF.64217.3__CUFF.64217.4__CUFF.64217.1	chr7:130910015-130968051	panTro3	1	23	58036	2184
Hde-Hsa	CUFF.19437.1	chr14:21311933-21312805	panTro3	2	2	872	655
Hde-Hsa	CUFF.64478.1__CUFF.64475.1	chrUn:25825909-25829338	ponAbe2	2	2	3429	3429
Hde-Hsa	CUFF.36498.1__CUFF.36498.2	chr19:56318609-56324555	panTro3	2	6	5946	2208
Hde-Hsa	CUFF.65257.1	chr8:912532-913202	panTro3	1	1	670	670
Hde-Hsa	CUFF.56595.1__CUFF.56595.2	chr5:138992707-139049493	panTro3	1	15	56786	641
Hde-Hsa	CUFF.8520.1	chr10:132713266-132713735	ponAbe2	1	1	469	469
Hde-Hsa	CUFF.58612.1	chr8:2039648-2040136	ponAbe2	1	1	488	488
Hde-Hsa	CUFF.11629.1__CUFF.11629.2	chr11:58119514-58130812	panTro3	1	7	11298	832
Hde-Hsa	CUFF.25300.1	chr16:21194196-21194660	panTro3	2	2	464	464
Hde-Hsa	CUFF.37240.3__CUFF.37240.5__CUFF.37240.4	chr22:42137786-42176624	ponAbe2	1	13	38838	407
Hde-Hsa	CUFF.27033.1	chr16:84136515-84137095	panTro3	1	1	580	580
Hde-Hsa	CUFF.59723.10	chr6:52707714-52798280	panTro3	1	36	90566	1950
Hde-Hsa	CUFF.45567.2	chr2B:171197850-171482406	panTro3	1	4	284556	1249
Hde-Hsa	CUFF.29469.1	chr17:26513424-26514673	panTro3	1	1	1249	990
Hde-Hsa	CUFF.38061.1	chr20:23615403-23618960	panTro3	3	3	3557	3458
Hde-Hsa	CUFF.11659.3__CUFF.11659.2__CUFF.11659.1	chr11:58361592-58406822	panTro3	2	7	45230	2832
Hde-Hsa	CUFF.59333.2__CUFF.59333.3	chr6:41677792-41687234	panTro3	2	9	9442	2220
Ptrs	CUFF.22083.1__CUFF.22083.2	chr16:4102852-4117860	ponAbe2	1	3	15008	1004
ils_Gbeg-Ggod-Ggog-Hde-Hsa	CUFF.5904.1	chr1:190489209-190523588	panTro3	1	5	34379	223
ils_ils_Gbeg-Ggod-Ggog-Hde-Hsa	CUFF.17029.1	chr12:112343791-112347550	panTro3	0	1	3759	1567
ils_ils_ils_Gbeg-Ggod-Ggog-Hde-Hsa	CUFF.17023.1	chr12:112338832-112347545	panTro3	0	2	8713	1562

Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.14292.1	chr12:96567618-96581976	ponAbe2	1	3	14358	799
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.60776.1	chr8:150981450-150982002	ponAbe2	1	1	552	552
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.36108.1	chr22:17209875-17273295	ponAbe2	1	2	63420	4313
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.46212.1__CUFF.46202.1__CUFF.46212.2	chr4:3811552-3817381	ponAbe2	1	3	5829	185
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.8505.1	chr10:132698644-132699493	ponAbe2	1	1	849	168
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.8508.1	chr10:132702048-132702523	ponAbe2	1	1	475	41
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.11378.1	chr11:101294540-101308678	ponAbe2	2	11	14138	1767
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.11378.5__CUFF.11378.2__CUFF.11378.3__CUFF.11378.4	chr11:101294540-101332244	ponAbe2	2	14	37704	1767
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.41914.1	chr2b:134898141-134898734	ponAbe2	2	2	593	593
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.48729.1	chr5:964198-966106	ponAbe2	1	1	1908	1177
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.39554.1	chr2a:112233060-112237384	ponAbe2	2	2	4324	4324
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.28964.1	chr18:76813671-76836046	ponAbe2	3	6	22375	6219
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.64170.1	chrUn:7438217-7439337	ponAbe2	1	1	1120	1120
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.28067.1	chr18:17861680-17863215	ponAbe2	1	1	1535	26
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.8960.1	chr11:1058088-1058715	ponAbe2	1	1	627	50
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.46154.1	chr4:3806726-3807539	ponAbe2	1	1	813	813
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.58111.1	chr7:153354665-153355626	ponAbe2	1	1	961	961
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.64392.1	chrUn:12924518-12927973	ponAbe2	2	3	3455	839
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.46180.1__CUFF.46183.1	chr4:3818601-3819247	ponAbe2	1	1	646	646
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.23551.1	chr16:61628302-61629469	ponAbe2	1	1	1167	584
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.64433.1	chrUn:22810038-22810872	ponAbe2	1	1	834	834
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.61033.1	chr8:153430204-153431184	ponAbe2	1	1	980	705
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.58103.1__CUFF.58103.3__CUFF.58103.2	chr7:149328103-149493757	ponAbe2	1	6	165654	1332
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.31191.2__CUFF.31194.1	chr19:40740646-40741970	ponAbe2	1	1	1324	1324
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.31191.1	chr19:40559291-40741970	ponAbe2	2	2	182679	9200

Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.52301.1	chr6:27923306-27923676	ponAbe2	1	1	370	370
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.46200.1	chr4:3821324-3822269	ponAbe2	1	1	945	676
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.28062.3	chr18:17828711-17861419	ponAbe2	4	4	32708	3279
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.1477.2	chr1:67094355-67101144	ponAbe2	1	3	6789	156
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.41905.1	chr2b:134846054-134846662	ponAbe2	1	1	608	608
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.21645.1	chr16:660451-660989	ponAbe2	1	1	538	318
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.31207.1	chr19:40600114-40600861	ponAbe2	1	1	747	747
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.32506.1	chr19:60394990-60395695	ponAbe2	1	1	705	704
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.32059.1	chr19:53342284-53344161	ponAbe2	1	1	1877	716
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.41949.2__CUFF.41949.1	chr2b:134910257-134919987	ponAbe2	4	4	9730	8847
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.46187.1__CUFF.46187.2	chr4:3580724-3603978	ponAbe2	1	8	23254	703
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.64168.1	chrUn:7437751-7438144	ponAbe2	1	1	393	393
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.41951.2__CUFF.41951.1	chr2b:134927186-134929749	ponAbe2	4	4	2563	2185
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.35821.1	chr21:47050673-47054499	ponAbe2	1	2	3826	984
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.55703.3	chr7:9394355-9397801	ponAbe2	1	3	3446	1318
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.41911.1	chr2b:134849552-134894238	ponAbe2	8	8	44686	29684
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.28062.7__CUFF.28062.5__CUFF.28062.6	chr18:17865303-17899684	ponAbe2	2	6	34381	7314
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.58131.1	chr7:149427810-149428488	ponAbe2	1	1	678	144
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.40996.1	chr2b:90877161-90887811	ponAbe2	1	2	10650	894
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.21873.1	chr16:2445452-2495943	ponAbe2	3	9	50491	4487
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	CUFF.64486.1	chrUn:28367950-28368823	ponAbe2	1	1	873	873

Homology of human lineage gene losses: To determine if any of the ORFs deleted along the human lineage had homology to genes identified in other organisms or had been previously predicted as genes, we searched them against the NCBI RefSeq protein database. Of the 86 ORFs lost along the human lineage, 60 contained homology to 42 different genes (**Table 6.7**).

Table 6.7: BLAST results of searching ORFs lost along the human lineage against the RefSeq protein database. We identified 42 significant hits to previously predicted genes/homologs.

lineage	reference	position	exons lost	BLAST result	e-value	HCEs
Hde-Hsa	panTro3	chr10:133269290-133306499	1	gi|397467791|ref|XP_003805587.1| PREDICTED: putative scavenger receptor cysteine-rich domain-containing protein LOC619207-like [Pan paniscus]	0.00E+00	30
Hde-Hsa	panTro3	chr16:21145204-21181787	3	gi|402907936|ref|XP_003916716.1| PREDICTED: ATP-binding cassette sub-family A member 3-like [Papio anubis]	8.65E-51	4
Hde-Hsa	panTro3	chr11:58411158-58441246	2	gi|402893240|ref|XP_003909808.1| PREDICTED: membrane-spanning 4-domains subfamily A member 5-like [Papio anubis]	3.52E-19	2
Hde-Hsa	ponAbe2	chr2a:16396704-16432475	1	gi|297667005|ref|XP_002811788.1| PREDICTED: sulfotransferase 1C3-like [Pongo abelii]	1.96E-119	6
Hde-Hsa	panTro3	chr19:43024841-43025176	1	gi|403293043|ref|XP_003937533.1| PREDICTED: WD repeat-containing protein 87-like [Saimiri boliviensis boliviensis]	6.79E-44	0
Hde-Hsa	panTro3	chr2B:171197850-171216575	1	gi|397507773|ref|XP_003824361.1| PREDICTED: xin actin-binding repeat-containing protein 2 [Pan paniscus]	8.54E-77	14
Hde-Hsa	ponAbe2	chr1:282764-292085	9	gi|297281586|ref|XP_002802121.1| PREDICTED: hypothetical protein LOC100427314 [Macaca mulatta]	1.51E-84	0
Hde-Hsa	ponAbe2	chr19:49725476-49755401	1	gi|426389531|ref|XP_004061173.1| PREDICTED: caspase recruitment domain-containing protein 8 [Gorilla gorilla gorilla]	0.00E+00	0
Hde-Hsa	ponAbe2	chr19:53072028-53077831	3	gi|110835743|ref|NP_001036087.1| sialic acid-binding Ig-like lectin 13 precursor [Pan troglodytes]	0.00E+00	4
Hde-Hsa	panTro3	chr6:28457142-28462847	1	gi|55626286|ref|XP_527293.1| PREDICTED: 60S ribosomal protein L30-like [Pan troglodytes]	6.23E-70	1
Hde-Hsa	panTro3	chr2A:49961715029725	1	gi|114575954|ref|XP_001152938.1| PREDICTED: trafficking protein particle complex subunit 12 isoform 6 [Pan troglodytes] >gi|114575956|ref|XP_001153001.1| PREDICTED:	5.05E-131	15

				trafficking protein particle complex subunit 12 isoform 7 [Pan troglodytes]		
Hde-Hsa	panTro3	chr19:43029614-43030908	1	gi|332206687|ref|XP_003252428.1| PREDICTED: hypothetical protein LOC100595862 [Nomascus leucogenys]	0.00E+00	1
Hde-Hsa	panTro3	chr7:130910015-130968051	1	gi|426357869|ref|XP_004046252.1| PREDICTED: protein FAM40B [Gorilla gorilla gorilla]	0.00E+00	62
Hde-Hsa	panTro3	chr14:21311933-21312805	2	gi|403264885|ref|XP_003924697.1| PREDICTED: uncharacterized protein LOC101031436 [Saimiri boliviensis boliviensis]	4.98E-12	1
Hde-Hsa	panTro3	chr19:56318609-56324555	2	gi|332241162|ref|XP_003269753.1| PREDICTED: LOW QUALITY PROTEIN: sialic acid-binding Ig-like lectin 13-like [Nomascus leucogenys]	8.88E-131	0
Hde-Hsa	panTro3	chr8:912532-913202	1	gi|156369634|ref|XP_001628080.1| predicted protein [Nematostella vectensis]	3.86E-11	0
Hde-Hsa	panTro3	chr5:138992707-139049493	1	gi|297676086|ref|XP_002815977.1| PREDICTED: M-phase inducer phosphatase 3 isoform 3 [Pongo abelii]	0.00E+00	40
Hde-Hsa	ponAbe2	chr8:2039648-2040136	1	gi|156389414|ref|XP_001634986.1| predicted protein [Nematostella vectensis]	4.87E-12	0
Hde-Hsa	panTro3	chr16:21194196-21194660	2	gi|403277334|ref|XP_003930322.1| PREDICTED: ATP-binding cassette sub-family A member 3-like [Saimiri boliviensis boliviensis]	1.76E-48	3
Hde-Hsa	ponAbe2	chr22:42137786-42176624	1	gi|426394877|ref|XP_004063711.1| PREDICTED: ceramide kinase [Gorilla gorilla gorilla]	0.00E+00	16
Hde-Hsa	panTro3	chr6:52707714-52798280	1	gi|297678352|ref|XP_002817041.1| PREDICTED: fibrocystin, partial [Pongo abelii]	0.00E+00	102
Hde-Hsa	panTro3	chr20:23615403-23618960	3	gi|332263715|ref|XP_003280898.1| PREDICTED: cystatin-12-like [Nomascus leucogenys]	3.58E-97	1
Hde-Hsa	panTro3	chr6:41677792-41687234	2	gi|332824003|ref|XP_003311331.1| PREDICTED: adenylate cyclase type 10-like [Pan troglodytes]	2.33E-153	16
ils_Gbeg-Ggod-Ggog-Hde-Hsa	panTro3	chr1:190489209-190523588	1	gi|397486258|ref|XP_003814247.1| PREDICTED: putative uncharacterized protein encoded by LINC00467-like [Pan paniscus] >gi|410034417|ref|XP_001170038.2| PREDICTED: putative uncharacterized protein encoded by LINC00467-like [Pan troglodytes]	2.93E-44	0
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr12:96567618-96581976	1	gi|359065249|ref|XP_002687265.2| PREDICTED: uncharacterized protein LOC100336892 [Bos taurus]	1.99E-24	4
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr22:17209875-17273295	1	gi|395753054|ref|XP_003779528.1| PREDICTED: immunoglobulin omega chain-like [Pongo abelii]	8.11E-64	30
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr10:132702048-132702523	1	gi|410044564|ref|XP_003951836.1| PREDICTED: uncharacterized protein LOC101059682 [Pan troglodytes]	4.63E-16	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr11:101294540-101308678	2	gi|395743436|ref|XP_003777927.1| PREDICTED: caspase-12-like [Pongo abelii]	0.00E+00	9
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2b:134898141-134898734	2	gi|291233023|ref|XP_002736453.1| PREDICTED: hypothetical protein [Saccoglossus kowalevskii]	2.14E-10	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2a:112233060-112237384	2	gi|259500926|ref|ZP_05743828.1| conserved hypothetical protein [Lactobacillus iners DSM 13335]	7.09E-10	1
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr18:76813671-76836046	3	gi|395749962|ref|XP_002828334.2| PREDICTED: uncharacterized protein LOC100442861 [Pongo abelii]	0.00E+00	9
Gbeg-Ggod-Ggog-Hde-	ponAbe2	chr18:17861680-	1	gi|338728618|ref|XP_003365712.1| PREDICTED: zinc finger protein 791-like [Equus caballus]	9.54E-25	2

Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv		17863215				
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr11:1058088-1058715	1	gi|50290505|ref|XP_447684.1| hypothetical protein [Candida glabrata CBS 138]	3.61E-09	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr16:61628302-61629469	1	gi|395748070|ref|XP_003778708.1| PREDICTED: exosome complex component MTR3 [Pongo abelii]	7.26E-80	2
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr8:153430204-153431184	1	gi|46094068|ref|NP_689818.2| zinc finger protein 781 [Homo sapiens]	1.26E-10	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr7:149328103-149493757	1	gi|33186925|ref|NP_057287.2| 5'-AMP-activated protein kinase subunit gamma-2 isoform a [Homo sapiens] >gi|332870104|ref|XP_003318972.1| PREDICTED: 5'-AMP-activated protein kinase subunit gamma-2 isoform 1 [Pan troglodytes]	7.45E-113	28
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:40740646-40741970	1	gi|402905533|ref|XP_003915572.1| PREDICTED: eosinophil lysophospholipase-like [Papio anubis]	1.75E-04	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr6:27923306-27923676	1	gi|350586447|ref|XP_001928408.4| PREDICTED: hypothetical protein LOC100155756 [Sus scrofa]	1.69E-07	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr18:17828711-17861419	4	gi|426387972|ref|XP_004060436.1| PREDICTED: zinc finger protein 14 [Gorilla gorilla gorilla]	2.92E-36	2
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr1:67094355-67101144	1	gi|114568366|ref|XP_001163353.1| PREDICTED: neutrophil cytosol factor 2 isoform 3 [Pan troglodytes] >gi|114568368|ref|XP_001163464.1| PREDICTED: neutrophil cytosol factor 2 isoform 5 [Pan troglodytes]	5.01E-49	3
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:40600114-40600861	1	gi|297277044|ref|XP_001088117.2| PREDICTED: galactoside-binding soluble lectin 13-like [Macaca mulatta]	3.79E-09	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:53342284-53344161	1	gi|426389932|ref|XP_004061370.1| PREDICTED: zinc finger protein 649 [Gorilla gorilla gorilla]	0.00E+00	3
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2b:90877161-90887811	1	gi|109100506|ref|XP_001089951.1| PREDICTED: formimidoyltransferase-cyclodeaminase-like [Macaca mulatta]	6.42E-112	16
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chrUn:28367950-28368823	1	gi|297265250|ref|XP_001107481.2| PREDICTED: hypothetical protein LOC716637 [Macaca mulatta]	4.68E-36	0

Section 7: Gene duplication analysis

Gene duplications: Lineage-specific duplication segments were each assessed for their underlying gene content. RefSeq gene models, collapsed by 80% reciprocal overlap, were used for this analysis. Since many duplicated genes are represented multiple times within the human reference, we avoid over counting lineage-specific gene duplicates for each set of genes identified in a lineage by performing pairwise *bl2seq* (NCBI) alignments between all genic, duplicated segments. If two duplicated genes were found to align to each other with >90% identity reciprocally over at least 10% of their length, they were clustered

together as paralogous. Finally, paralogous gene clusters were manually curated.

In total, 405 genes were identified as lineage-specific duplications with the criteria that at least 50% of the gene model be duplicated (**Figure 7.1, Figure 7.2, Table 7.1**). An increased rate of gene duplication was observed in the African great ape, gorilla, and the human-chimpanzee ancestors with the rate in the human-chimpanzee ancestor showing an ~1.5-fold higher rate than the next highest rate of duplication—the African great ape ancestor (19.3 genes/MY versus 13 genes/MY; **Table 7.2**). **Table 7.1** contains a list of all lineage-specific duplicated genes in great apes. We additionally identified 103 genes duplicated in multiple lineages yet expanded to markedly higher copy in one particular lineage (**Table 7.3**). A heatmap of gene duplications, deletions, and expansions is shown in **Figure 7.3.**

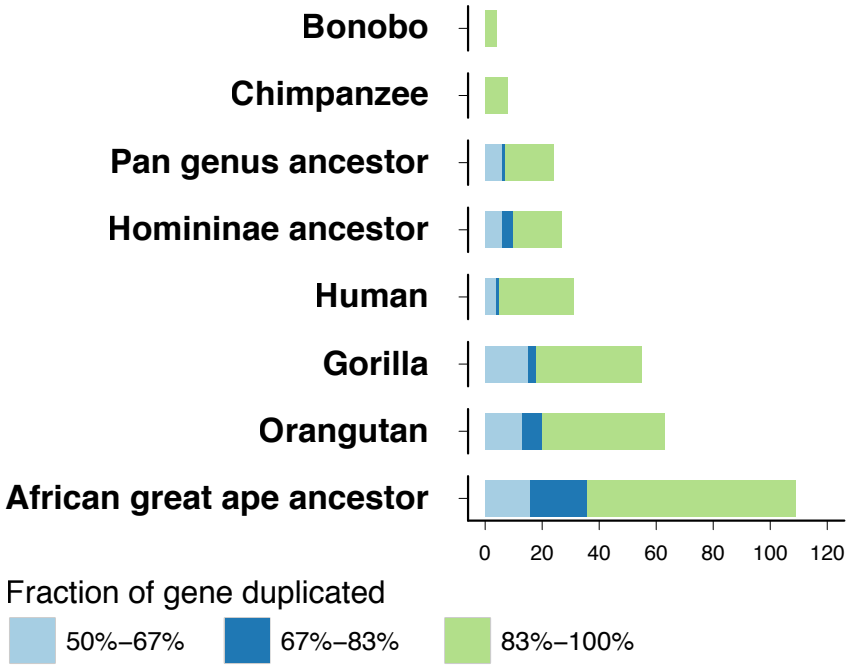


Figure 7.1: Bar chart of the number of lineage-specific gene duplications stratified by the fraction of the gene duplicated.
Table 7.1 is included on an accompanying CD

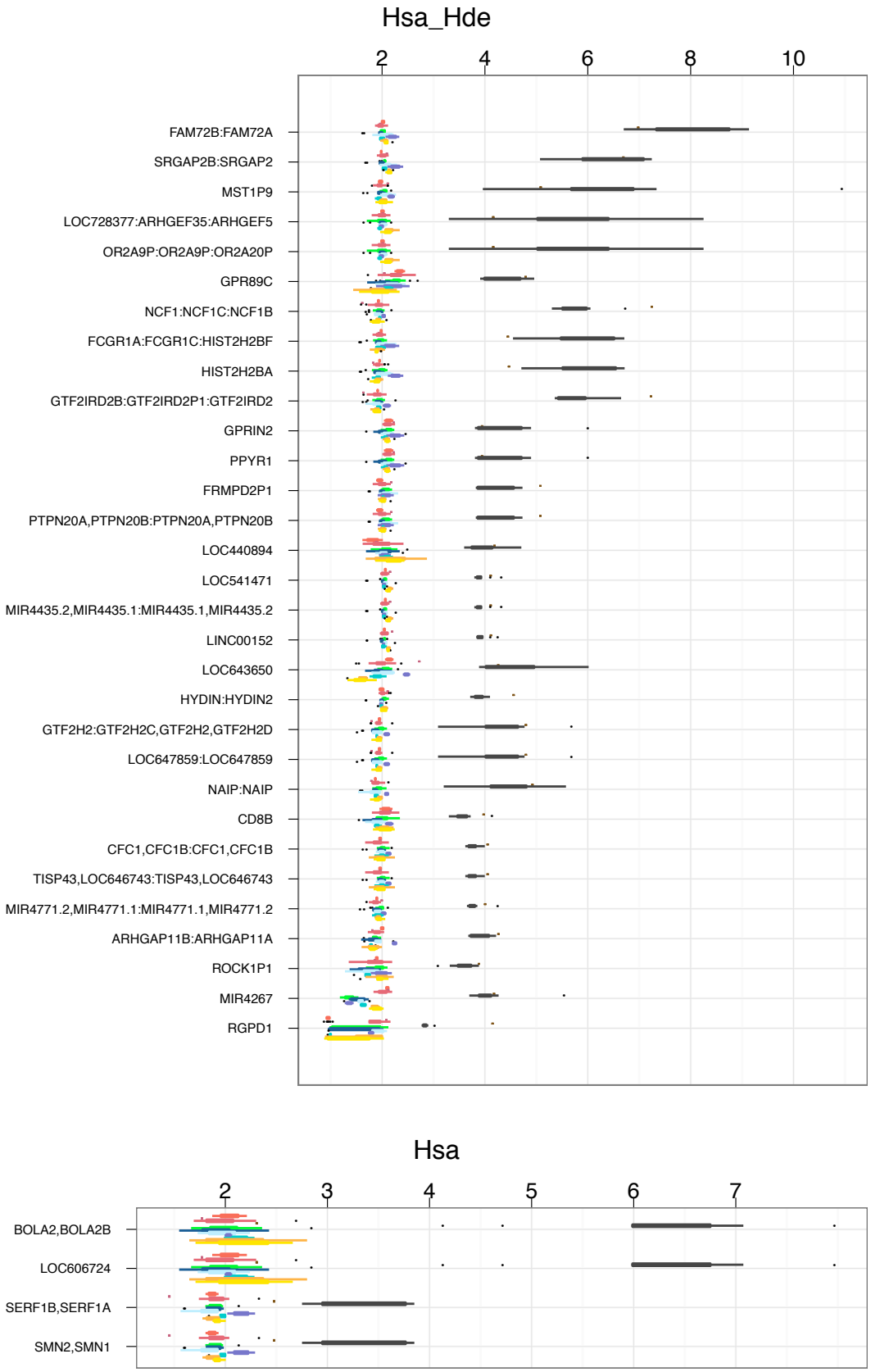
Table 7.2: Counts of lineage-specific gene duplications and rates per million years (MY) of gene duplication in each lineage.

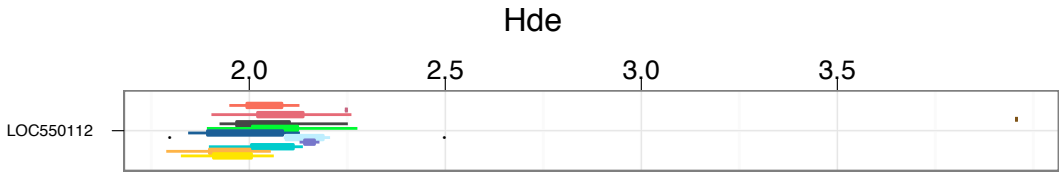
lineage	duplicated_genes	duplications/MY
Ggod-Ggog	2	10
Gbeg-Ggod-Ggog	55	8.333333333
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	109	12.97619048
Hde	1	1.25
Hde-Hsa	31	5.166666667
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	27	19.28571429
Ppa-Ptre-Ptrs-Ptrt-Ptrv	24	5.454545455
Ppa	4	2
Ppy	1	0.555555556
Pab-Ppy	63	4.344827586

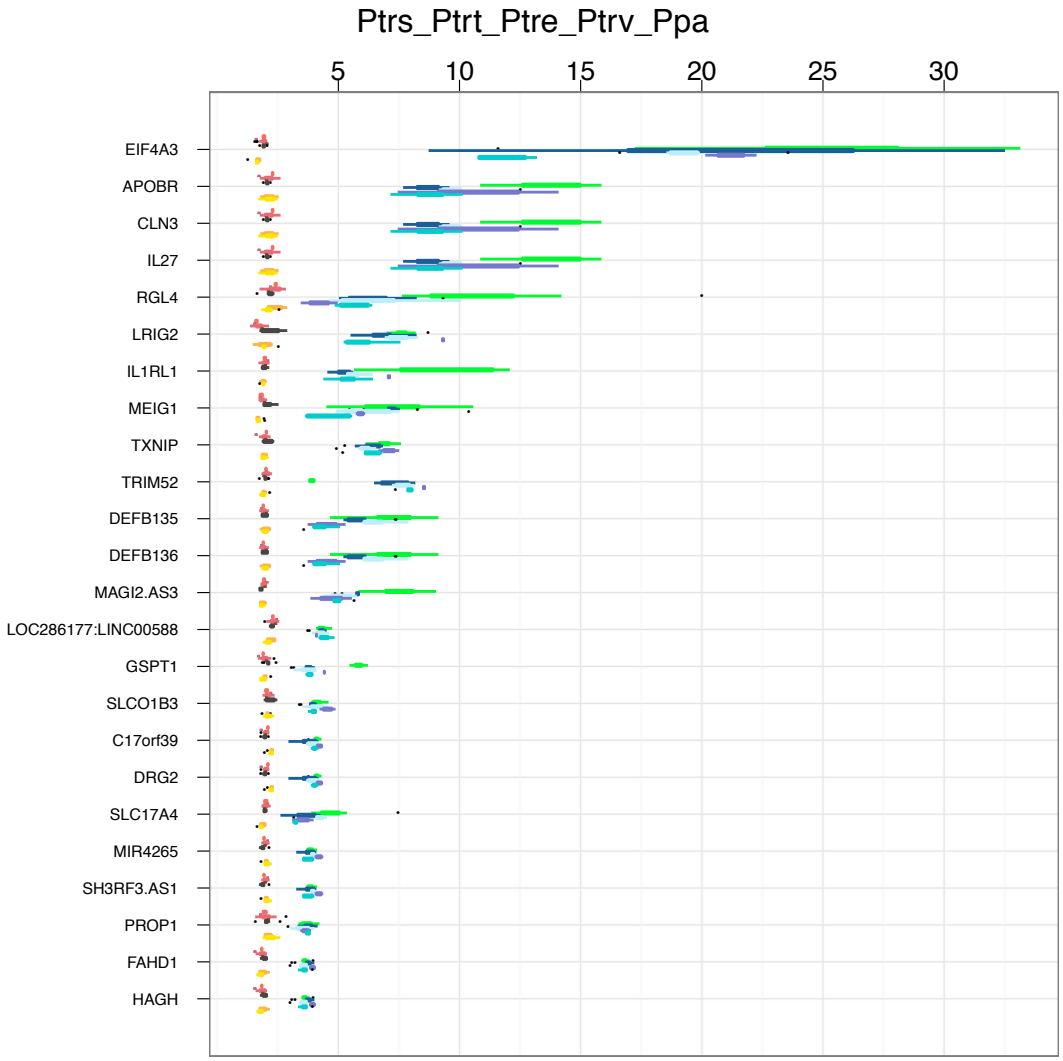
Gbeg	1	0.833333333
Ptre-Ptrs-Ptrt-Ptrv	8	8.888888889
Hsa	4	5
Pab	2	1.111111111

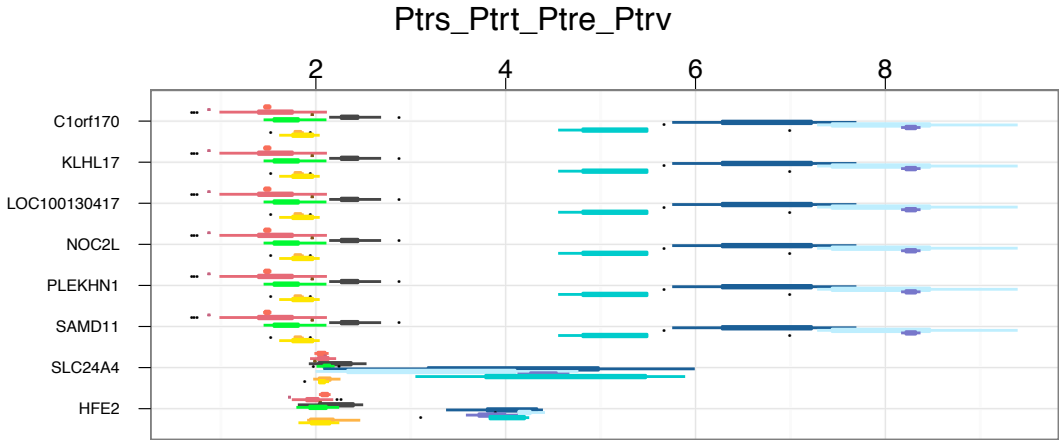
Table 7.3 is included on an accompanying CD

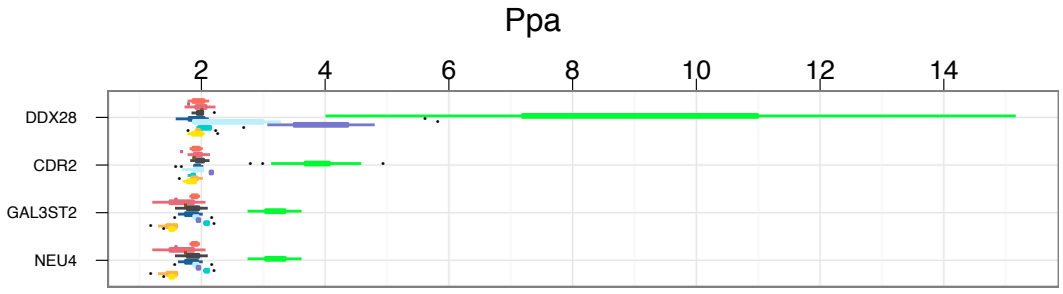
Figure 7.2 (following pages): Lineage-specific gene duplication boxplots for the copy number of each gene in all species assessed in this study.

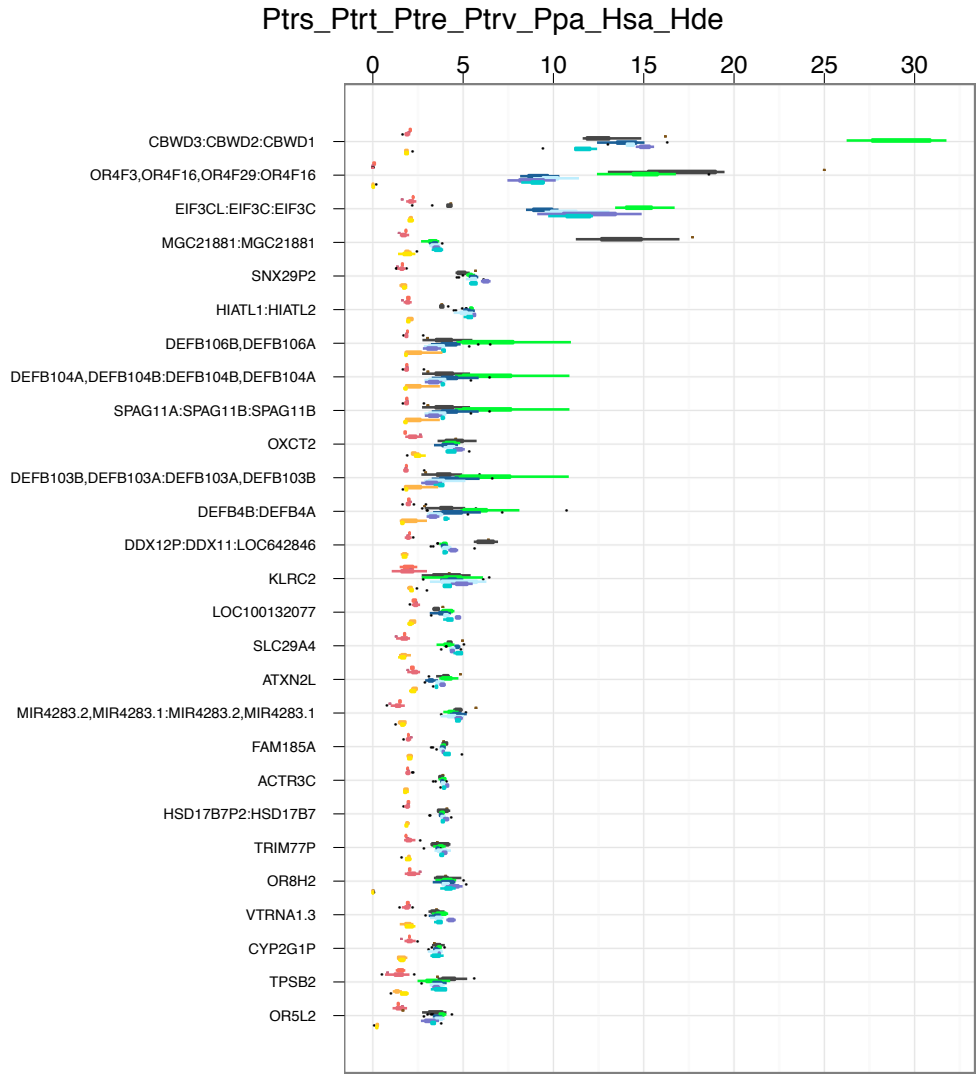


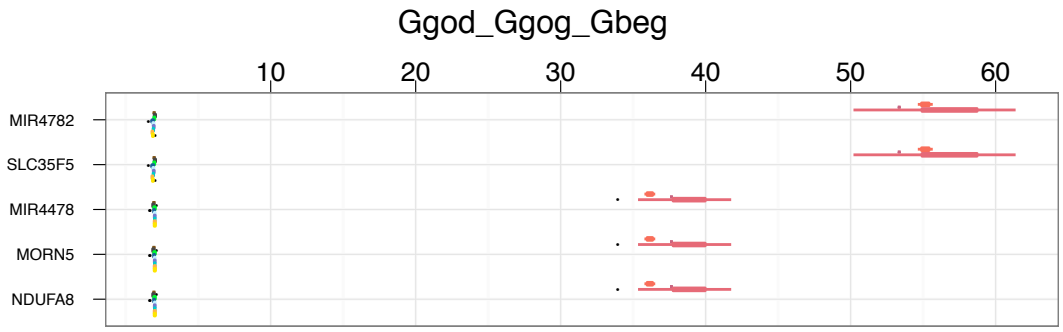

Hde
2.0
2.5
3.0
3.5
LOC550112

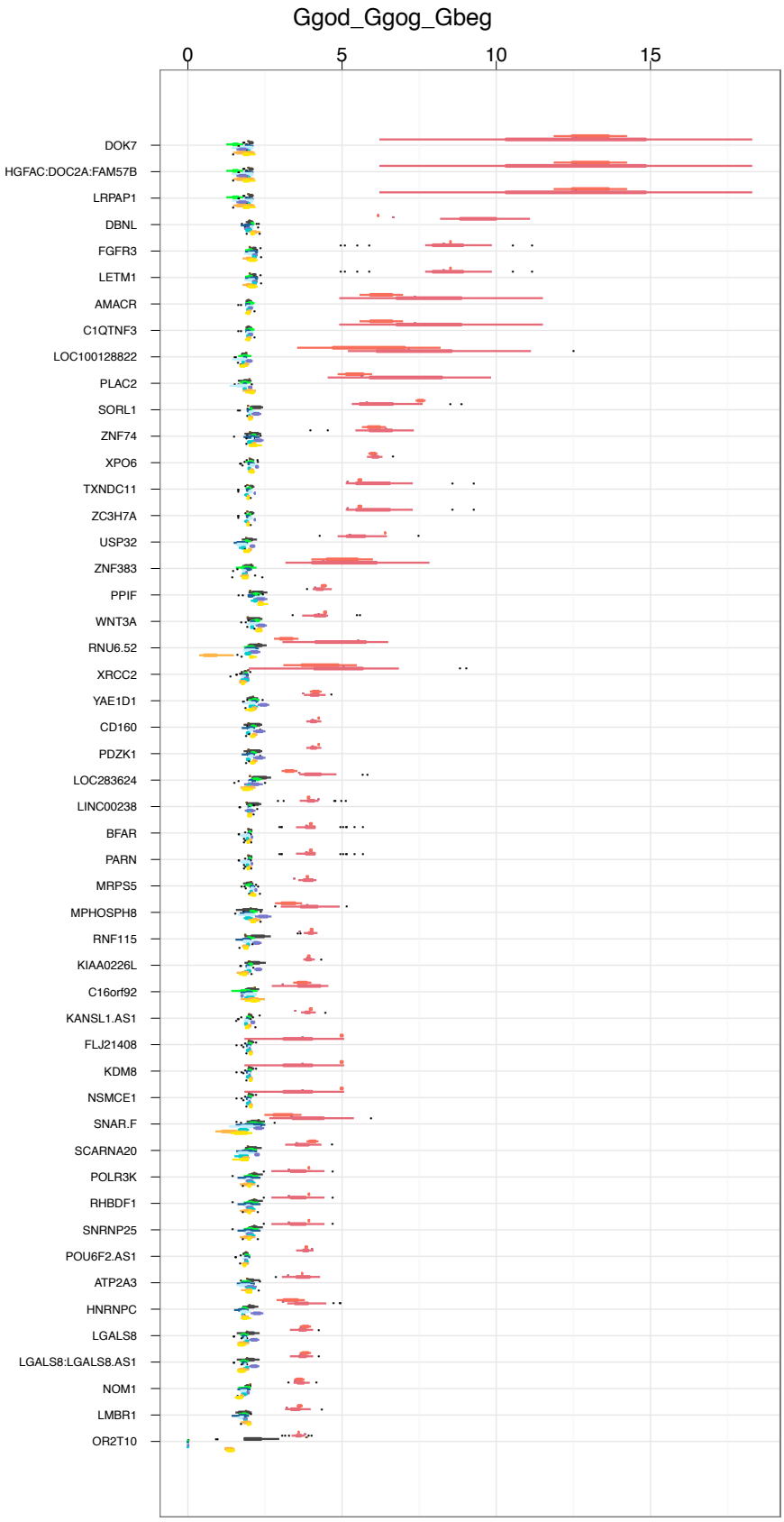

Ptrs_Ptrt_Ptre_Ptrv_Ppa
5
10
15
20
25
30
EIF4A3
APOBR
CLN3
IL27
RGL4
LRIG2
IL1RL1
MEIG1
TXNIP
TRIM52
DEFB135
DEFB136
MAGI2.AS3
LOC286177:LINC00588
GSPT1
SLCO1B3
C17orf39
DRG2
SLC17A4
MIR4265
SH3RF3.AS1
PROP1
FAHD1
HAGH

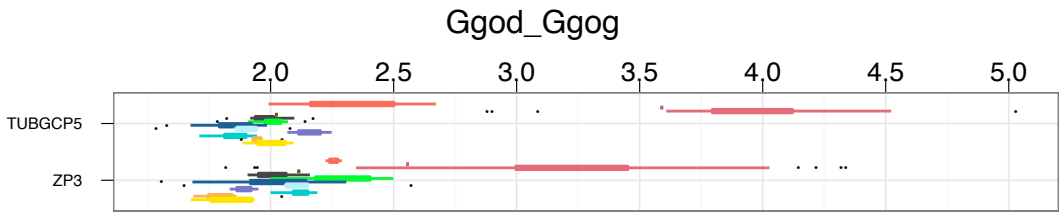

Ptrs_Ptrt_Ptre_Ptrv
2
4
6
8
C1orf170
KLHL17
LOC100130417
NOC2L
PLEKHN1
SAMD11
SLC24A4
HFE2

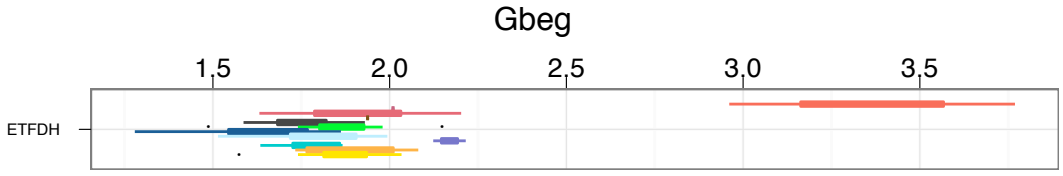

Ppa
2
4
6
8
10
12
14
DDX28
CDR2
GAL3ST2
NEU4

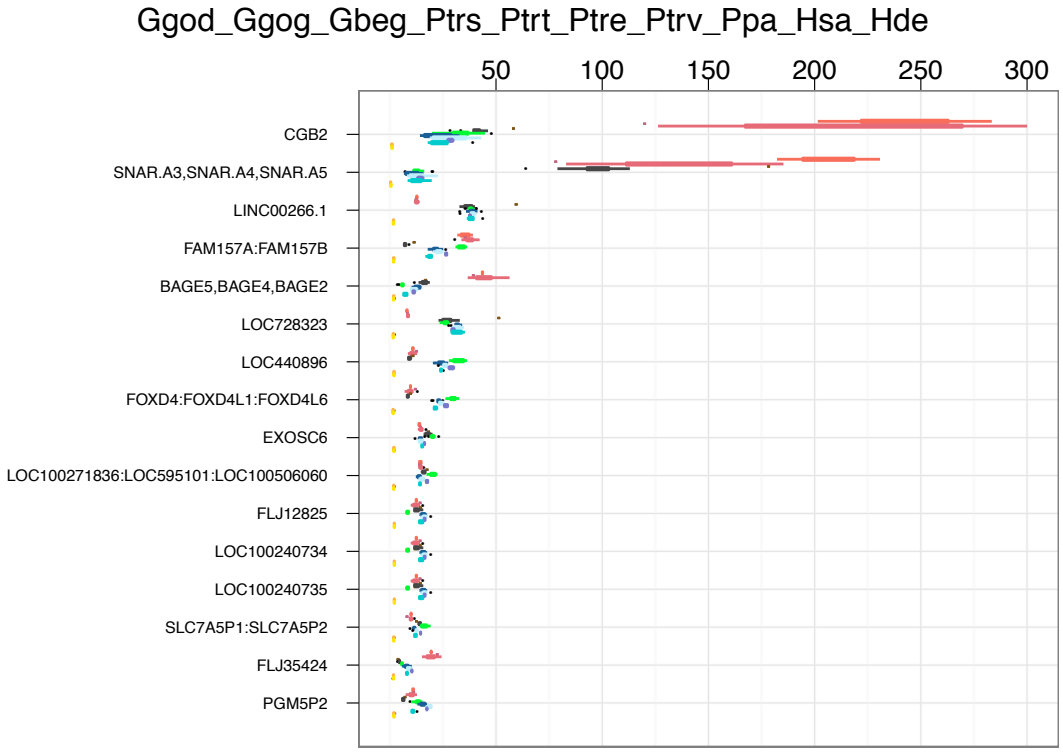

Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
0
5
10
15
20
25
30
CBWD3:CBWD2:CBWD1
OR4F3,OR4F16,OR4F29:OR4F16
EIF3CL:EIF3C:EIF3C
MGC21881:MGC21881
SNX29P2
HIATL1:HIATL2
DEFB106B,DEFB106A
DEFB104A,DEFB104B:DEFB104B,DEFB104A
SPAG11A:SPAG11B:SPAG11B
OXCT2
DEFB103B,DEFB103A:DEFB103A,DEFB103B
DEFB4B:DEFB4A
DDX12P:DDX11:LOC642846
KLRC2
LOC100132077
SLC29A4
ATXN2L
MIR4283.2,MIR4283.1:MIR4283.2,MIR4283.1
FAM185A
ACTR3C
HSD17B7P2:HSD17B7
TRIM77P
OR8H2
VTRNA1.3
CYP2G1P
TPSB2
OR5L2

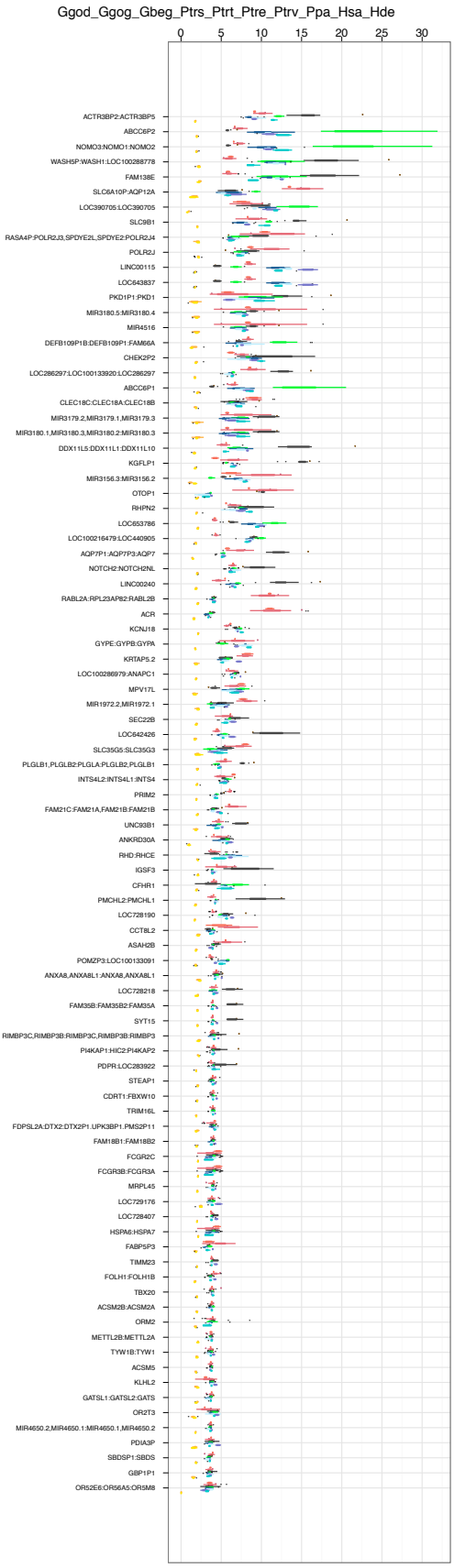

Ggod_Ggog_Gbeg
10
20
30
40
50
60
MIR4782
SLC35F5
MIR4478
MORN5
NDUFA8

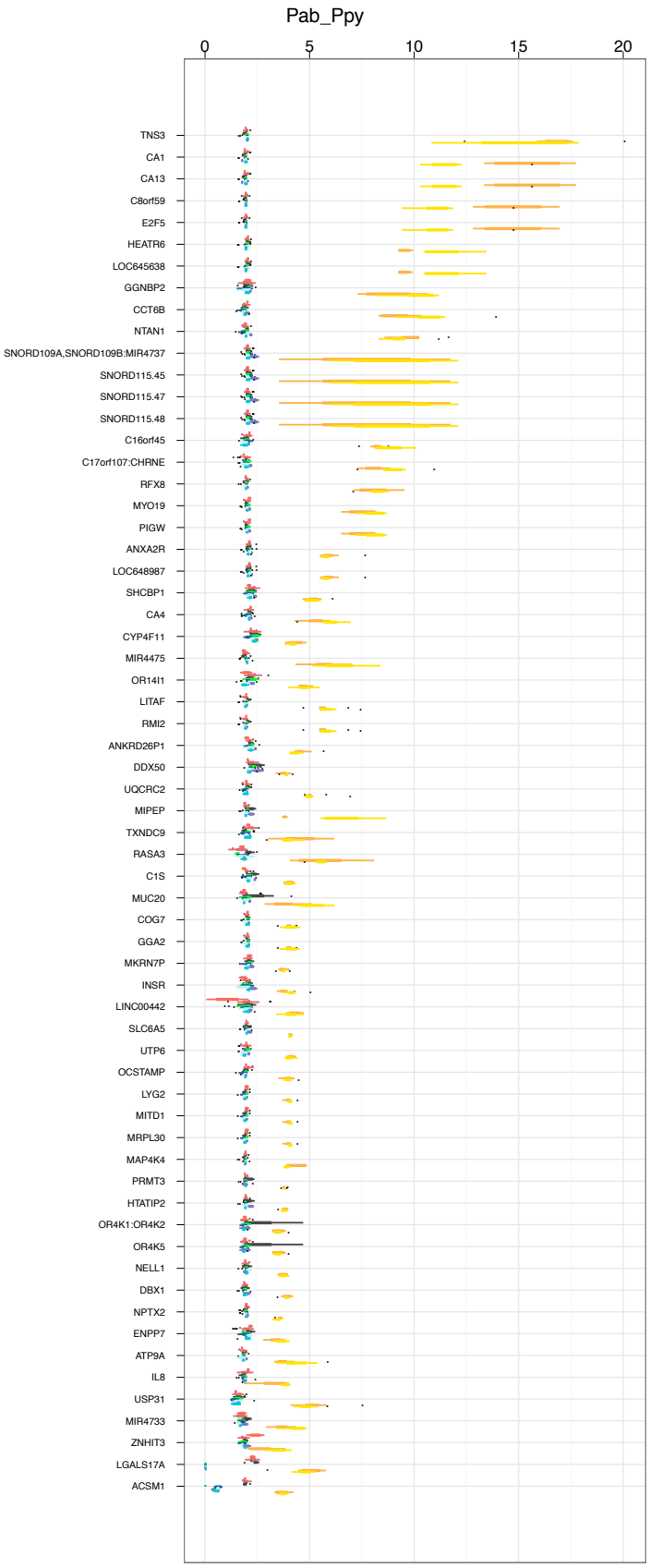

Ggod_Ggog_Gbeg
0
5
10
15
DOK7
HGFAC:DOC2A:FAM57B
LRPAP1
DBNL
FGFR3
LETM1
AMACR
C1QTNF3
LOC100128822
PLAC2
SORL1
ZNF74
XPO6
TXNDC11
ZC3H7A
USP32
ZNF383
PPIF
WNT3A
RNU6.52
XRCC2
YAE1D1
CD160
PDZK1
LOC283624
LINC00238
BFAR
PARN
MRPS5
MPHOSPH8
RNF115
KIAA0226L
C16orf92
KANSL1.AS1
FLJ21408
KDM8
NSMCE1
SNAR.F
SCARNA20
POLR3K
RHBDF1
SNRNP25
POU6F2.AS1
ATP2A3
HNRNPC
LGALS8
LGALS8:LGALS8.AS1
NOM1
LMBR1
OR2T10

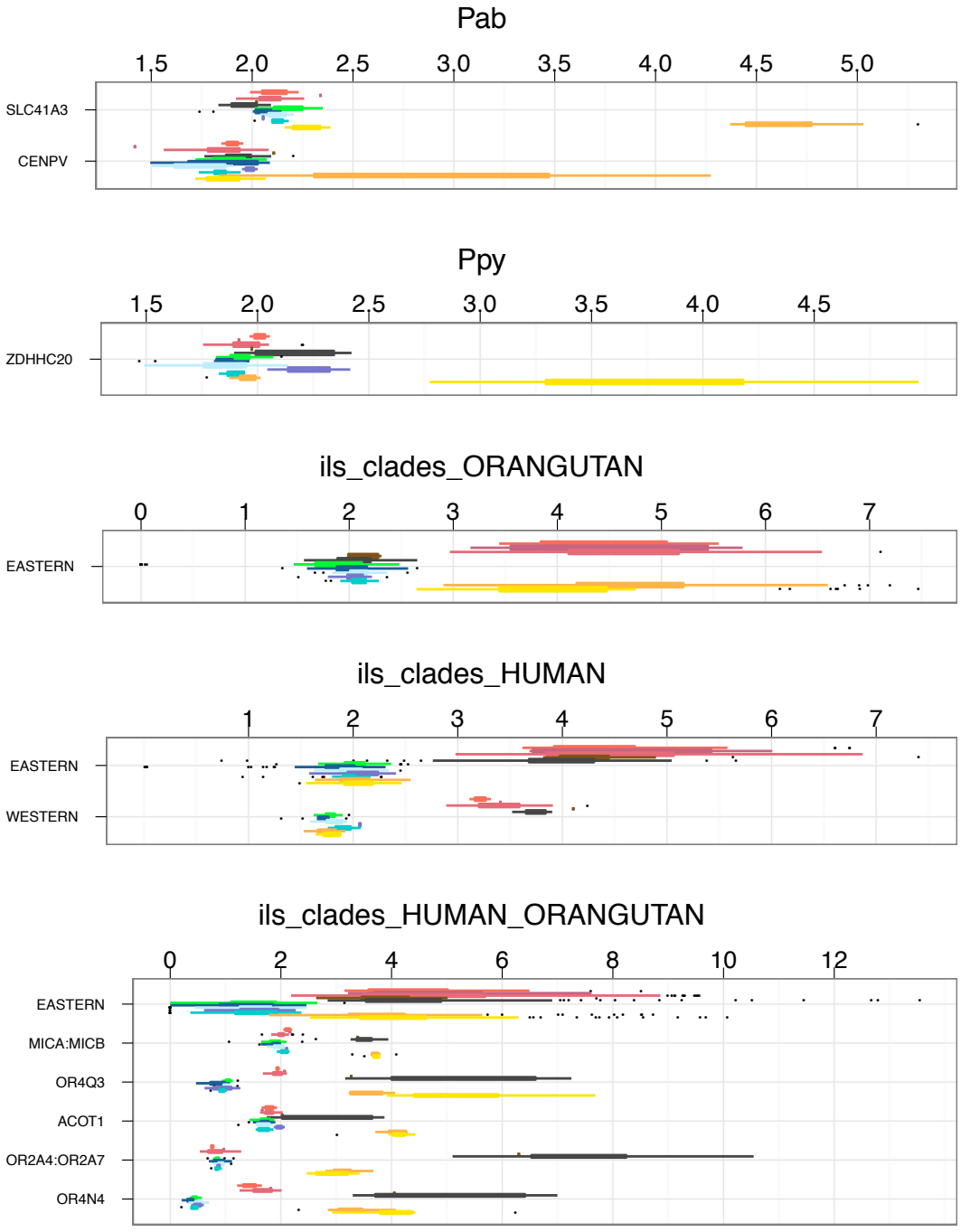

Ggod_Ggog
2.0
2.5
3.0
3.5
4.0
4.5
5.0
TUBGCP5
ZP3


Gbeg
1.5
2.0
2.5
3.0
3.5
ETFDH

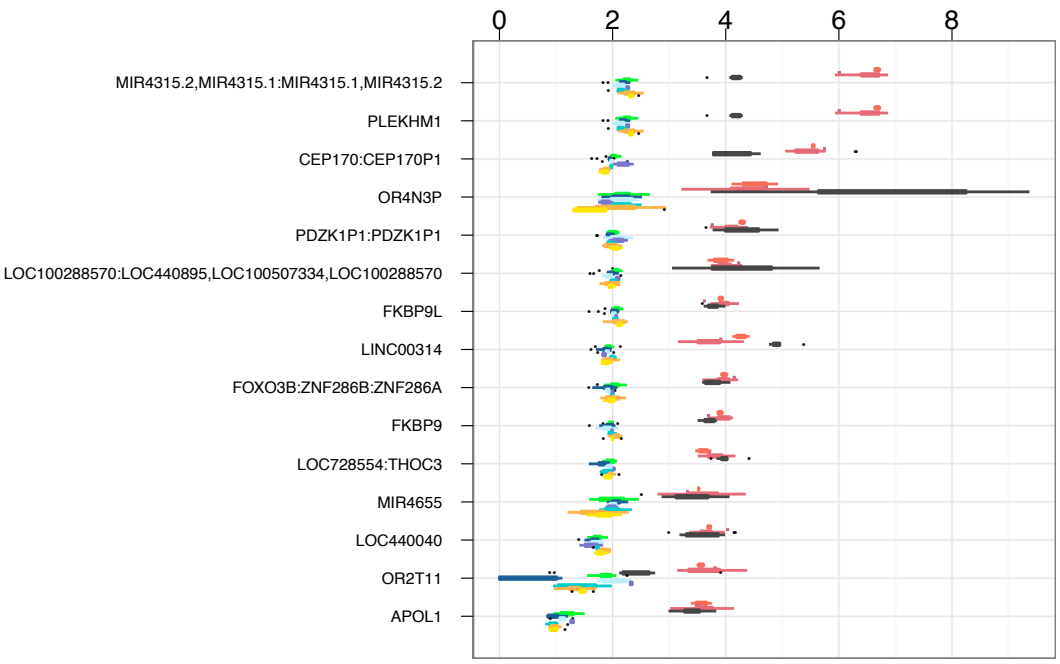

Ggod_Ggog_Gbeg_Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
50
100
150
200
250
300
CGB2
SNAR.A3,SNAR.A4,SNAR.A5
LINC00266.1
FAM157A:FAM157B
BAGE5,BAGE4,BAGE2
LOC728323
LOC440896
FOXD4:FOXD4L1:FOXD4L6
EXOSC6
LOC100271836:LOC595101:LOC100506060
FLJ12825
LOC100240734
LOC100240735
SLC7A5P1:SLC7A5P2
FLJ35424
PGM5P2


Ggod_Ggog_Gbeg_Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
0
5
10
15
20
25
30
ACTR3BP2:ACTR3BP5
ABCC6P2
NOMO3:NOMO1:NOMO2
WASH5P:WASH1:LOC100288778
FAM138E
SLC6A10P:AQP12A
LOC390705:LOC390705
SLC9B1
RASA4P:POLR2J3,SPDYE2L,SPDYE2:POLR2J4
POLR2J
LINC00115
LOC643837
PKD1P1:PKD1
MIR3180.5:MIR3180.4
MIR4516
DEFB109P1B:DEFB109P1:FAM66A
CHEK2P2
LOC286297:LOC100133920:LOC286297
ABCC6P1
CLEC18C:CLEC18A:CLEC18B
MIR3179.2,MIR3179.1,MIR3179.3
MIR3180.1,MIR3180.3,MIR3180.2:MIR3180.3
DDX11L5:DDX11L1:DDX11L10
KGFLP1
MIR3156.3:MIR3156.2
OTOP1
RHPN2
LOC653786
LOC100216479:LOC440905
AQP7P1:AQP7P3:AQP7
NOTCH2:NOTCH2NL
LINC00240
RABL2A:RPL23AP82:RABL2B
ACR
KCNJ18
GYPE:GYPB:GYPA
KRTAP5.2
LOC100286979:ANAPC1
MPV17L
MIR1972.2,MIR1972.1
SEC22B
LOC642426
SLC35G5:SLC35G3
PLGLB1,PLGLB2:PLGLA:PLGLB2,PLGLB1
INTS4L2:INTS4L1:INTS4
PRIM2
FAM21C:FAM21A,FAM21B:FAM21B
UNC93B1
ANKRD30A
RHD:RHCE
IGSF3
CFHR1
PMCHL2:PMCHL1
LOC728190
CCT8L2
ASAH2B
POMZP3:LOC100133091
ANXA8,ANXA8L1:ANXA8,ANXA8L1
LOC728218
FAM35B:FAM35B2:FAM35A
SYT15
RIMBP3C,RIMBP3B:RIMBP3C,RIMBP3B:RIMBP3
PI4KAP1:HIC2:PI4KAP2
PDPR:LOC283922
STEAP1
CDRT1:FBXW10
TRIM16L
FDPSL2A:DTX2:DTX2P1,UPK3BP1,PMS2P11
FAM18B1:FAM18B2
FCGR2C
FCGR3B:FCGR3A
MRPL45
LOC729176
LOC728407
HSPA6:HSPA7
FABP5P3
TIMM23
FOLH1:FOLH1B
TBX20
ACSM2B:ACSM2A
ORM2
METTL2B:METTL2A
TYW1B:TYW1
ACSM5
KLHL2
GATSL1:GATSL2:GATS
OR2T3
MIR4650.2,MIR4650.1:MIR4650.1,MIR4650.2
PDIA3P
SBDSP1:SBDS
GBP1P1
OR52E6:OR56A5:OR5M8

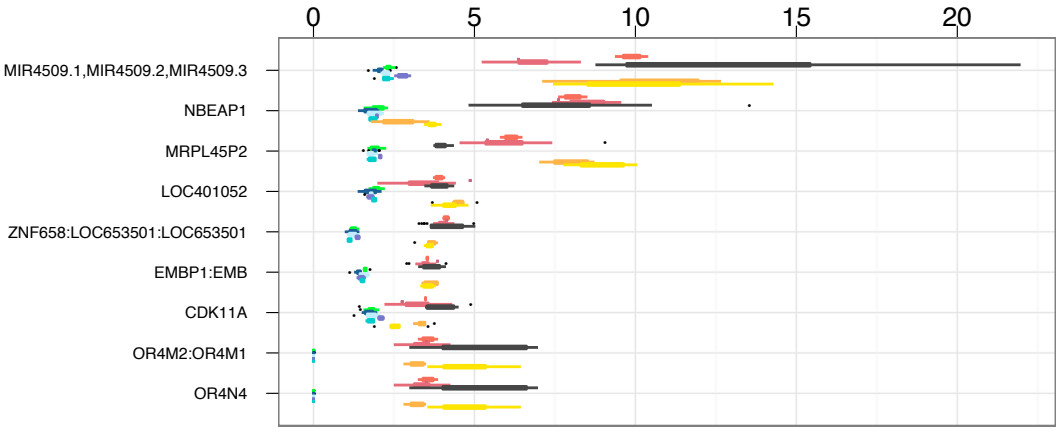

Pab_Ppy
0
5
10
15
20
TNS3
CA1
CA13
C8orf59
E2F5
HEATR6
LOC645638
GGNBP2
CCT6B
NTAN1
SNORD109A,SNORD109B:MIR4737
SNORD115.45
SNORD115.47
SNORD115.48
C16orf45
C17orf107:CHRNE
RFX8
MYO19
PIGW
ANXA2R
LOC648987
SHCBP1
CA4
CYP4F11
MIR4475
OR14I1
LITAF
RMI2
ANKRD26P1
DDX50
UQCRC2
MIPEP
TXNDC9
RASA3
C1S
MUC20
COG7
GGA2
MKRN7P
INSR
LINC00442
SLC6A5
UTP6
OCSTAMP
LYG2
MITD1
MRPL30
MAP4K4
PRMT3
HTATIP2
OR4K1:OR4K2
OR4K5
NELL1
DBX1
NPTX2
ENPP7
ATP9A
IL8
USP31
MIR4733
ZNHIT3
LGALS17A
ACSM1


Pab
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SLC41A3
CENPV
Ppy
1.5
2.0
2.5
3.0
3.5
4.0
4.5
ZDHHC20
ils_clades_ORANGUTAN
0
1
2
3
4
5
6
7
EASTERN
ils_clades_HUMAN
1
2
3
4
5
6
7
EASTERN
WESTERN
ils_clades_HUMAN_ORANGUTAN
0
2
4
6
8
10
12
EASTERN
MICA:MICB
OR4Q3
ACOT1
OR2A4:OR2A7
OR4N4

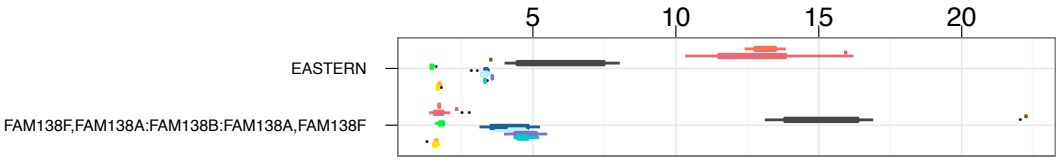
ils_clades_HUMAN_GORILLA



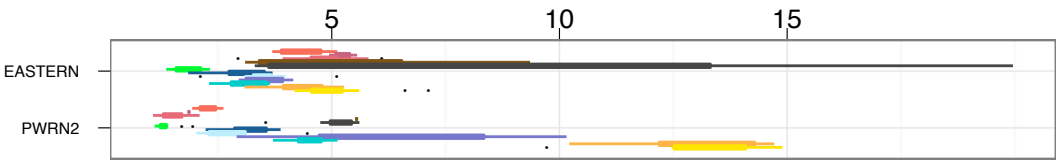
ils_clades_HUMAN_GORILLA_ORANGUTAN

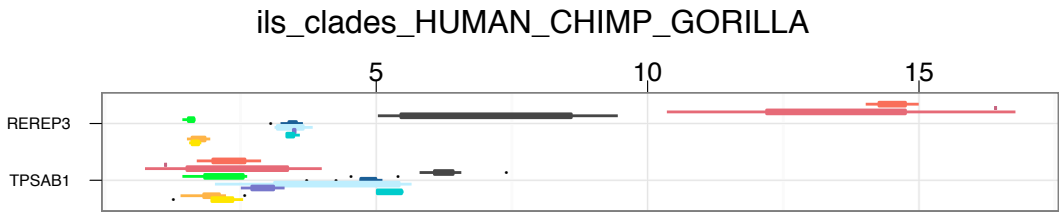


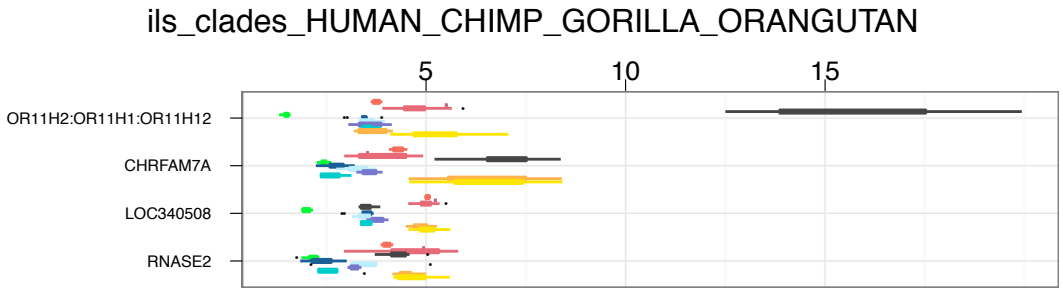
ils_clades_HUMAN_CHIMP

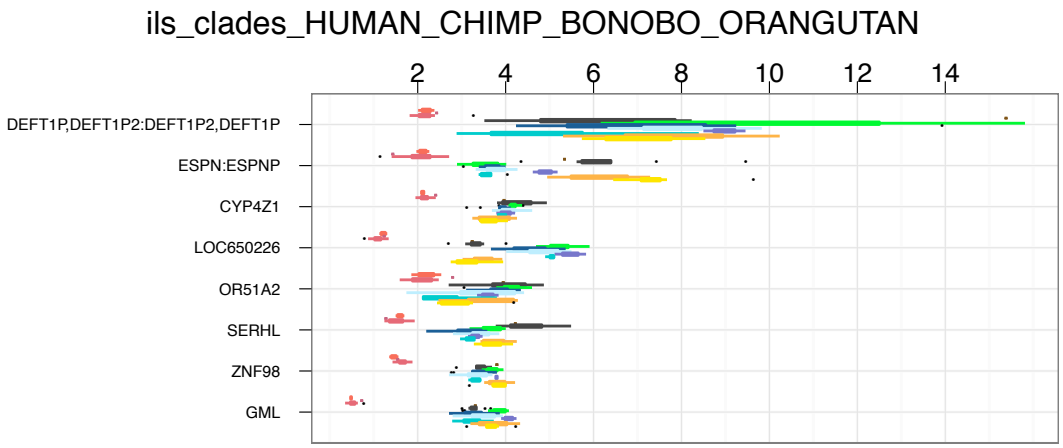


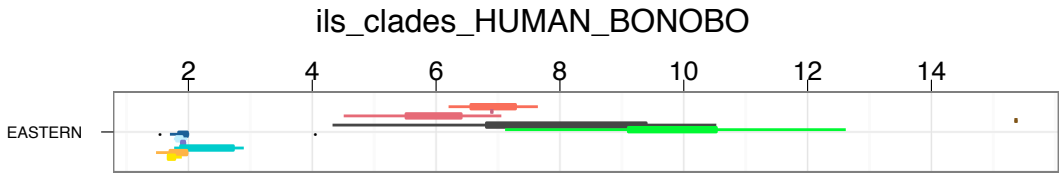
ils_clades_HUMAN_CHIMP_ORANGUTAN

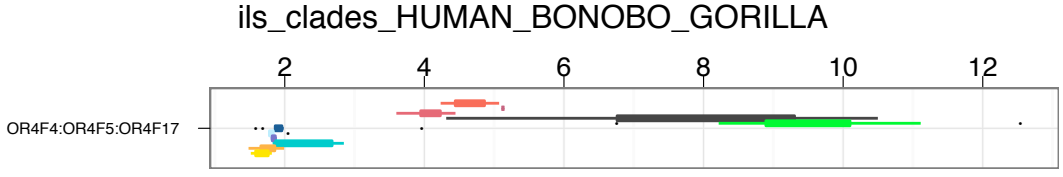


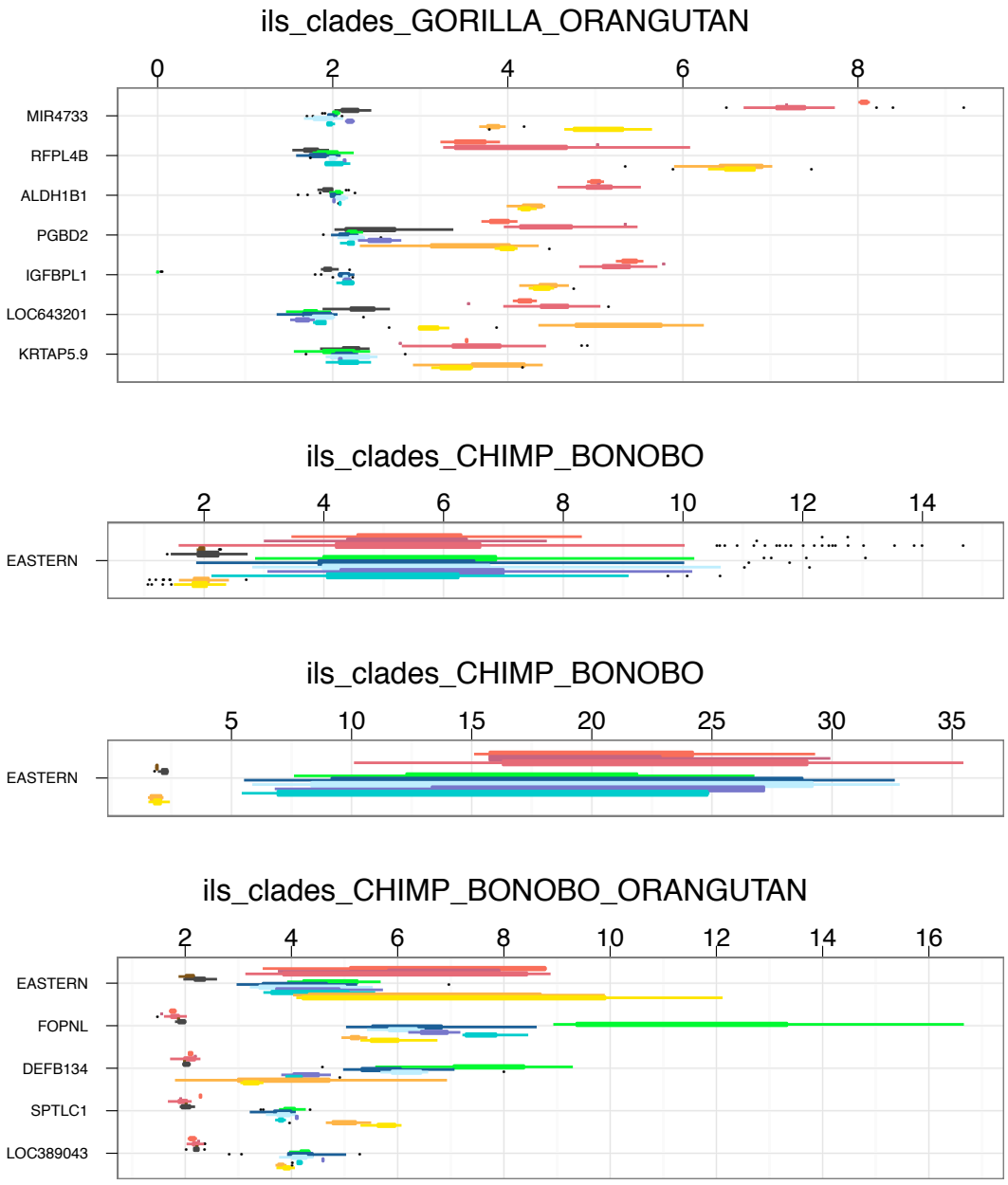

ils_clades_HUMAN_CHIMP_GORILLA
5
10
15
REREP3
TPSAB1

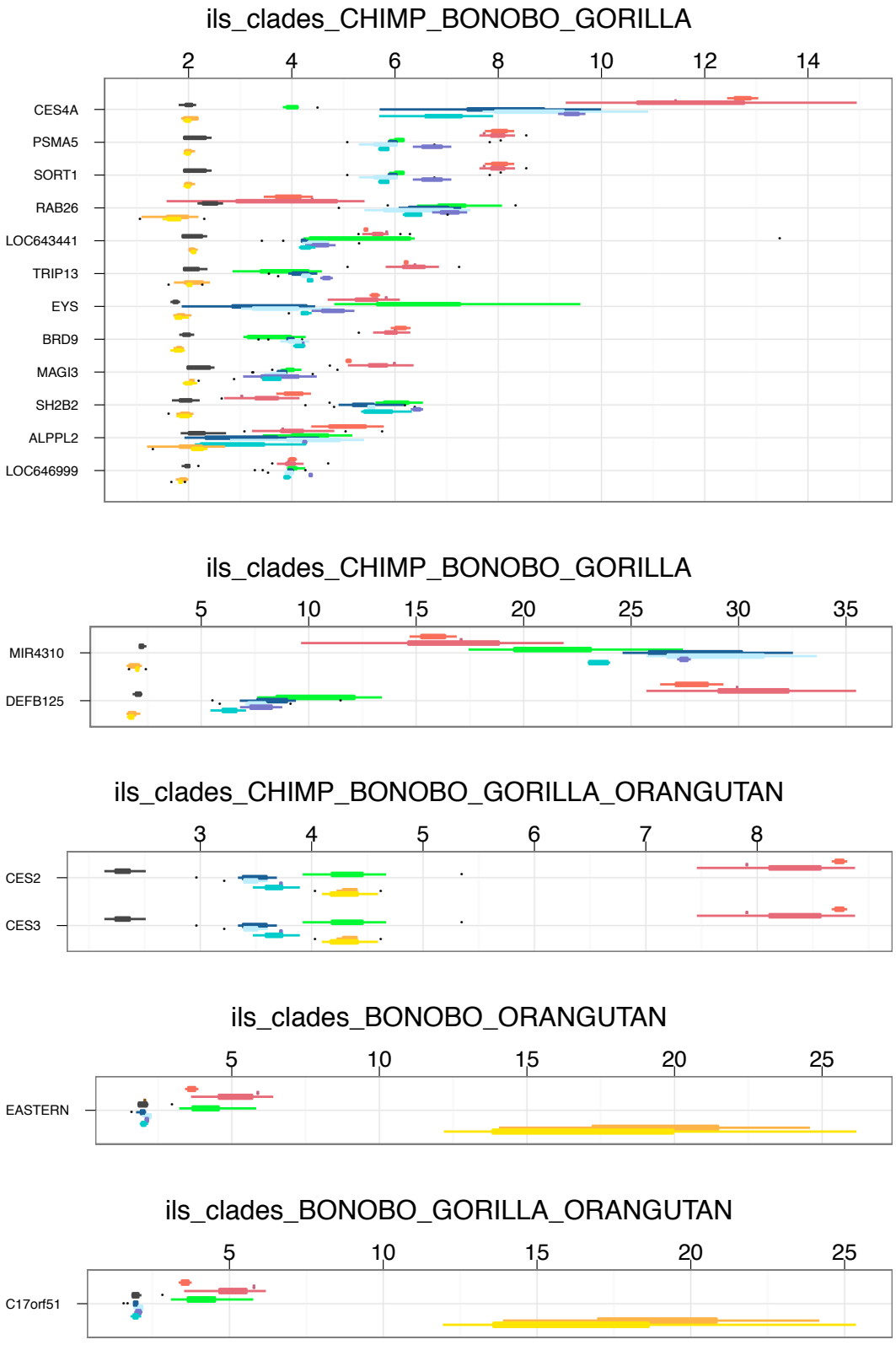

ils_clades_HUMAN_CHIMP_GORILLA_ORANGUTAN
5
10
15
OR11H2:OR11H1:OR11H12
CHRFAM7A
LOC340508
RNASE2


ils_clades_HUMAN_CHIMP_BONOBO_ORANGUTAN
2
4
6
8
10
12
14
DEFT1P,DEFT1P2:DEFT1P2,DEFT1P
ESPN:ESPNP
CYP4Z1
LOC650226
OR51A2
SERHL
ZNF98
GML


ils_clades_HUMAN_BONOBO
2
4
6
8
10
12
14
EASTERN


ils_clades_HUMAN_BONOBO_GORILLA
2
4
6
8
10
12
OR4F4:OR4F5:OR4F17


ils_clades_GORILLA_ORANGUTAN
0
2
4
6
8
MIR4733
RFPL4B
ALDH1B1
PGBD2
IGFBPL1
LOC643201
KRTAP5.9
ils_clades_CHIMP_BONOBO
2
4
6
8
10
12
14
EASTERN
ils_clades_CHIMP_BONOBO
5
10
15
20
25
30
35
EASTERN
ils_clades_CHIMP_BONOBO_ORANGUTAN
2
4
6
8
10
12
14
16
EASTERN
FOPNL
DEFB134
SPTLC1
LOC389043


ils_clades_CHIMP_BONOBO_GORILLA
2
4
6
8
10
12
14
CES4A
PSMA5
SORT1
RAB26
LOC643441
TRIP13
EYS
BRD9
MAGI3
SH2B2
ALPPL2
LOC646999
ils_clades_CHIMP_BONOBO_GORILLA
5
10
15
20
25
30
35
MIR4310
DEFB125
ils_clades_CHIMP_BONOBO_GORILLA_ORANGUTAN
3
4
5
6
7
8
CES2
CES3
ils_clades_BONOBO_ORANGUTAN
5
10
15
20
25
EASTERN
ils_clades_BONOBO_GORILLA_ORANGUTAN
5
10
15
20
25
C17orf51

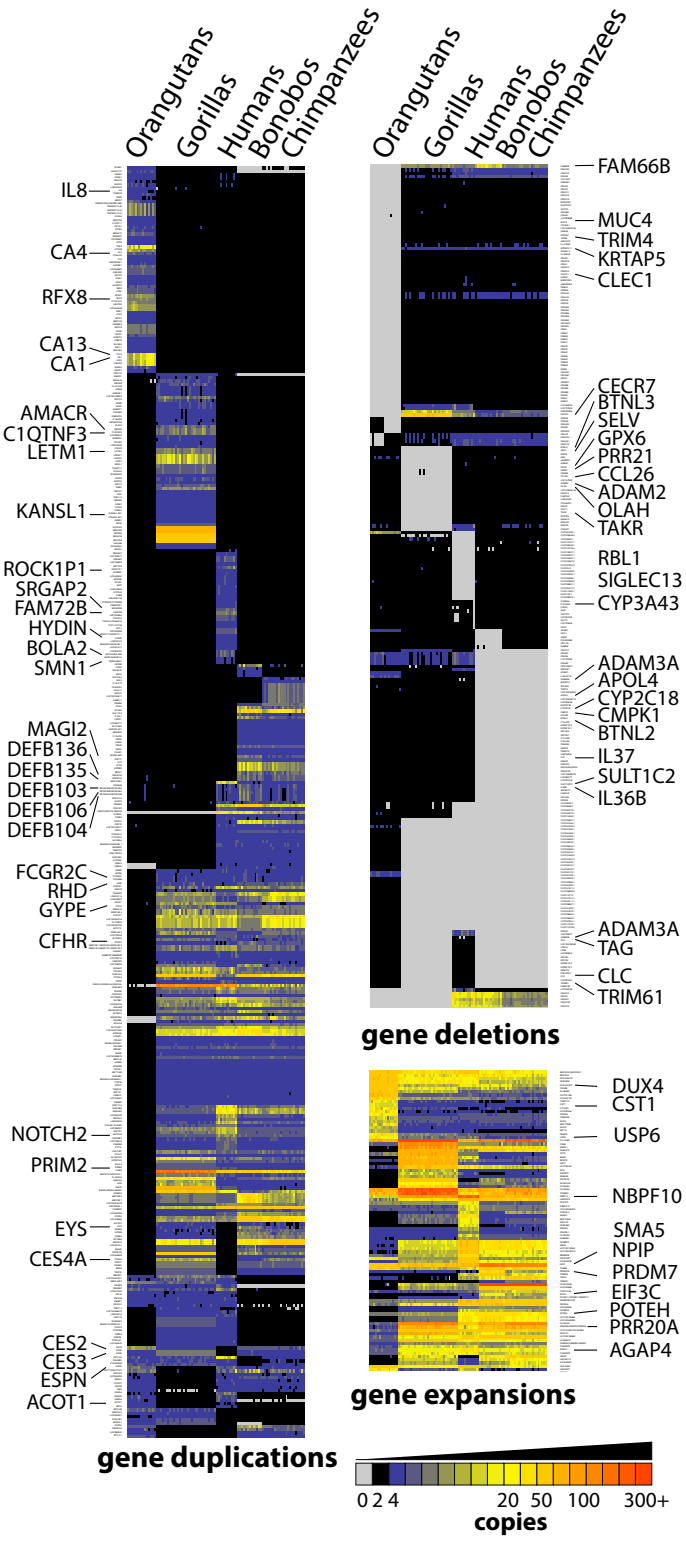


Figure 7.3: Heatmap representation of all lineage-specific gene duplications, deletions, and expansions identified throughout the great ape lineage.

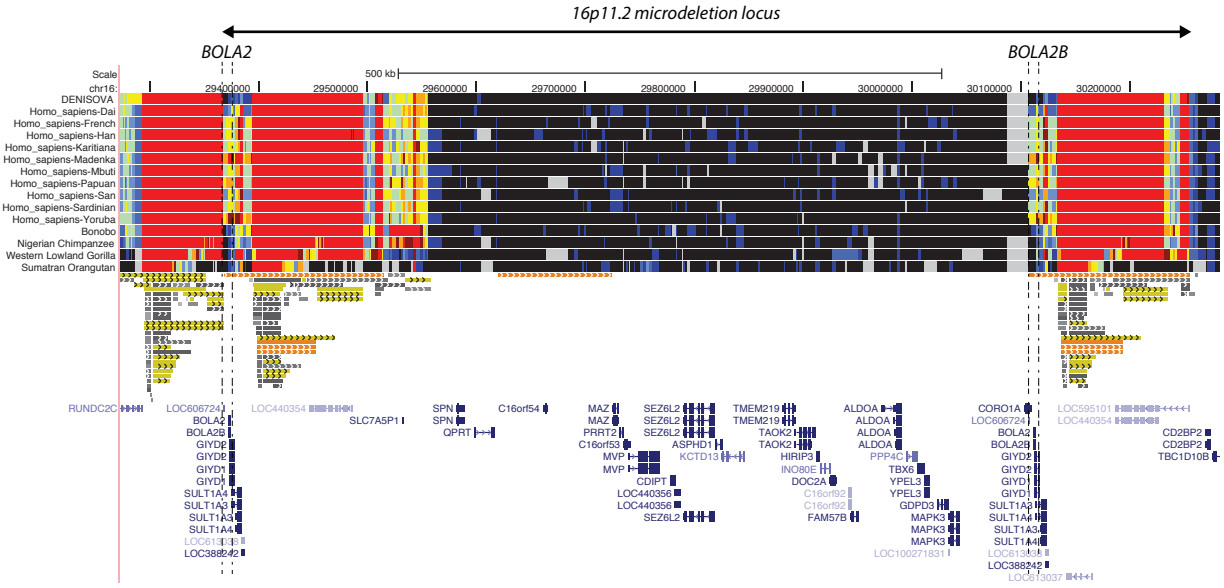


Figure 7.4: Heatmap of the 16p11.2 locus. The *BOLA2* and *BOLA2B* gene are highlighted in addition to the approximate breakpoints of the 16p11.2 micro-duplication/micro-deletion. *BOLA2* exhibits the ancestral copy number state.

Section 8: Distribution of duplication and deletion events

We assessed the position of lineage-specific gene duplications in relation to each other along the genome and found them to be significantly non-uniformly distributed (**Figures 8.1, 8.2 a,b,c,d,e**). Duplicated genes instead clustered into distinct cores of activity.

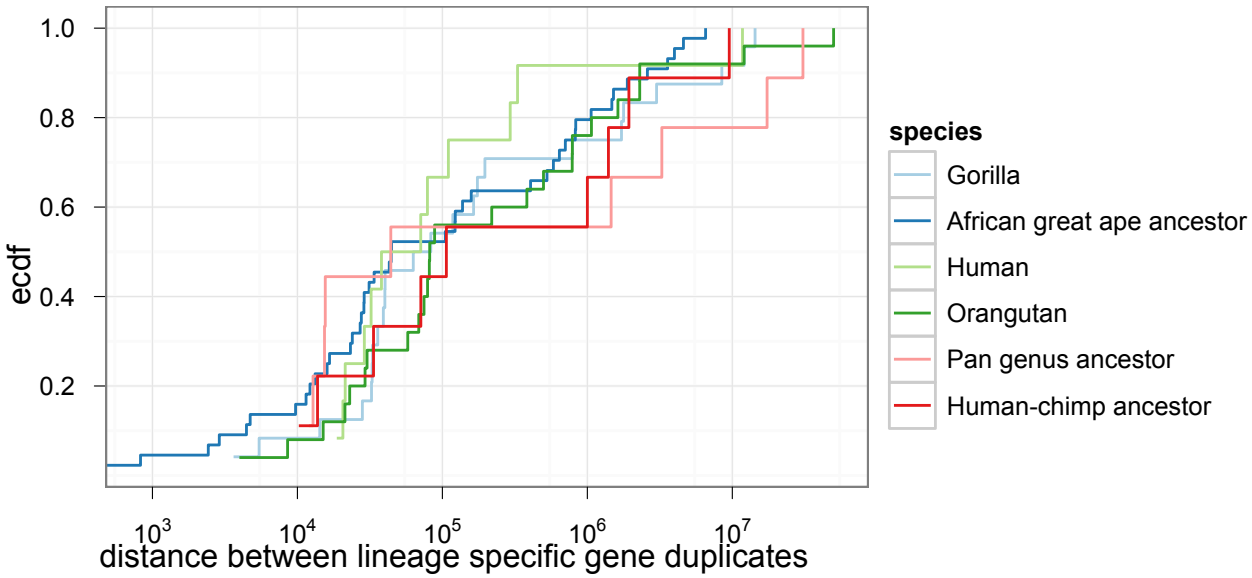


Figure 8.1: Cumulative histogram of the distance between lineage-specific gene duplicates. More than 50% of all gene duplicates lie within 100 kbp of each other.

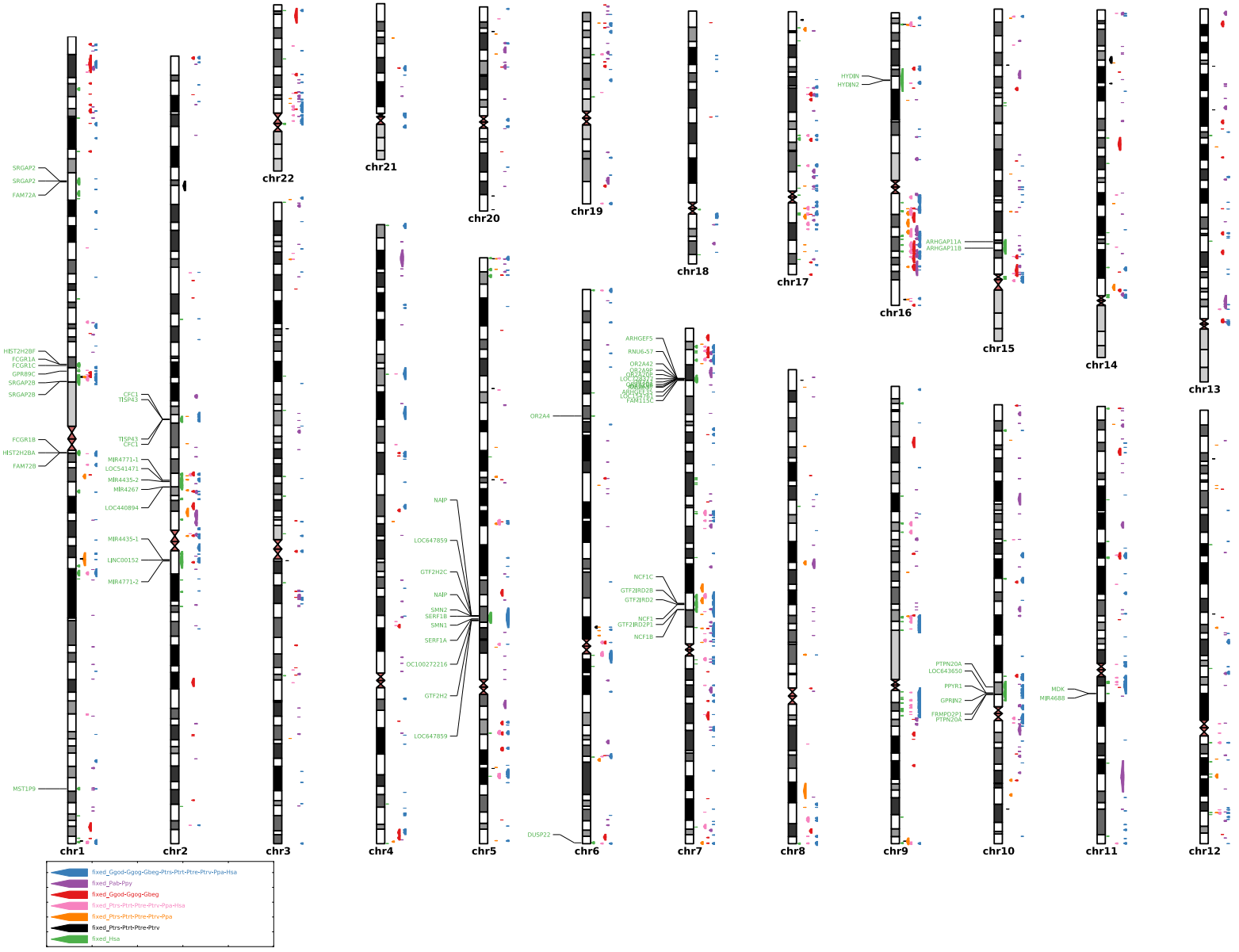

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
SRGAP2
SRGAP2
FAM72A
HIST2H2BF
FCGR1A
FCGR1C
GPR89C
SRGAP2B
SRGAP2B
FCGR1B
HIST2H2BA
FAM72B
MST1P9
CFC1
TISP43
TISP43
CFC1
MIR4771-1
LOC541471
MIR4435-2
MIR4267
LOC440894
MIR4435-1
LINC00152
MIR4771-2
NAIP
LOC647859
GTF2H2C
NAIP
SMN2
SERF1B
SMN1
SERF1A
OC100272216
GTF2H2
LOC647859
OR2A4
DUSP22
ARHGEF5
RNU6-57
OR2A42
OR2A9P
NCF1C
GTF2IRD2B
GTF2IRD2
NCF1
GTF2IRD2P1
NCF1B
HYDIN
HYDIN2
ARHGAP11A
ARHGAP11B
PTPN20A
LOC643650
PPYR1
GPRIN2
FRMPD2P1
PTPN20A
MDK
MIR4688
fixed_Ggod-Ggog-Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Pab-Ppy
fixed_Ggod-Ggog-Gbeg
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa
fixed_Ptrs-Ptrt-Ptre-Ptrv
fixed_Hsa

Figure 8.2a: Human-specific gene duplications plotted on a chromosome ideogram show gene duplications clustered together. Lineage-specific duplicated segments are plotted to the right of each chromosome ideogram as arrows with their bases scaled by 10 for visibility.

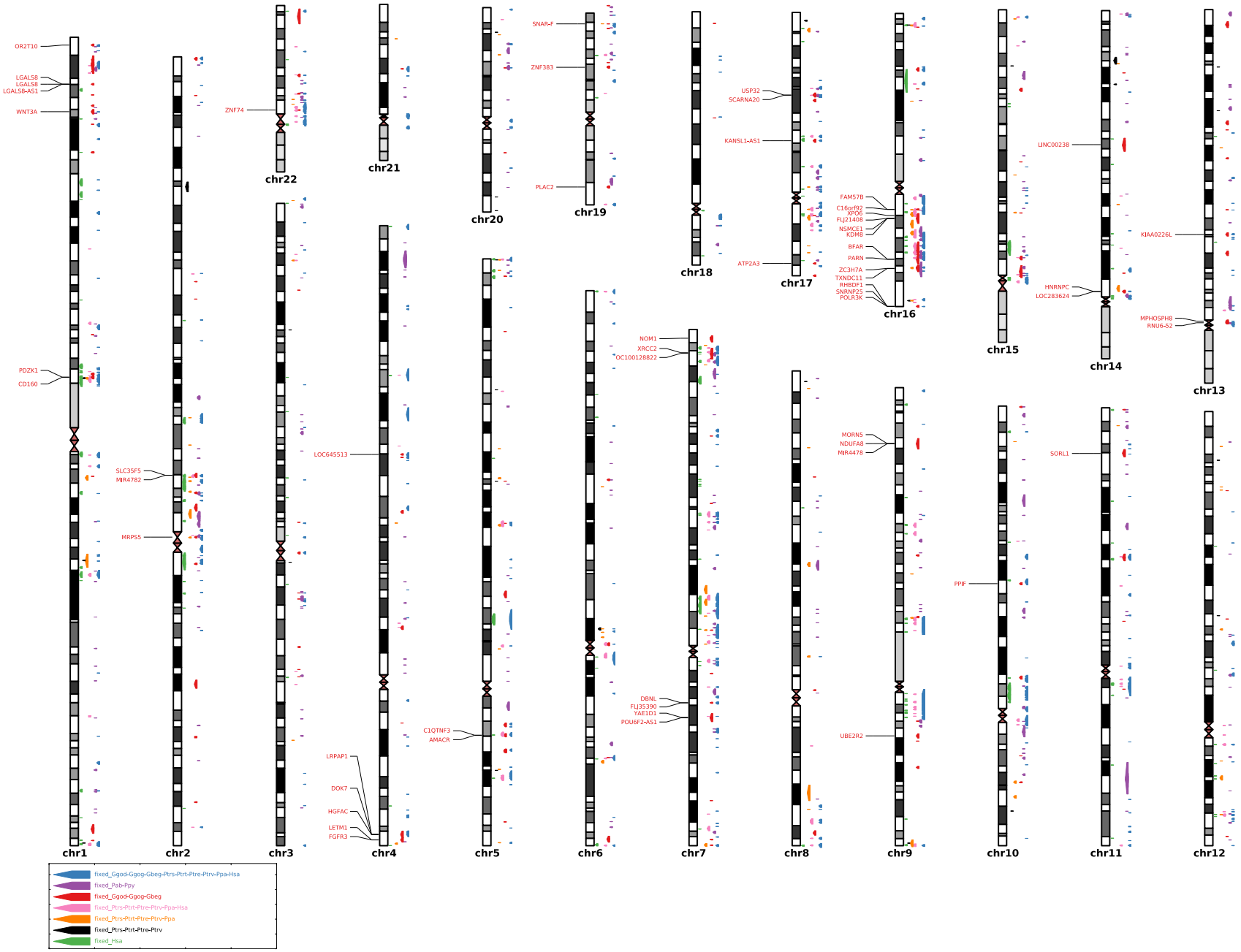

OR2T10
LGALS8
LGALS8
LGALS8-AS1
WNT3A
PDZK1
CD160
chr1
SLC35F5
MIR4782
MRPS5
chr2
ZNF74
chr22
chr3
chr21
LOC645513
LRPAP1
DOK7
HGFAC
LETM1
FGFR3
chr4
chr20
C1QTNF3
AMACR
chr5
SNAR-F
ZNF383
PLAC2
chr19
chr6
chr18
NOM1
XRCC2
OC100128822
DBNL
FLJ35390
YAE1D1
POU6F2-AS1
chr7
USP32
SCARNA20
KANSL1-AS1
ATP2A3
chr17
chr8
FAM57B
C16orf92
XPO6
FLJ21408
NSMCE1
KDM8
BFAR
PARN
ZC3H7A
TXNDC11
RHBDF1
SNRNP25
POLR3K
chr16
MORN5
NDUFA8
MIR4478
UBE2R2
chr9
chr15
PPIF
chr10
LINC00238
HNRNPC
LOC283624
chr14
SORL1
chr11
KIAA0226L
MPHOSPH8
RNU6-52
chr13
chr12
fixed_Ggod-Ggog-Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Pab-Ppy
fixed_Ggod-Ggog-Gbeg
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Ptrs-Ptrt-Ptre-Ppa
fixed_Ptrs-Ptrt-Ptre-Ptrv
fixed_Hsa

Figure 8.2b: Gorilla-specific gene duplications plotted on a chromosome ideogram are also clustered together.

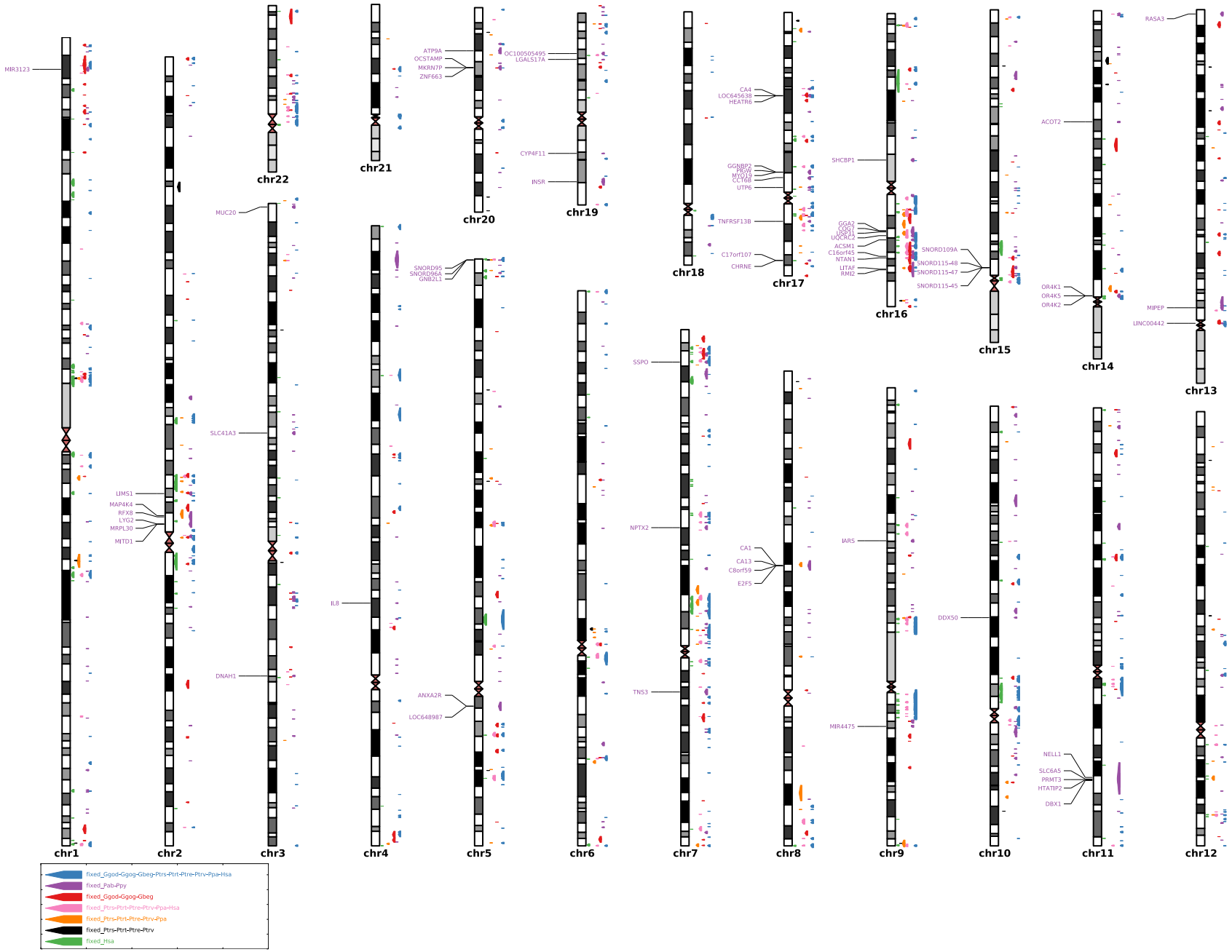

MIR3123
LIMS1
MAP4K4
RFX8
LYG2
MRPL30
MITD1
chr1
chr2
MUC20
SLC41A3
DNAH1
chr3
chr22
chr21
ATP9A
OCSTAMP
MKRN7P
ZNF663
chr20
OC100505495
LGALS17A
CYP4F11
INSR
chr19
IL8
chr4
SNORD95
SNORD96A
GNB2L1
ANXA2R
LOC648987
chr5
chr6
SSPO
NPTX2
TNS3
chr7
chr18
CA4
LOC645638
HEATR6
GGNBP2
PIGW
MYO1B
CCT6B
UTP6
TNFRSF13B
C17orf107
CHRNE
chr17
CA1
CA13
C8orf59
E2F5
chr8
SHCBP1
GGA2
CDC1
USP31
UQCRC1
ACSM1
C16orf45
NTAN1
LITAF
RMI2
chr16
IAR5
MIR4475
chr9
SNORD109A
SNORD115-48
SNORD115-47
SNORD115-49
chr15
DDX50
chr10
ACOT2
OR4K1
OR4K5
OR4K2
chr14
NELL1
SLC6A5
PRMT3
HTATIP2
DBX1
chr11
RASA3
MIPEP
LINC00442
chr13
chr12
fixed_Ggod-Ggog-Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Pab-Ppy
fixed_Ggod-Ggog-Gbeg
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa
fixed_Ptrs-Ptrt-Ptre-Ptrv
fixed_Hsa

Figure 8.2c: Orangutan-specific gene duplications plotted on a chromosome ideogram.

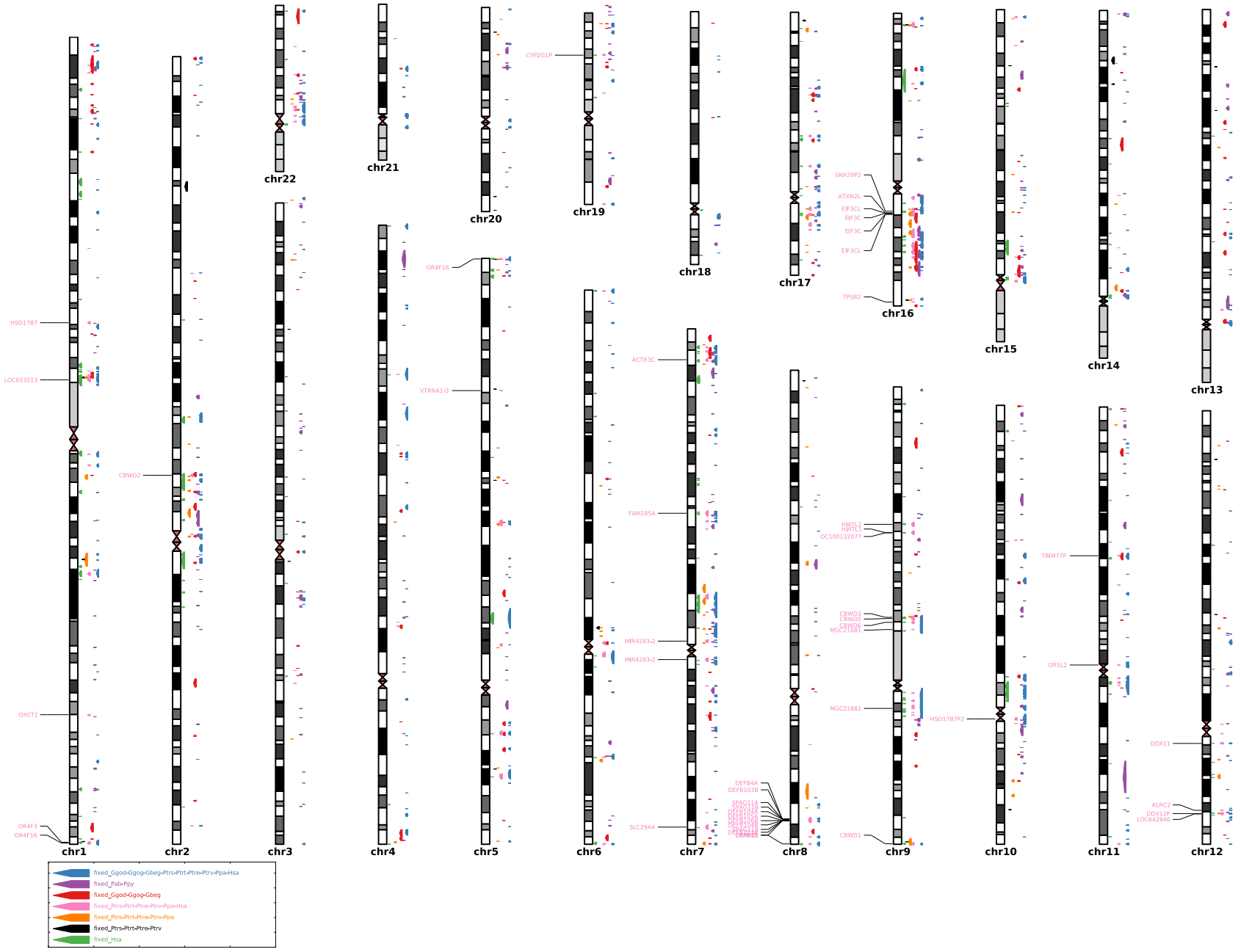

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
HSD17B7
LOC653513
CBWD2
OXCT2
OR4F3
OR4F16
OR4F16
VTRNA1-3
CYP2G1P
ACTR3C
FAM185A
MIR4283-2
MIR4283-2
SLC29A4
DEFB4A
DEFB103B
SNX29P2
ATXN2L
EIF3CL
EIF3C
EIF3C
EIF3CL
TPSB2
HIATL2
HIATL1
OC100132077
CBWD3
CBWD5
CBWD4
MGC21881
MGC21881
CBWD1
HSD17B7P2
TRIM77P
OR5L2
DDX11
KLRC2
DDX12P
LOC642846
fixed_Ggod-Ggog-Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Pab-Ppy
fixed_Ggod-Ggog-Gbeg
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Ptrs-Ptrt-Ptre-Ptrv-Ppa
fixed_Ptrs-Ptrt-Ptre-Ptrv
fixed_Hsa

Figure 8.2d: Human, chimpanzee, and bonobo shared gene duplications plotted on a chromosome ideogram.

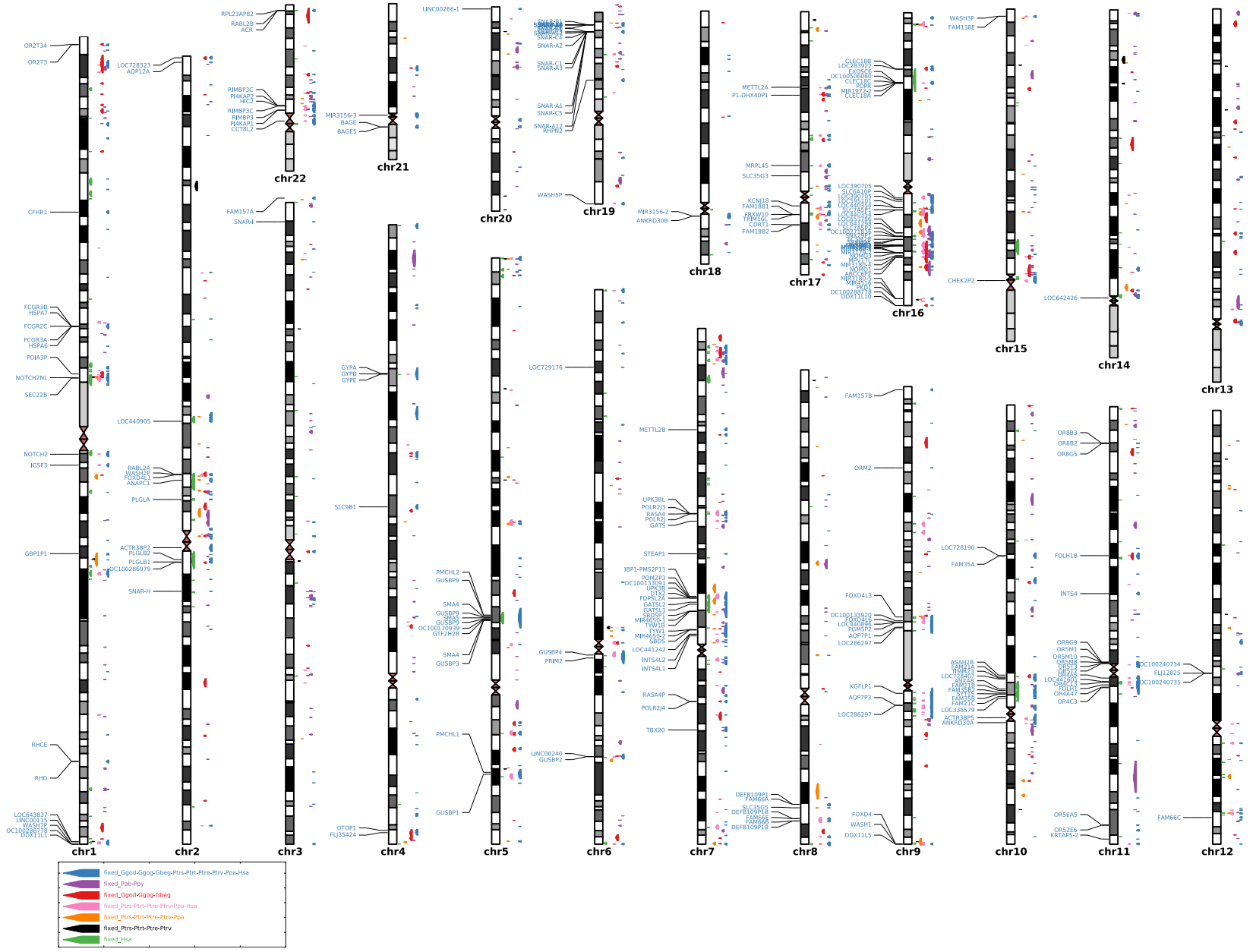


Figure 8.2e: African great ape-specific gene duplications plotted on a chromosome ideogram.

Distribution of SDs and deletions: We have previously observed that lineage-specific duplications tend to arise adjacent to ancestral duplications, a phenomenon we have termed *duplication shadowing*(Cheng et al. 2005; Marques-Bonet et al. 2009). We tested that association again here (using a permutation test method)(Marques-Bonet et al. 2009) and found that indeed lineage-specific duplications were highly enriched near to ancestral duplications (P<0.0002, **Table 8.1**). We were curious whether this phenomenon held true for deletions as well. Lineage-specific deletions did not tend to arise adjacent to ancestral deletions (P>0.2, **Table 8.1**), and additionally, no association with SDs was observed either (P>0.2, **Table 8.1**).

Table 8.1: Association of lineage-specific duplication and deletion events with each other. As we previously reported, we find that lineage-specific SDs tend to emerge adjacent to ancestral SDs. We also tested if deletions emerged adjacent to ancestral deletions or SDs and found no association. For loci lost along the human lineage, only deletions flanked by regions that could unambiguously be placed in HG18 were considered.

Comparison	test events	reference events	test within 5 kbp of reference	P-value
human duplications with human/chimp duplications	279	268	30	P<0.0002
chimp duplications with human/chimp duplications	115	268	12	P<0.0002
Bonobo duplications with human/chimp duplications	51	268	2	P=0.0214
human/chimp duplications with human/chimp/gorilla duplications	268	1172	125	P<0.0002
gorilla duplications with human/chimp/gorilla duplications	238	1172	28	P<0.0002
orangutan duplications with human/chimp/gorilla/orangutan ancestral duplications	549	7077	116	P<0.0002
gorilla deletions with gorilla duplications	469	238	1	P=0.8200
gorilla deletions with human/chimp/gorilla duplications	469	1172	5	P=0.8900
gorilla deletions with human/chimp duplications	459	268	2	P=0.5500
human deletions with human/chimp/ duplications	331	268	1	P=0.7600
human/chimp deletions with human/chimp/gorilla duplications	65	1172	0	P=1.0000
human/chimp/gorilla deletions with human/chimp/gorilla/orangutan duplications	372	7077	47	P=1.0000
human deletions with human/chimp ancestral deletions	331	74	7	P=0.2300
gorilla deletions with human/chimp/gorilla ancestral deletions	469	422	39	P=1.0000
human/chimp deletions with human/chimp/gorilla ancestral deletions	65	422	10	P=0.6500

Section 9: Rates of segmental duplication and deletion and underlying genes

As we have identified SNPs and substitutions in the same individuals assessed for structural variants, we can compute precise estimates of the rates of

duplication and deletion as a function of the number of substitutions along a branch. Additionally, the divergence times of each of the species assessed can be computed precisely from genetic divergence estimates. We have previously described a likelihood framework for testing the significance of changes in duplication rate observed throughout a phylogeny(Marques-Bonet et al. 2009). In this likelihood framework, the null model assumes a constant rate among all branches while in the test model the rates vary among different branches. All tests were performed in units of duplications per million years, with divergence times estimated from exact substitution rates, and in units of duplications per substitution.

Rates of SD: We have previously shown an acceleration in the rate of duplication in the African great ape ancestor(Marques-Bonet et al. 2009) and additionally an acceleration of the duplication rate in the gorilla lineage(Ventura et al. 2011). We confirmed these two findings by assessing the rate of ancient duplication among all branches in the tree. The four subspecies of chimpanzee were considered as a single population. Three models were tested against a null model of a constant duplication rate of duplication along all branches:

- **An accelerated rate in the African great ape ancestor,**
- **An accelerated rate in the African great ape ancestor, the gorilla ancestor, and the human-chimpanzee ancestor, and**
- **Differing rates of duplication among all branches of the tree.**

Each of these three models is highly significant compared the null model (**Table 9.1**); however, the tree is best explained by a model in which there is an accelerated rate of duplication in the African great ape and gorilla ancestors in addition to the ancestor of human and chimpanzee. This supports a model in which a burst of duplication activity occurred in the African great ape ancestor which subsequently rapidly declined in the ancestor of human and chimpanzee and declined more slowly in the gorilla lineage.

Table 9.1: Duplication rates along the great ape lineage.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in African great ape against all other branches	Mb/My	λ=3.030	λ1 = 2.127 λ2 = 7.501	1	6.73E-13
	bp/sub	λ=1.197	λ1 = 0.841 λ2 = 2.961	1	7.00E-13
	sites/MY	λ=30.006	λ1 = 17.659 λ2 = 91.155	1	<1.138e-202
acceleration in African great ape and gorilla and human-chimp linages against all branches	Mb/My	λ=3.030	λ1 = 1.496 λ2 = 6.173	1	1.21E-17
	bp/sub	λ=1.197	λ1 = 0.592 λ2 = 2.432	1	1.53E-17

	sites/ MY	λ=30.006	λ1 = 14.943 λ2 = 60.866	1	<7.529e-157
different rates among all branches	Mb/ My	λ=3.030	λ1 = 0.266 λ2 = 7.501 λ3 = 4.984 λ4 = 3.809 λ5 = 1.649 λ6 = 1.119 λ7 = 0.547 λ8 = 1.210 λ9 = 0.470 λ10 = 0.528 λ11 = 2.147 λ12 = 0.341 λ13 = 2.104	12	3.00E-14
	bp/sub	λ=1.197	λ1 = 0.108 λ2 = 2.961 λ3 = 1.965 λ4 = 1.476 λ5 = 0.647 λ6 = 0.434 λ7 = 0.216 λ8 = 0.484 λ9 = 0.178 λ10 = 0.215 λ11 = 0.849 λ12 = 0.132 λ13 = 0.863	12	3.48E-14

Gene duplication rates: We next tested if the increase in duplication rate in African great ape ancestor, gorilla and human-chimpanzee ancestral lineages had led to a significant increase in the rate of gene duplication (**Table 9.2**). Indeed, the rate of gene duplication in these lineages was ~2.8-fold higher than throughout the rest of the tree (P=1.66e-20). We also observed a ~1.5-fold excess in the rate of gene duplication in the human-chimpanzee ancestral branch compared to the African great ape ancestral branch and the gorilla branch, which was weakly significant (P=0.0107).

Table 9.2: Rates of lineage-specific gene duplication.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in the rate of gene duplication in the African great ape, gorilla and human-chimpanzee lineages against all branches	genes/MY	λ=6.620	λ1 = 4.167 λ2 = 11.646	1	1.66E-20
	genes/sub	λ=2.616	λ1 = 1.649 λ2 = 4.588	1	2.42E-20
acceleration in the rate of gene duplication in human-chimp lineage compared to the gorilla and African great ape lineages	genes/MY	λ=11.646	λ1 = 10.933 λ2 = 19.286	1	0.0107
	genes/sub	λ=4.588	λ1 = 4.313 λ2 = 7.474	1	0.01323

Rates of lineage-specific deletion and gene loss: We observed a ~2-fold increase in the rate of deletion along the chimpanzee-bonobo lineage. We found this increase to be highly statistically significant (**Table 9.3,** $P=4.79X10^{-9}$). In addition, the associated rate of gene loss along this branch is also highly statistically significant ($P=4.4X10^{-8}$). Amongst the other branches, the rate of deletion and the associated rate of gene loss appear to be more clocklike. We cannot reject the null model (of clocklike) in either of these cases (P=0.03 and P=0.6). Analyses were also performed just for deletions >5kbp, (**Table 9.4**), showing a strong trend

Table 9.3: Rates of lineage-specific deletion and gene loss.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in the rate of deletion along the chimpanzee-bonobo ancestral branch compared to all other branches	ckbp/MY	λ=9.072	λ1 = 8.189	1	4.79E-09
			λ2 = 18.054		
	bp/kb-sub	λ=3.587	λ1 = 3.240	1	6.63E-09
			λ2 = 7.086		
	sites/MY	λ=114.842	λ1 = 110.829	1	7.13E-13
			λ2 = 151.136		
	genes/MY	λ=6.034	λ1 = 5.318	1	4.40E-08
			λ2 = 12.955		
	genes/subs	λ=2.385	λ1 = 2.103	1	5.57E-08
			λ2 = 5.085		
differing rate of deletion in all non-chimpanzee-bonobo ancestor branches (reject clocklike)	kbp/MY	λ=8.918	λ1 = 7.626	5	0.03123
			λ2 = 8.919		
			λ3 = 8.801		
			λ4 = 10.392		
			λ5 = 7.172		
			λ6 = 11.668		
	bp/kb-sub	λ=3.526	λ1 = 3.012	5	0.03557
			λ2 = 3.660		
			λ3 = 3.411		
			λ4 = 4.102		
			λ5 = 2.868		
			λ6 = 4.600		
	genes/MY	λ=5.660	λ1 = 6.207	5	0.6271
			λ2 = 10.000		
			λ3 = 4.286		
			λ4 = 4.762		
			λ5 = 6.167		
			λ6 = 5.303		
	sites/MY	λ=113.767	λ1 = 76.621	4	1.82E-72
			λ2 = 152.143		
			λ3 = 141.071		
			λ4 = 153.167		
			λ5 = 116.667		
excluding orangutan	sites/MY	λ=137.812	λ1 = 116.667	3	1.12E-07
			λ2 = 152.143		
			λ3 = 141.071		
			λ4 = 153.167		

Table 9.4: Rates of lineage-specific deletion for sites >5kb.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in the rate of deletion along the chimpanzee-bonobo ancestral branch compared to all other branches	ckbp/MY	λ=7.778	λ1 = 6.836	1	2.08E-09
			λ2 = 16.304		
	bp/10-sub	λ=3.072	λ1 = 2.702	1	2.72E-09
			λ2 = 6.399		
	sites/MY	λ=50.837	λ1 = 45.704	1	1.85E-38
			λ2 = 97.273		
	genes/MY	λ=6.282	λ1 = 5.527	1	1.60E-07
			λ2 = 12.955		
	genes/subs	λ=2.482	λ1 = 2.186	1	2.00E-07
			λ2 = 5.085		
differing rate of deletion in all non-chimpanzee-bonobo ancestor branches (reject clock like)	kbp/MY	λ=6.812	λ1 = 6.365	4	0.0003297
			λ2 = 5.761		
			λ3 = 7.478		
			λ4 = 3.655		
			λ5 = 10.100		
	bp/10-sub	λ=2.693	λ1 = 2.514	4	0.000414
			λ2 = 2.233		
			λ3 = 2.952		
			λ4 = 1.462		
			λ5 = 3.981		
	genes/MY	λ=5.637	λ1 = 6.207	4	0.5693
			λ2 = 4.286		
			λ3 = 4.762		
			λ4 = 6.167		
			λ5 = 5.303		
	sites/MY	λ=46.152	λ1 = 38.759	4	3.35E-20

			λ2 = 47.143		
			λ3 = 46.548		
			λ4 = 37.333		
			λ5 = 69.697		
excluding orangutan	sites/MY	λ=50.938	λ1 = 69.697	3	8.61E-15
			λ2 = 47.143		
			λ3 = 46.548		
			λ4 = 37.333		

Section 10: Copy number variation and duplication diversity

Copy number variants and copy number polymorphic loci: Copy number polymorphic loci and copy number variants were identified using a digital array comparative genomic hybridization (dCGH) strategy(Sudmant et al. 2010), in combination with the above scale-space filtering-based segmentation approach. Briefly, for a particular species, one individual is assigned to be a *reference* sample. A comparative copy number signal between *test* individuals from each species and their respective references is then generated from the log2 ratio of the 500 bp tiled window copy number estimates between the two samples. These log2 ratios are then segmented into blocks of respective 'gain' and 'loss' using the above segmentation procedure. CNPs identified in high GC loci were then filtered out if the GC content of the CNP was >0.55 or if the mean copy number of individuals with known GC biases differed by >0.5 from unbiased individuals. 18 individuals were excluded from this analysis as GC bias prevented them from being compared to nonbiased samples. CNPs encompassing fewer than four 500 bp windows of unmasked sequence were discarded to enrich for true positives effectively imposing a 2 kbp size constraint on discovered events. To maximize sensitivity, for each species every individual in turn was selected as the *reference* genome and compared against all other genomes. A total of 2062 individual dCGH comparisons were performed and a total of 6406 events discovered (**Table 10.1**). Though the minimum size of CNPs discovered was 2 kbp, as a minimum of four 500 bp windows of unmasked sequence was required to discover a CNP, the median size was much larger (8061 bp; **Figure 10.1**).

Table 10.1: CNPs discovered among 80 great apes.

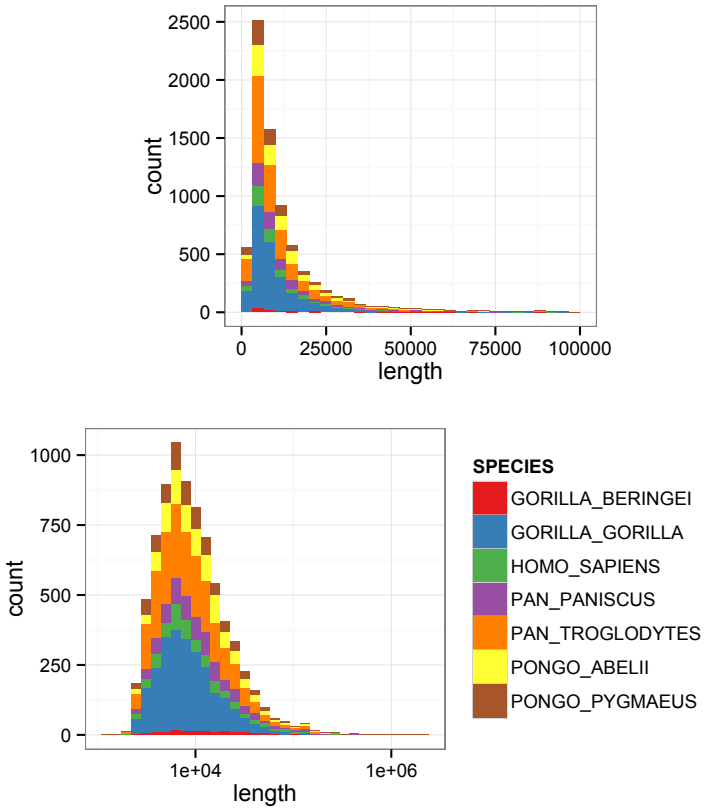


Figure 10.1: Distribution of CNP sizes shown in standard and log-scale. The median CNP size was 8061 bp.

	private		segregating			total		θ	
species	deletion	duplication	duplication	deletion	duplication /deletion	private	segregating	CNVs (per genome)	bp (pe
Sumatran_orangutan	430	71	226	114	65	501	405	122.05	0.0029
Bornean_orangutan	296	61	149	108	53	357	310	97.48	0.002
Human	203	182	113	38	29	385	180	49.38	0.0013
Bonobo	340	68	147	89	29	408	265	77.04	0.0012
Eastern_chimpanzee	466	135	107	78	21	601	206	68.21	0.002
Western_chimpanzee	242	61	49	37	20	303	106	37.47	0.001
Nigerian-Cameroon_chimpanzee	198	43	132	35	30	241	197	55.53	0.0018
Central_chimpanzee	581	232	-	-	-	813	-	-	0.0024
Western_gorilla	402	130	287	92	39	532	418	98.27	0.002
Eastern_gorilla	114	60	-	-	-	174	0	-	0.0009

We next assessed the number, and size distribution of discovered CNVs on a per-individual basis, to determine if any individuals in particular showed aberrantly more or larger CNVs (**Table 10.2**). The results are summarized in **Figure 10.2** with all individuals of a particular species demonstrating a similar number of CNVs on average, with four potential outliers showing excess numbers of deletions. The first outlier, Nyango, is the only Cross River Gorilla sample assessed in our study. We note that this individual shows both signals of recent inbreeding and exceedingly low diversity, (Prado and Sudmant *et al – in* press) in addition to a unique demographic history and reduced N_e. Though interesting, it is hard to draw conclusions from this single individuals excess of deletions. The other three individuals are a

central chimpanzee, of which only two individuals were assessed, a Bornean orangutan and a Sumatran orangutan. The two orangutans are not outliers among the orangutan samples with respect to sequence coverage, nor is the central chimpanzee. We thus find no apparent reason to purge these individuals from our analyses, however, we note that all reported results on population diversity are robust to the removal of these individuals, and indeed, none of the significant signals reported is related to any of these individuals. Finally, we assessed the distribution of CNV lengths among the 80 great apes analyzed for CNV diversity in this study (**Figure 10.3**). We noted that for almost all individuals the length distributions of deletions and duplications were very similar as one would expect for a random sampling of CNVs. Additionally, size distributions were consistent among individuals of the same species and largely between species as well. Interestingly, the human individuals consistently exhibited a slight excess of longer deletions, perhaps as a result of the human reference genome being biased towards a particular set of CNV alleles.

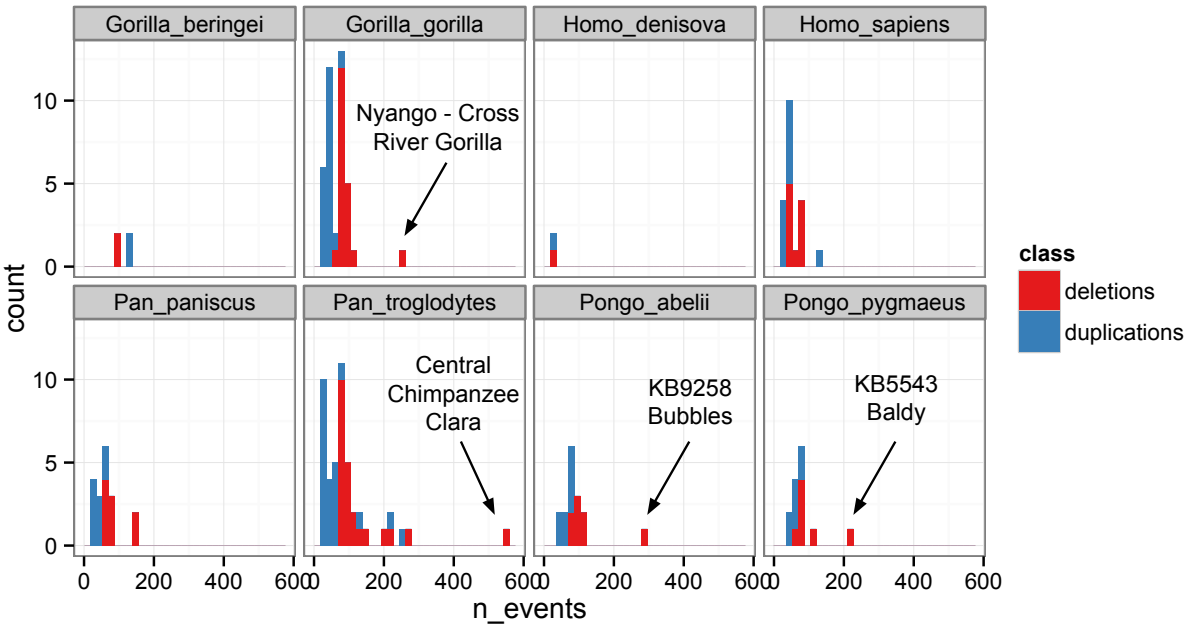


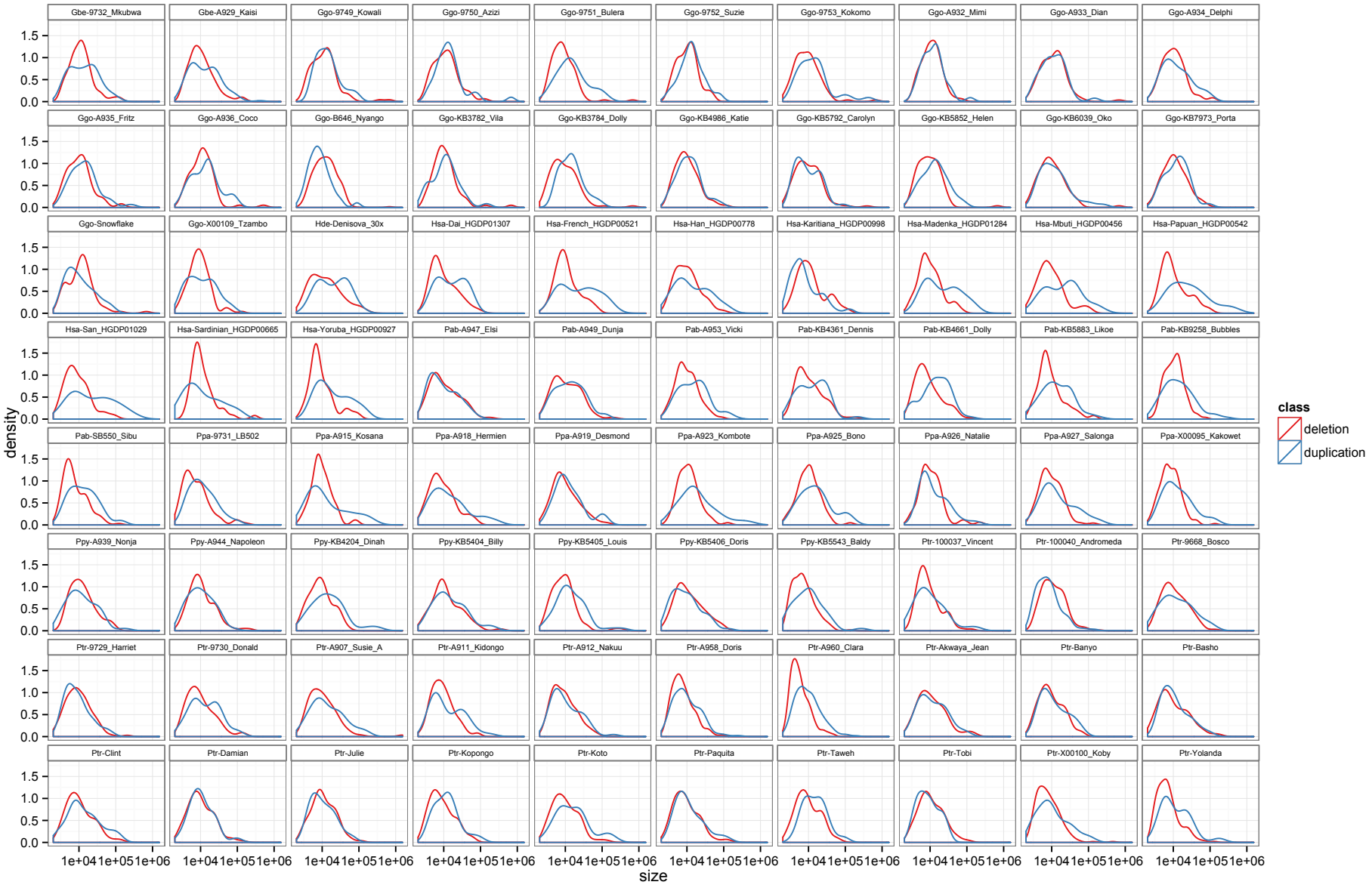
Figure 10.2: Summary of the number of duplications and deletions discovered per individual and grouped by species.

Table 10.2: The number of CNV duplications and deletions discovered per individual.

indiv	species	duplications	deletions
Pongo_abelii-KB9258_Bubbles	Pongo_abelii	77	293
Pongo_abelii-A953_Vicki	Pongo_abelii	87	90
Pongo_abelii-A949_Dunja	Pongo_abelii	70	70
Pongo_abelii-KB4661_Dolly	Pongo_abelii	66	107
Pongo_abelii-KB5883_Likoe	Pongo_abelii	70	97
Pongo_abelii-A947_Elsi	Pongo_abelii	67	70
Pongo_abelii-SB550_Sibu	Pongo_abelii	50	109
Pongo_abelii-KB4361_Dennis	Pongo_abelii	45	97
Pongo_pygmaeus-KB5543_Baldy	Pongo_pygmaeus	83	212

Pongo_pygmaeus-KB4204_Dinah	Pongo_pygmaeus	44	73
Pongo_pygmaeus-KB5406_Doris	Pongo_pygmaeus	55	107
Pongo_pygmaeus-A939_Nonja	Pongo_pygmaeus	81	80
Pongo_pygmaeus-KB5404_Billy	Pongo_pygmaeus	48	69
Pongo_pygmaeus-KB5405_Louis	Pongo_pygmaeus	64	87
Pongo_pygmaeus-A944_Napoleon	Pongo_pygmaeus	61	84
Homo_denisova-Denisova_30x	Homo_denisova	32	28
Homo_sapiens-Han_HGDP00778	Homo_sapiens	36	45
Homo_sapiens-Dai_HGDP01307	Homo_sapiens	18	50
Homo_sapiens-Yoruba_HGDP00927	Homo_sapiens	35	73
Homo_sapiens-Mbuti_HGDP00456	Homo_sapiens	35	76
Homo_sapiens-San_HGDP01029	Homo_sapiens	36	74
Homo_sapiens-Karitiana_HGDP00998	Homo_sapiens	129	42
Homo_sapiens-Sardinian_HGDP00665	Homo_sapiens	37	48
Homo_sapiens-French_HGDP00521	Homo_sapiens	23	55
Homo_sapiens-Papuan_HGDP00542	Homo_sapiens	20	47
Homo_sapiens-Madenka_HGDP01284	Homo_sapiens	29	86
Pan_paniscus-A925_Bono	Pan_paniscus	43	78
Pan_paniscus-A923_Kombote	Pan_paniscus	31	73
Pan_paniscus-9731_LB502	Pan_paniscus	59	140
Pan_paniscus-A918_Hermien	Pan_paniscus	38	66
Pan_paniscus-A919_Desmond	Pan_paniscus	30	68
Pan_paniscus-A927_Salonga	Pan_paniscus	40	83
Pan_paniscus-X00095_Kakowet	Pan_paniscus	62	153
Pan_paniscus-A926_Natalie	Pan_paniscus	28	58
Pan_paniscus-A915_Kosana	Pan_paniscus	26	67
Pan_troglodytes_ellioti-Tobi	Pan_troglodytes	21	96
Pan_troglodytes_schweinfurthii-A911_Kidongo	Pan_troglodytes	53	147
Pan_troglodytes_verus-9668_Bosco	Pan_troglodytes	46	98
Pan_troglodytes_verus-9730_Donald	Pan_troglodytes	53	103
Pan_troglodytes_schweinfurthii-Yolanda	Pan_troglodytes	48	223
Pan_troglodytes_ellioti-Julie	Pan_troglodytes	26	85
Pan_troglodytes_ellioti-Taweh	Pan_troglodytes	23	81
Pan_troglodytes_ellioti-Basho	Pan_troglodytes	28	84
Pan_troglodytes_schweinfurthii-A912_Nakuu	Pan_troglodytes	53	105
Pan_troglodytes_ellioti-Banyo	Pan_troglodytes	30	76
Pan_troglodytes_schweinfurthii-9729_Harriet	Pan_troglodytes	77	116
Pan_troglodytes_ellioti-Paquita	Pan_troglodytes	23	84
Pan_troglodytes_schweinfurthii-100040_Andromeda	Pan_troglodytes	127	262
Pan_troglodytes_schweinfurthii-100037_Vincent	Pan_troglodytes	55	98
Pan_troglodytes_troglodytes-A960_Clara	Pan_troglodytes	214	542
Pan_troglodytes_ellioti-Koto	Pan_troglodytes	25	77
Pan_troglodytes_ellioti-Kopongo	Pan_troglodytes	21	90
Pan_troglodytes_verus-A907_Susie_A	Pan_troglodytes	53	84
Pan_troglodytes_verus-X00100_Koby	Pan_troglodytes	50	198
Pan_troglodytes_ellioti-Akwaya_Jean	Pan_troglodytes	29	79
Pan_troglodytes_ellioti-Damian	Pan_troglodytes	26	86
Pan_troglodytes_troglodytes-A958_Doris	Pan_troglodytes	247	126
Pan_troglodytes_verus-Clint	Pan_troglodytes	43	86
Gorilla_gorilla_gorilla-9753_Kokomo	Gorilla_gorilla	26	92
Gorilla_gorilla_gorilla-A936_Coco	Gorilla_gorilla	35	68
Gorilla_gorilla_gorilla-A935_Fritz	Gorilla_gorilla	38	84
Gorilla_gorilla_gorilla-KB5852_Helen	Gorilla_gorilla	39	87
Gorilla_gorilla_gorilla-9751_Bulera	Gorilla_gorilla	33	93
Gorilla_gorilla_gorilla-A932_Mimi	Gorilla_gorilla	38	91
Gorilla_gorilla_gorilla-Snowflake	Gorilla_gorilla	57	86
Gorilla_gorilla_gorilla-KB7973_Porta	Gorilla_gorilla	33	86
Gorilla_gorilla_gorilla-KB5792_Carolyn	Gorilla_gorilla	39	87
Gorilla_gorilla_gorilla-KB4986_Katie	Gorilla_gorilla	36	75
Gorilla_beringei_graueri-A929_Kaisi	Gorilla_beringei	130	101
Gorilla_gorilla_gorilla-9752_Suzie	Gorilla_gorilla	35	94
Gorilla_gorilla_gorilla-9749_Kowali	Gorilla_gorilla	20	87
Gorilla_gorilla_gorilla-X00109_Tzambo	Gorilla_gorilla	41	108
Gorilla_beringei_graueri-9732_Mkubwa	Gorilla_beringei	127	90
Gorilla_gorilla_dielhi-B646_Nyango	Gorilla_gorilla	77	252
Gorilla_gorilla_gorilla-KB3782_Vila	Gorilla_gorilla	35	79
Gorilla_gorilla_gorilla-9750_Azizi	Gorilla_gorilla	37	101
Gorilla_gorilla_gorilla-A934_Delphi	Gorilla_gorilla	27	80
Gorilla_gorilla_gorilla-KB6039_Oko	Gorilla_gorilla	40	84
Gorilla_gorilla_gorilla-KB3784_Dolly	Gorilla_gorilla	25	87

Gorilla_gorilla_gorilla-A933_Dian	Gorilla_gorilla	36	81


Gbe-9732_Mkubwa
Gbe-A929_Kaisi
Ggo-9749_Kowali
Ggo-9750_Azizi
Ggo-9751_Bulera
Ggo-9752_Suzie
Ggo-9753_Kokomo
Ggo-A932_Mimi
Ggo-A933_Dian
Ggo-A934_Delphi
Ggo-A935_Fritz
Ggo-A936_Coco
Ggo-B646_Nyango
Ggo-KB3782_Vila
Ggo-KB3784_Dolly
Ggo-KB4986_Katie
Ggo-KB5792_Carolyn
Ggo-KB5852_Helen
Ggo-KB6039_Oko
Ggo-KB7973_Porta
Ggo-Snowflake
Ggo-X00109_Tzambo
Hde-Denisova_30x
Hsa-Dai_HGDP01307
Hsa-French_HGDP00521
Hsa-Han_HGDP00778
Hsa-Karitiana_HGDP00998
Hsa-Madenka_HGDP01284
Hsa-Mbuti_HGDP00456
Hsa-Papuan_HGDP00542
Hsa-San_HGDP01029
Hsa-Sardinian_HGDP00665
Hsa-Yoruba_HGDP00927
Pab-A947_Elsi
Pab-A949_Dunja
Pab-A953_Vicki
Pab-KB4361_Dennis
Pab-KB4661_Dolly
Pab-KB5883_Likoe
Pab-KB9258_Bubbles
Pab-SB550_Sibu
Ppa-9731_LB502
Ppa-A915_Kosana
Ppa-A918_Hermien
Ppa-A919_Desmond
Ppa-A923_Kombote
Ppa-A925_Bono
Ppa-A926_Natalie
Ppa-A927_Salonga
Ppa-X00095_Kakowet
Ppy-A939_Nonja
Ppy-A944_Napoleon
Ppy-KB4204_Dinah
Ppy-KB5404_Billy
Ppy-KB5405_Louis
Ppy-KB5406_Doris
Ppy-KB5543_Baldy
Ptr-100037_Vincent
Ptr-100040_Andromeda
Ptr-9668_Bosco
Ptr-9729_Harriet
Ptr-9730_Donald
Ptr-A907_Susie_A
Ptr-A911_Kidongo
Ptr-A912_Nakuu
Ptr-A958_Doris
Ptr-A960_Clara
Ptr-Akwaya_Jean
Ptr-Banyo
Ptr-Basho
Ptr-Clint
Ptr-Damian
Ptr-Julie
Ptr-Kopongo
Ptr-Koto
Ptr-Paquita
Ptr-Taweh
Ptr-Tobi
Ptr-X00100_Koby
Ptr-Yolanda
density
0.0
0.5
1.0
1.5
1e+04
1e+05
1e+06
size
class
deletion
duplication

Diversity of CNPs in great apes: In order to explore the diversity of great ape copy number variants, we first compared the copy number diversity of each of the populations analyzed to the SNP diversity (Prado and Sudmant, companion under review). Using segregating CNVs, we computed an estimate of Watterson's θ for segregating CNVs/genome and compared this to Watterson's theta for SNPs/bp. As expected, the two were highly correlated (r^2=0.52, p=0.026; **Figure 10.4**).

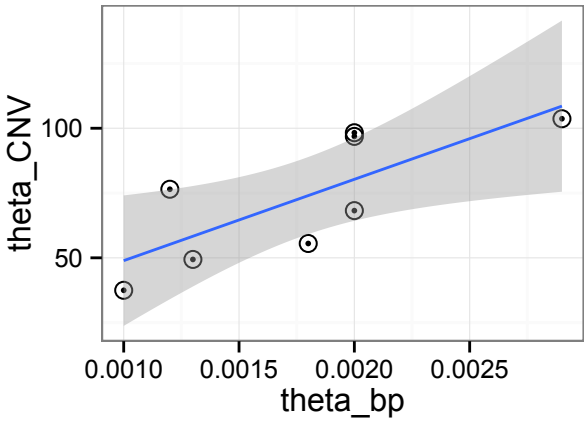


Figure 10.4: Estimates Watterson's theta (θ) for segregating CNVs/genome plotted versus segregating SNPs/bp shows a strong correlation between the diversity of SNP and CNP diversity (r^2 =0.52, p=0.026).

We next performed a Principle Components Analysis (PCA(Patterson et al. 2006)) using only deletion CNVs that could be assigned a state of 0, 1, or 2 (**Figure 10.5**). All known species' and subspecies' relationships and classifications are observed as clusters in the PCAs. Additionally, the relative levels of diversity are captured in the clusterings with more diverse populations, such as the Western gorilla, showing a more dispersed pattern. Sumatran orangutan show more diversity than Bornean orangutan, as suggested by the estimates of θ in both SNPs and CNPs. The PCAs additionally show two clusters in Sumatran orangutans, possibly demonstrating Sumatran orangutan subpopulations. Among human, the first PC discriminates African from non-African populations with the Denisova individual falling intermediate. The Denisova is closest to the Papuan and Asian individuals along PC2, which is interesting as gene flow from Denisovans to Papuan individuals has been reported(Meyer et al. 2012).

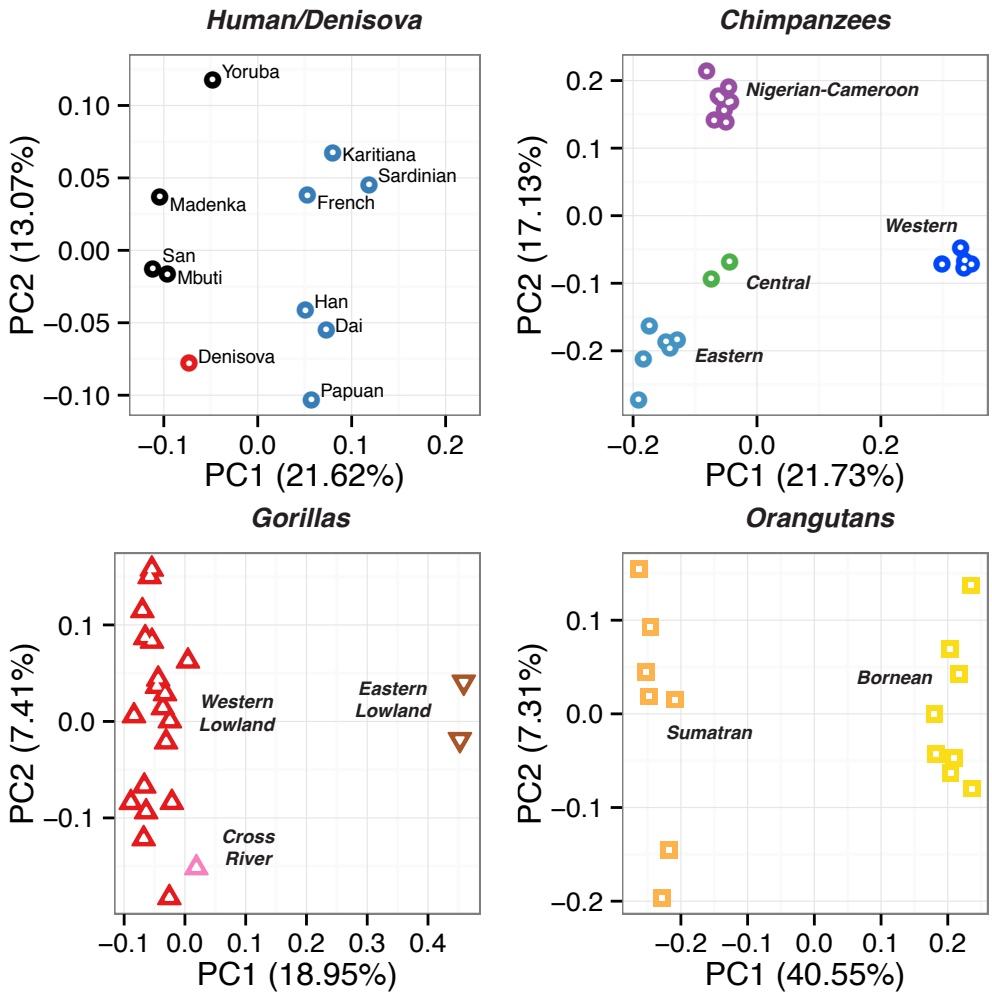


Figure 10.5: PCA plots based on CNV genotypes for chimpanzee, gorilla, and orangutan, respectively. All individual species' and subspecies' relationships are confirmed in addition to relative diversity between populations. Western lowland gorillas, for example, show far more diversity than Eastern lowland gorillas. Between orangutan species, the Sumatran orangutans show more diversity and appear to cluster into two groups, potentially a signature of Sumatran orangutan subpopulations.

We next constructed neighbor-joining trees using the genotypes of fixed and segregating variant calls to compute the hamming distance between all genomes. Species' trees were first considered using all variants >2 kbp, 3.5 kbp, and 5 kbp, respectively (**Figure 10.6**). As a function of millions of years since divergence, all trees show a significant increase in the number fixed deletions in the chimpanzee-bonobo ancestor (**Section 9**); however, we find there is a marked enrichment in the chimpanzee-bonobo branch for larger events.

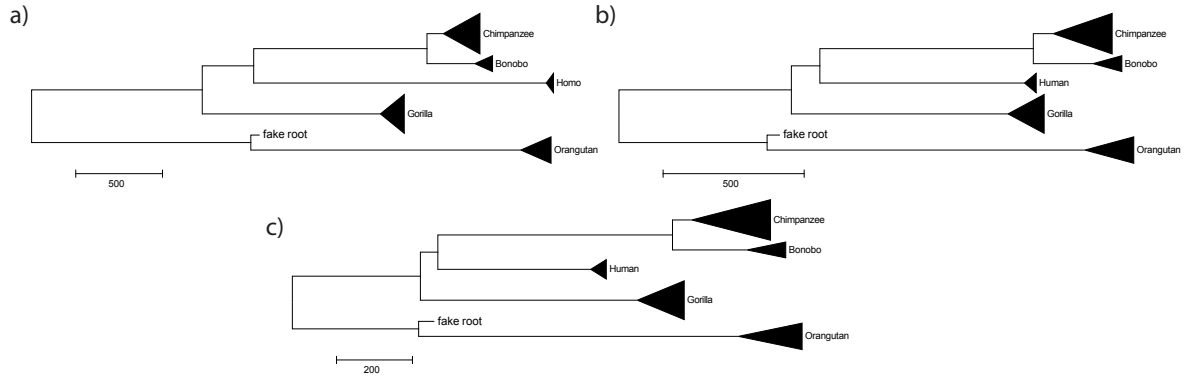


Figure 10.6: Neighbor-joining species trees constructed from segregating variants and fixed variants ≥2 kbp, ≥3.5 kbp and ≥5 kbp (a,b,c), respectively. Though the chimpanzee-bonobo branch shows an increased rate of deletion for each of these size thresholds, the effect is more pronounced for larger deletions.

We next constructed a neighbor-joining tree for all great ape individuals (**Figure 10.7**). Confidence values on branches are generated from repeatedly randomly subsampling 50% of the variants and re-computing the tree. All known species' relationships were captured with 100% confidence along with the correct species-tree topology. All the individual subspecies' topologies were additionally reconstructed with high confidence and were identical to those identified by SNPs with the exception of Eastern chimpanzee, which are placed as an outgroup to all chimpanzee subspecies. Eastern lowland gorillas place as an outgroup to all Western gorillas and the single Cross River gorilla individual as an outgroup to all Western lowland gorillas. Within the chimpanzee phylogeny, the two Nigerian-Cameroon individuals Tobi and Julie cluster together, as observed in the SNP-based tree, and the individuals Yolanda, Andromeda and Vincent, all Eastern chimpanzees from Gombe reserve in Tanzania, cluster together with 100% support. Within the human phylogeny, Africans and non-Africans form two high-confidence clades and the Denisova archaic hominid places as an outgroup to all humans robustly with 92% confidence.

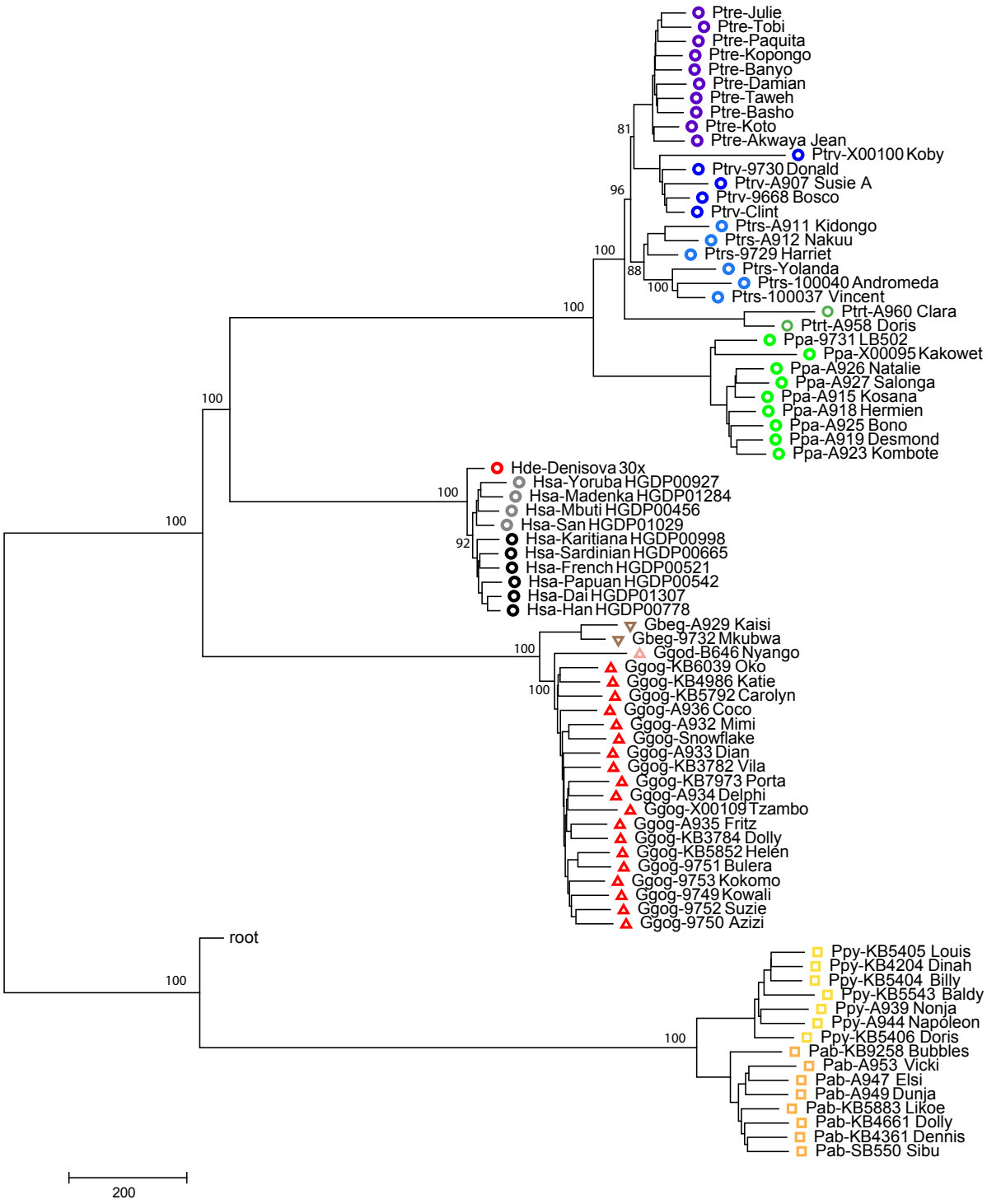


Figure 10.7: A neighbor-joining tree of great ape individuals constructed from segregating structural variation and fixed deletions >5 kbp. Confidence values on branches are generated from repeatedly randomly subsampling 50% of the variants and re-computing the tree.

CNV burden among great ape populations: Though our per-population sample sizes are small, we attempted to assess the CNV burden among different primate populations by comparing the total load of deletion and duplication events between different populations. Analysis was limited to non-segmentally duplicated regions of the genome and events >30 kbp to limit false positives. We first assessed the relative duplication load of different populations (**Figure 10.8**).

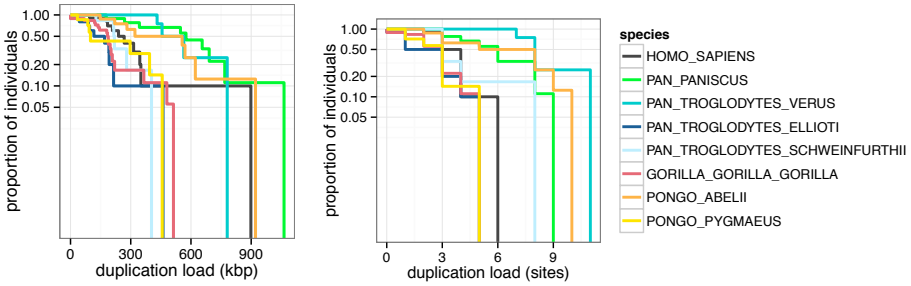


Figure 10.8: Survival functions of the proportion of individuals with at least a given total base-pair load of duplications plotted by total kbp and by number of sites. Bonobo, Western chimpanzee, and Sumatran orangutan have an increased load of duplications. The human Papuan individual harbors two large private duplications increasing the total kbp load in this individual.

Bonobo, Western chimpanzee, and Sumatran orangutan all showed a significantly higher burden of duplications compared to human, Eastern chimpanzee, Nigerian-Cameroon chimpanzee, gorilla, and Bornean orangutan (**Table 10.3**). A single human Papuan individual showed an increased duplication load as measured in base pairs as a result of two large private duplications.

Table 10.3: Statistical significance of the relative base-pair load and the relative site load of duplications between different primate populations.

Species 1	Species 2	p-value – bp load	p-value # of sites load
Bonobos	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.0014	0.005
Sumatran orangutans	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.0088	0.014
Western chimpanzees	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.02	0.004
Bonobos, Sumatran orangutans and Western chimpanzees	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	8.66e-6	3.09e-5

We next assessed the relative load of deletions between different populations identifying a significantly increased deletion burden in Western chimpanzees

compared to all other primate populations (**Figure 10.9**; p=0.00289, for base-pair load, p=0.04 for site load). One particular Western chimpanzee, Susie-A, harbored a single large ~1.6 Mbp deletion event (**Section 11**). To ensure that this single individual was not driving the significance of the load comparison, we removed her from the analysis and tested the relative load again. Though this left only three Western chimpanzees in the analysis, the result was still significant by base-pair load (p=0.02) though not by number of sites (p=0.055).

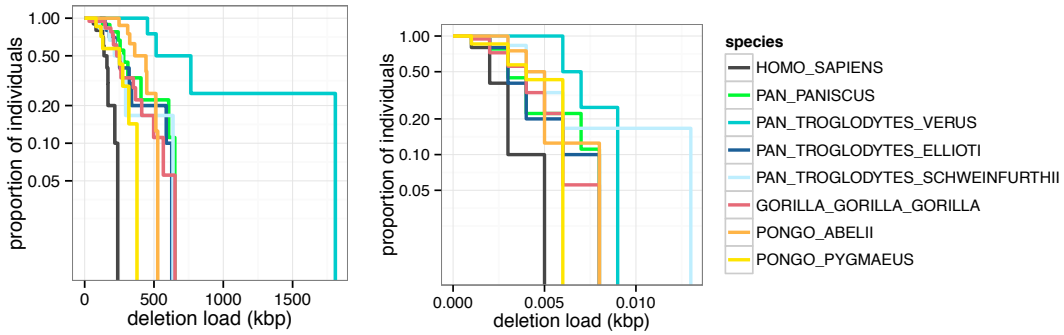


Figure 10.9 Survival functions of the proportion of individuals with at least a given total base-pair load of deletions plotted by total kbp and by number of sites. Western chimpanzees have an increased load of deletions.

Section 11: A human genomic disorder identified in a nonhuman primate

One particularly striking structural variant we identified was a ~1.6 Mbp microdeletion on 17p11.2 in the Western chimpanzee Susie-A (chr17:16,658,625-18,271,593). This deletion event encompasses 29 genes, notably including the gene retinoic acid-induced 1 (*RAI1*), deletions of which cause Smith-Magenis syndrome (SMS) in humans, a rare syndrome with an incidence of 1 in 15,000-25,000 human births. SMS is a complex neurobehavioral disorder resulting mental retardation and developmental delay, behavioral abnormalities, and facial and skeletal dismorphologies(Elsea and Girirajan 2008). We validated this event by arrayCGH against the Western chimpanzee Rappa (**Figure 11.1**).

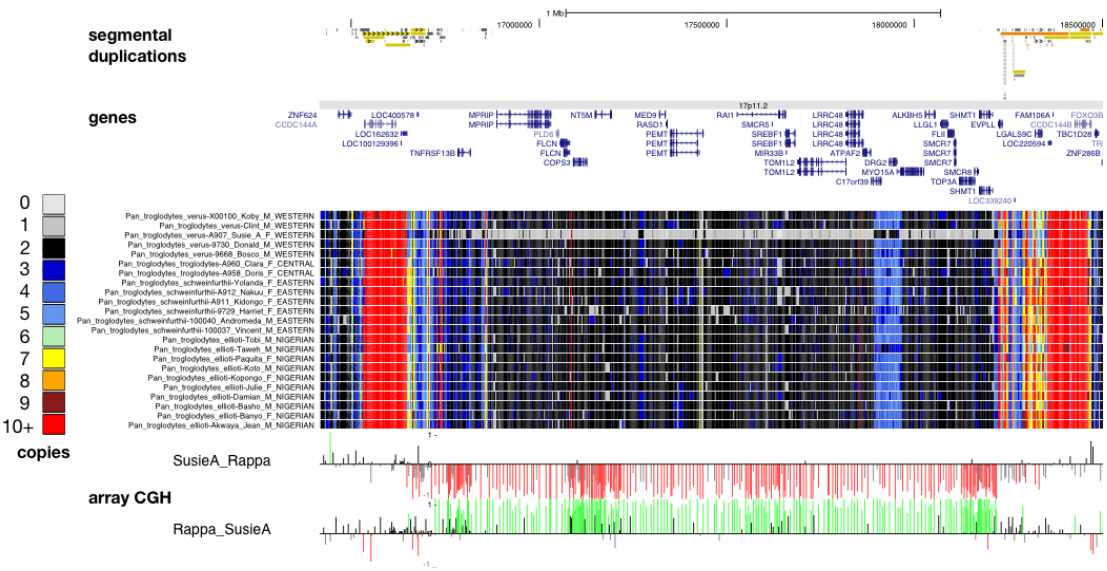


Figure 11.1: A 1.74 Mbp microdeletion on 17p11.2 in the chimpanzee Susie-A. The deletion overlaps 29 genes including *RAI1*; haploinsufficiency in *RAI1* results in SMS in humans, a severe neurobehavioral.

Though Susie-A, who was captured in 1975, died in 1997, we were able to obtain a report of her clinical features from the records of her handlers and veterinarians. These descriptions bare striking similarity to many of the phenotypes observed in SMS patients **(Table 11.1**); notably, maladaptive behaviors such as aggression and disobedience were identified in Susie-A, who was described a "mean and more aggressive than usual" chimpanzee. Susie-A also presented a hump on her back similar to many SMS patients who exhibit abnormal spinal curvatures, was obese—a common characteristic of SMS patients, and died of interstitial nephritis. Renal abnormalities are common in SMS patients. Finally, Susie-A exhibited tracheitis and had grossly overlapping tracheal cartilage ends, which contributed to noted breathing noises she would make; 50-75% of SMS patients exhibit tracheobronchial problems and velopharyngeal insufficiency.

Table 11.1: Common clinical features of SMS and related features of Susie-A.

Clinical features of Smith-Magenis syndrome	Related clinical features of Susie-A
Maladaptive behavioral issues including: • **Frequent outbursts and tantrums** • **Aggression** • **Disobedience** • **Emotional volatility** • **Tendency toward attention-**	Susie-A is described as having exhibiting • 'marked impairment in her behavioral skills,' • 'mean and more aggressive than usual' behavior • When in the close proximity to people Susie-A would palpate her vagina, which is known to be a challenging "culturally

seeking behaviors • **Lack of respect for personal space during conversation.**	abnormal" sexualized behavior in chimpanzees.
Abnormal curvature of the spine and scoliosis are present in 50-75% of individuals with SMS(Greenberg et al. 1996; Pagon et al. 1993).	Susie-A had kyphoscoliosis as a result of abnormal spine curvature.
Edelman *et al*(Edelman et al. 2007) found that individuals with *RAI1* mutations and deletions were likely to be obese. Mouse models of *RAI1* deletions additionally demonstrated an obese phenotype for deletions but not duplications of the critical locus(Walz et al. 2003; Yan et al. 2004); duplications conferred an underweight phenotype. A null *RAI1* allele in mice generated by Bi *et al*(Bi et al. 2005) also exhibited obesity.	Susie-A was an obese chimp with a body weight of ~90 kg. The normal body weight of a mature chimpanzee is between 50 kbp and 65 kg and less for female Western chimpanzees.
>75% of SMS patients exhibit otolaryngologic abnormalities including a hoarse, deep voice and 50-75% exhibit tracheobronchial problems and velopharyngeal insufficiency(Pagon et al. 1993).	Susie-A exhibited tracheitis and had grossly overlapping tracheal cartilage ends, which are suspected to contribute to documented breathing noises she would make.
Renal abnormalities have been shown to occur in 20-35% of SMS patients(Greenberg et al. 1996; Pagon et al. 1993). Additionally, 25%-50% of SMS patients have been found to have cardiac abnormalities.	Susie-A died of chronic interstitial nephritis and was found to have increased creatinine and leukocytosis consistent with renal failure. Numerous eosinophils were found in her gastrointestinal tract with no detection of parasites. Obesity and tracheitis presumably played a role in her final cardiorespiratory failure as well.

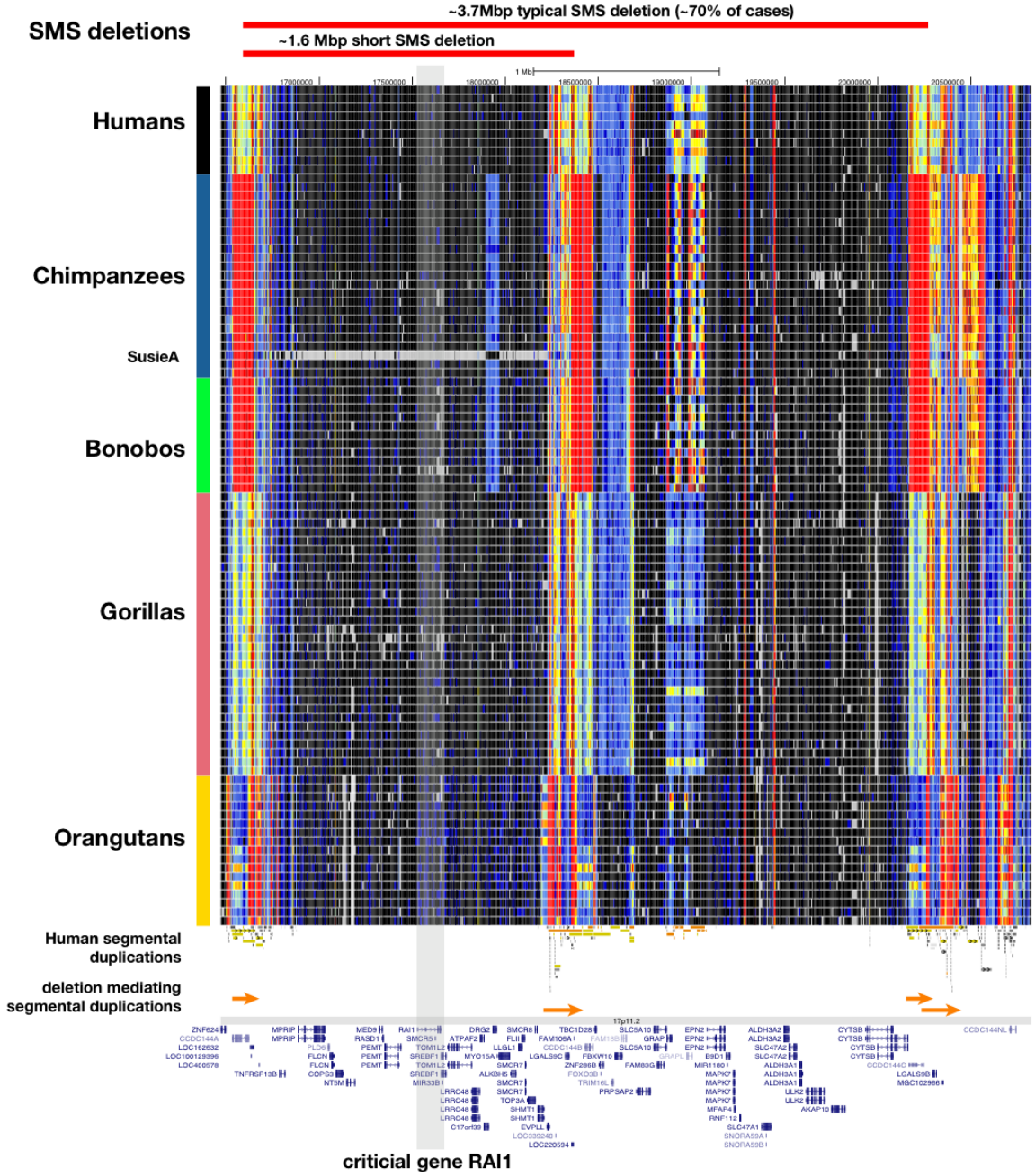


Figure 11.2: Architecture of the SMS/PTLS locus 17q11.2 in great apes. In humans, SDs in direct orientation mediate two different deletions sharing a distal breakpoint, both of which delete the critical gene *RAI1*. Susie-A exhibits the rarer of the two deletions.

SMS is caused by haploinsufficiency of *RAI1*, which is most often the result of a ~3.7 Mbp deletion mediated by two highly homologous SDs (~70% of cases(Elsea and Girirajan 2008)). Atypical deletions encompassing *RAI1* and *RAI1* mutations are responsible for an additional fraction of cases. The event we identify in Susie-A is much smaller than the typical human deletion (**Figure 11.2**).

Notably, the SDs of the 17p11.2 locus have increased in copy number specifically in the *Pan* genus (**Figure 11.3, Figure 11.4**), perhaps modifying the architecture of bonobo and chimpanzee SDs to predispose to this smaller deletion event.

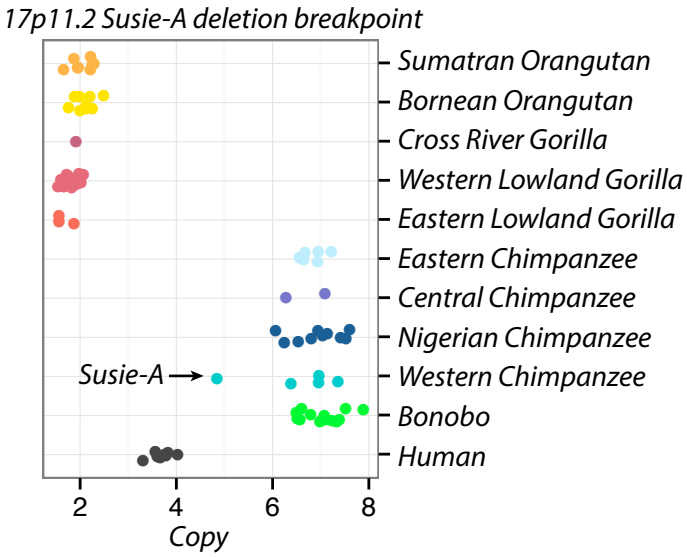


Figure 11.4: Copy number of the H duplicon at 17p11.2 is shown for different individuals clustered by great ape species. Chimpanzees and bonobos have increased copy number of the duplication ranging from 6-8 copies, with the exception of Suzie-A who has lost a copy. Humans exhibit 4 copies of the region.

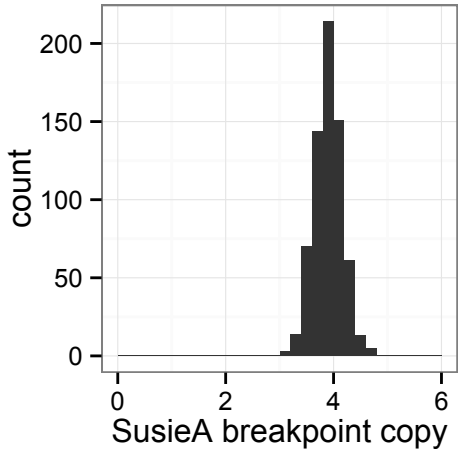


Figure 11.4: Copy number of the H duplicon, which is lost in Susie-A, genotyped across 675 individuals. 100% of humans robustly show 4 copies of this duplication.

To resolve the architecture of the chimpanzee 17p11.2 locus, we turned to large-insert clone sequencing. We began by searching for chimpanzee BACs

that mapped to the SDs of the SMS region in 17p11.2 of HG18 to determine the structure of SDs present in that region. We identified 37 such BACs that mapped to this locus and were annotated as mapping to chromosome 17 or unknown. Two of these BACs had previously been Sanger sequenced and we sequenced the remaining clones using a PacBio RS system, successfully assembling 20 more nonredundant clones (**Table 11.2**). To determine true anchor positions for each BAC, we ran whole-genome shotgun sequence detection (WSSD(Bailey et al. 2002)) of duplication content in BACs using human, chimpanzee, gorilla, and orangutan reads and determined regions in BACs with no coverage by WSSD in any species. We searched HG18 with these unique regions to find the highest identity location and considered these positions the "anchor" for each BAC. We were able to anchor 13 BACs in this manner. Aligning BACs to each other, we were then able to construct three super-contigs from seven BACs. Finally, we used DupMasker(Jiang et al. 2008) to determine the content of the duplication blocks in humans and chimpanzees and aligned our contigs to both the human and chimpanzee reference genomes to assess the differences in content and structure (**Figure 11.5**). BACs that we could not anchor with 100% certainty were placed to their most likely location given their structure and content in the context of Susie-A's deletion.

Table 11.2 : Sequenced BACs assessed and sequencing technology used.

Clone name	Source
CH251-545A24 (AC183294.3)	Sanger
CH251-59G18	PacBio
CH251-60C22 (AC183837.3)	Sanger
CH251-21N5	PacBio
CH251-182P19	PacBio
CH251-425F18	PacBio
CH251-33G11	PacBio
CH251-179D15	PacBio
CH251-133B4	PacBio
CH251-200P13	PacBio
CH251-433E19	PacBio

Clone name	Source
CH251-35C10	PacBio
CH251-48B19	PacBio
CH251-354I21	PacBio
CH251-118C16	PacBio
CH251-13D21	PacBio
CH251-173I3	PacBio
CH251-184C19	PacBio
CH251-18I21	PacBio
CH251-251H7	PacBio
CH251-30P23	PacBio
CH251-75B17	PacBio

Table 11.3: Summary of BACs used to resolve the architecture of the SMS region in chimpanzees. Clone information is grouped by sequencing technology (PacBio and Sanger). PacBio clones were sequenced and assembled in-house; Sanger clones were sequenced at The Genome Institute at Washington University.

Clone assembly	Number	Size (bp)	Unique by WSSD (bp)	Discontiguous wrt HG18 (bp)	Contiguous and inverted wrt HG18 (bp)	Collapsed duplications in assembly (bp)
PacBio	20	2,935,821	2,181,309	1,194,224	71,582	268,000
Sanger	2	413,710	95,750	219,862	0	N/A
Total	22	3,349,531	2,277,059	1,414,086	71,582	268,000

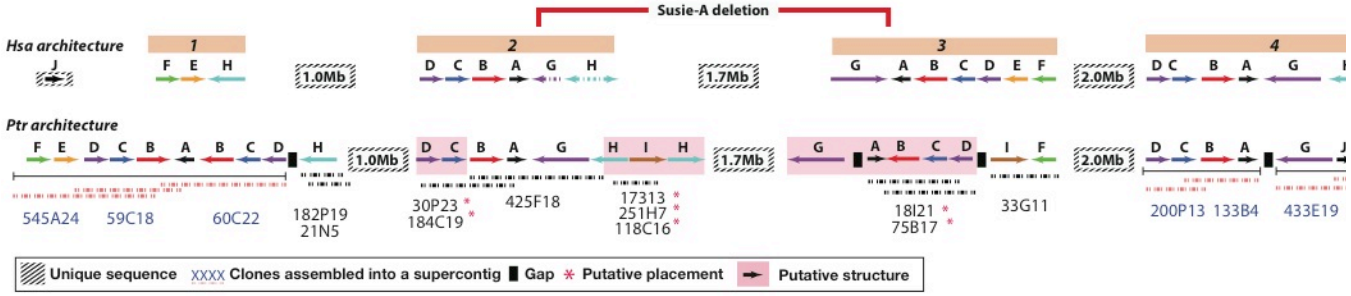


Figure 11.5: Duplicon structure of the human and chimpanzee 17p11.2 locus. Duplication blocks are labeled 1-4 and annotated with the underlying structure of SDs. We hypothesize that the directly oriented G-duplicon in chimpanzees is responsible for mediating the Susie-A deletion.

Section 12: References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bi W, Ohyama T, Nakamura H, Yan J, Visvanathan J, Justice MJ, Lupski JR. 2005. Inactivation of Rai1 in mice recapitulates phenotypes observed in chromosome engineered mouse models for Smith-Magenis syndrome. *Hum Mol Genet* **14**: 983–995.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

Edelman EA, Girirajan S, Finucane B, Patel PI, Lupski JR, Smith ACM, Elsea SH. 2007. Gender, genotype, and phenotype differences in Smith-Magenis syndrome: a meta-analysis of 105 cases. *Clin Genet* **71**: 540–550.

Elsea SH, Girirajan S. 2008. Smith–Magenis syndrome. *Eur J Hum Genet* **16**: 412–421.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**: 1629–1631.

Greenberg F, Lewis RA, Potocki L, Glaze D, Parke J, Killian J, Murphy MA, Williamson D, Brown F, Dutton R, et al. 1996. Multi-disciplinary clinical study of Smith-Magenis syndrome (deletion 17p11.2). *Am J Med Genet* **62**: 247–254.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat*

Methods **7**: 576–577.

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**: 1270–1278.

Jiang Z, Hubley R, Smit A, Eichler EE. 2008. DupMasker: a tool for annotating primate segmental duplications. *Genome Res* **18**: 1362–1368.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci USA* **109**: 15716–15721.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lichter P, Tang CJ, Call K, Hermanson G, Evans GA, Housman D, Ward DC. 1990. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**: 64–69.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529–533.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**: 216–219.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**: 222–226. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22936568&retmode=ref&cmd=prlinks.

Miller DT, Nasir R, Sobeih MM, Shen Y, Wu B-L, Hanson E. 1993. *16p11.2 Microdeletion - GeneReviews*. University of Washington, Seattle, Seattle

(WA) http://www.ncbi.nlm.nih.gov/books/NBK11167/.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**: 18–23.

Pagon RA, Bird TD, Dolan CR, Stephens K, Adam MP, Smith AC, Boyd KE, Elsea SH, Finucane BM, Haas-Givler B, et al. 1993. *Smith-Magenis Syndrome*. University of Washington, Seattle, Seattle (WA).

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161–1165.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**: 562–578.

Varki A, Geschwind DH, Eichler EE. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat Rev Genet* **9**: 749–763.

Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, et al. 2011. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* **21**: 1640–1649.

Walz K, Caratini-Rivera S, Bi W, Fonseca P, Mansouri DL, Lynch J, Vogel H, Noebels JL, Bradley A, Lupski JR. 2003. Modeling del(17)(p11.2p11.2) and dup(17)(p11.2p11.2) contiguous gene syndromes by chromosome

engineering in mice: phenotypic consequences of gene dosage imbalance. *Mol Cell Biol* **23**: 3646–3655.

Witkin A. 1984. Scale-space filtering: A new approach to multi-scale description. **9**: 150–153.

Yan J, Keener VW, Bi W, Walz K, Bradley A, Justice MJ, Lupski JR. 2004. Reduced penetrance of craniofacial anomalies as a function of deletion size and genetic background in a chromosome engineered partial mouse model for Smith-Magenis syndrome. *Hum Mol Genet* **13**: 2613–2624.

VITA

Peter James Heshedahl Sudmant was born in Vancouver, Canada, in 1985. He attended Queen Mary Elementary School and then Lord Byng Secondary where he developed a passion for math, computer science and literature. In 2003 he enrolled in Computer Engineering at the University of Waterloo in Waterloo, Ontario. In 2005 he switched his major to Science and shortly after switched again to the Department of Computer Science (CS) and Mathematics to focus on a CS degree with a focus on bioinformatics and biology. Peter graduated with honors in 2008 and, having been inspired to pursue research in the lab of Dr. Wyeth W. Wasserman at the University of British Columbia, he then enrolled in a PhD program at the University of Washington under the excellent mentorship of Dr. Evan E. Eichler. Peter very much enjoyed his time as a student in the Eichler lab and graduated with a Doctor of Philosophy in Genome Sciences in 2013. In addition to his studies, Peter is avid runner, cross-country skier and hiker who enjoys listening to and playing music and reading the New Yorker.