Human-specific duplicate genes: new frontiers for disease and evolution

Xander Nuttle

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

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Evan E. Eichler, Chair
Chris T. Amemiya
Christine Queitsch

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Human-specific duplicate genes: new frontiers for disease and evolution

Xander Nuttle

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

Gene duplication is a fundamental force contributing to the evolution of novel traits, genomic diversity among species and individuals, and disease. In this dissertation, I characterize the evolutionary history, diversity, functional potential, and disease relevance of gene families that emerged specifically along the lineage leading to human. I leveraged a haploid clone library to resolve the sequence and structure of four human *SRGAP2* paralogs, adding ~380 kbp of sequence to the human reference genome. Analyzing this high-quality sequence, I found that the promoter and first nine exons of *SRGAP2* duplicated three times across chromosome 1, ~3.4-1 million years ago. All paralogs produce mRNA transcripts, but *SRGAP2C* is most highly expressed and has fixed in copy number in the human population, making it the most likely functional duplicate. To screen large cohorts of autism and intellectual disability patients for mutations that disrupt *SRGAP2C*, I developed a method to genotype paralog-specific copy number and sequence variation using molecular inversion probes. I demonstrated that this method was broadly applicable to large-scale genotyping of previously inaccessible duplicated genes. Using this method, I also discovered regions of interlocus gene conversion between duplicated sequences >80 Mbp apart on the same chromosome and refined unequal crossover breakpoints for copy number polymorphisms at the *RH* locus. Finally, I employed my genotyping method and strategies used to characterize *SRGAP2* duplications to study *BOLA2*, a gene duplicated specifically in *Homo sapiens* located at chromosome 16p11.2. Sequencing this region in orangutan and chimpanzee revealed drastic rearrangements between species, including six inversions affecting 47 genes. I determined that an ~95 kbp segment including *BOLA2* duplicated ~282 thousand years ago, specifically predisposing humans to

recurrent microdeletions and microduplications associated with autism. I demonstrate that despite its young age and its conferring susceptibility to rearrangements, the *BOLA2* duplication has nearly fixed in the human lineage. I show that *BOLA2* duplication resulted in a *Homo sapiens*-specific in-frame fusion transcript and that expression correlates with genomic copy number. Collectively, my work provides new insights into the birth, evolution, and disease relevance of duplicate genes, pioneers new genotyping technology, and identifies specific gene innovations as novel candidates for the evolution of uniquely human traits.

Table of Contents

List of Figures

List of Tables

Acknowledgements

It has been an incredible journey chasing my scientific childhood dream, from reading articles about the natural world in *Discover* and *Scientific American* growing up to seeing my research covered in these very places. Along the way, I have been blessed with a wonderful family, incredible mentors, collaborators, and friends whose love, wisdom, passion, inspiration, and encouragement have been invaluable. To everyone who has shared in my voyage of learning and discovery: thank you.

First and foremost, I would like to thank my parents, Allen and Barb, and my brothers Jon, Mike, and Ben. I also thank my grandparents Grandma, Papa, and Grandma Pat. My love of learning was fostered by my family and by many teachers growing up, especially my elementary school teacher Mrs. Wilshire and high school teachers Mr. Krotec and Br. Ernest Miller. Marisa Pedulla and Graham Hatfull provided me my first research experience in high school through the Phagehunting Program at the University of Pittsburgh. Hunt Willard organized and taught a riveting course for the Genome Revolution Focus Program at Duke, getting me really excited to study genome sciences and always taking keen interest in my scientific pursuits. My freshman year of college, Greg Wray and lab members David Garfield, Courtney Babbitt, and B.J. Nielsen welcomed me into their research group studying genetic bases for human evolution—my scientific home as an undergraduate. To all these people and others who made these experiences so interesting, fun, and fulfilling: thank you.

Here at the University of Washington, I have enjoyed the company of great classmates, as well as wonderful people and scientists every day in the Eichler Lab. I'd like to recognize Megan Dennis for teaming up on the *SRGAP2* project, as well as postdocs Santhosh Girirajan, Can Alkan, Francesca Antonacci, Emre Karakoc, Brad Coe, Fereydoun Hormozdiari, Mark Chaisson, Osnat Penn, Bo Xiong, and Stuart Cantsilieris. I am also thankful to all my fellow students during my time with the lab: Michael Duyzend, Max Dougherty, Peter Sudmant, Nik Krumm, and Madeleine Geisheker. I appreciate the friendship and efforts of other lab members over the years, including Maika Malig, Laura Vives, Carl Baker, John Huddleston, Giorgia Chiatante, Archana Raja, Arthur Ko, and Tonia Brown.

The Genome Center at Washington University in St. Louis was instrumental in performing sequencing for the *SRGAP2* project, and I acknowledge Jay Shendure for helpful discussion and guidance in developing my molecular inversion probe genotyping method. I thank Giuliana Giannuzzi and Alex Reymond for teaming up on the *BOLA2* project, as well as Josh Schraiber, Iñigo Narvaiza, and Josh Akey for their contributions and helpful conversations. My committee has provided useful guidance and feedback throughout my doctoral studies. Thanks to Harmit Malik, Christine Queitsch, Chris Amemiya, and Bob Waterston.

Finally, I would like to express my deepest gratitude to my mentor, Evan Eichler. He allowed me to work on a fantastic project combining genome science, disease, and evolution—a synthesis that has always fascinated me. He provided unwavering support, challenged me to work hard to answer difficult questions and make the most of my abilities and opportunities, and celebrated new findings, successful oral presentations, and publications. I have been very blessed to be able to learn from someone who embodies the passion, hard work, dedication, and love I hope to carry forward.

1. Introduction

1.1 Duplicate Genes and Evolution

Gene duplication, a mutational event producing an extra copy of a gene within a genome, is a potent evolutionary force [1-10]. In his seminal 1970 book [1], Susumu Ohno argued that gene duplication fosters the evolution of novel function. By creating new sequence with functional redundancy, gene duplication enables evolutionary exploration. Duplicate genes can undergo many different fates [1, 2, 5, 8, 11-18] (**Figure 1.1**). Most commonly, they are lost from the population or pseudogenized via deleterious mutations. If retained, they can function as the original copy, serve to partition multiple functions of the original copy, or acquire a new function. This last possibility was especially compelling to Ohno, leading him to propose gene duplication as the major mutational mechanism driving the evolution of complex life [1].

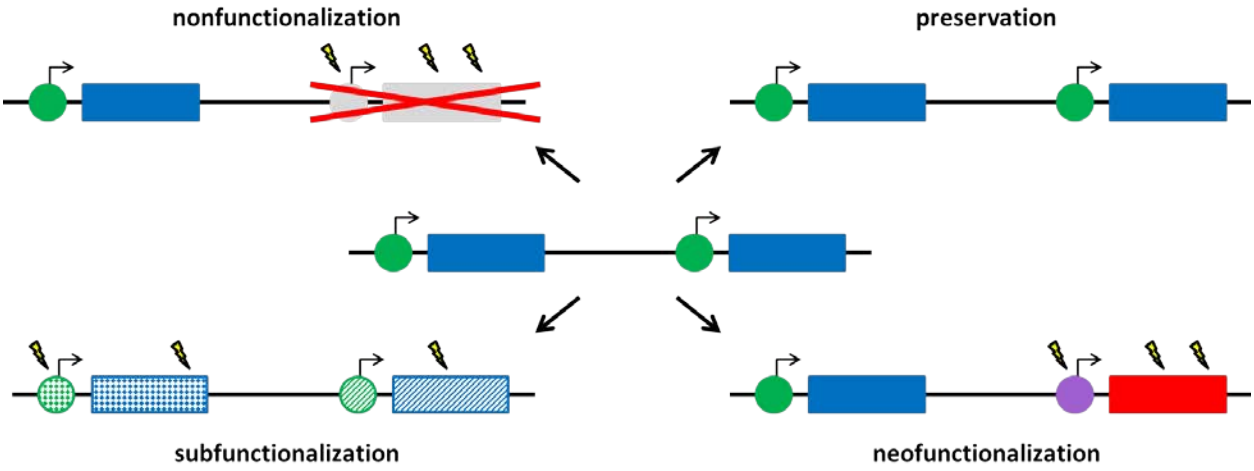


Figure 1.1. Potential Fates for a Duplicate Gene Pair. A newly formed duplicate gene pair (center) can undergo one of several possible fates [1, 2, 5, 8, 11-18]. Most frequently, one copy is eliminated from the population by genetic drift (not shown) or inactivated by disruptive mutations in regulatory (green circle) or coding (blue rectangle) sequence (upper left). Occasionally, two functional copies are retained. If no mutations alter the function of either copy, the extra copy results in in increased dosage of the original gene product (upper right). Otherwise, the two copies may evolve to optimize different functions both performed by the original gene (lower left), one copy may evolve a novel function (lower right), or retained copies may undergo some combination of these scenarios, often involving the evolution of multifunctional proteins [19].

In addition to evolutionary innovation, gene duplication immediately and sometimes drastically modifies genome structure [20-22]. Whole-genome duplication events have been inferred to have occurred in yeast [23] and in early vertebrates [24], doubling the size and gene content of corresponding genomes. Segmental duplications, events involving a contiguous chromosomal region at least 1 kbp in size [25], have also played a major role in karyotypic evolution [3, 4, 21, 26-28]. Hominoid genomes are enriched for such duplications in pericentromeric and subtelomeric regions that participated in chromosomal rearrangements distinguishing different species [29, 30]. Rearrangement breakpoints often map within duplicated sequences, underscoring their importance for genomic restructuring.

Gene families are abundant in organisms from myriad branches of the tree of life, testifying to the breadth and importance of gene duplication throughout evolution [5, 7, 9, 31, 32]. Duplication still occurs in genomes today—recent segmental duplication events and associated copy number variation have been documented in many species [25, 33-39]. Although gene duplication is widespread, its extent and spatial distribution within the genome differ among different lineages [40, 41]. For example, most duplicated sequences in mouse and other non-hominoid mammals occur in tandem [35, 37, 42]. In contrast, human and African great ape duplications are commonly interspersed [3, 4, 10, 40, 43]. This interspersed architecture predisposes our genomes and those of our closest evolutionary relatives to microdeletions, microduplications, and inversions, underlying structural genomic differences among species as well as genomic disorders [3, 4, 20, 22, 30, 44-46].

1.2 Duplicate Genes and Disease

Genomic disorders are defined as diseases resulting from nonallelic homologous recombination (NAHR) between duplicated sequences [44]. Because of their homology, duplicated sequences at different genomic locations sometimes align with each other during meiosis. Crossover between them results in the duplication, deletion, or inversion of intervening sequence (**Figure 1.2**). Such mutations typically affect hundreds of kbp to a few Mbp of sequence, often altering the copy number of dosage-sensitive genes [44]. Even inversions can have adverse consequences if they disrupt a gene [47] or create

new regulatory contexts [48]. Collectively, genomic disorders have been implicated in several human diseases [49-53], including Charcot-Marie-Tooth disease type 1A [54], Smith-Magenis syndrome (SMS) [55], velocardiofacial syndrome [56], and Williams syndrome [57]. Microdeletions and microduplications resulting from NAHR have also been associated with autism spectrum disorder [58-61], schizophrenia [62-64], macrocephaly [65, 66], microcephaly [58, 65, 66], and extremes of body mass index [67, 68], among other conditions. Finally, genomic disorders do not appear to be exclusive to humans. A chimpanzee with an NAHR-mediated deletion including sequence orthologous to that deleted in SMS exhibited striking behavioral anomalies reminiscent of the SMS phenotype in humans [41].

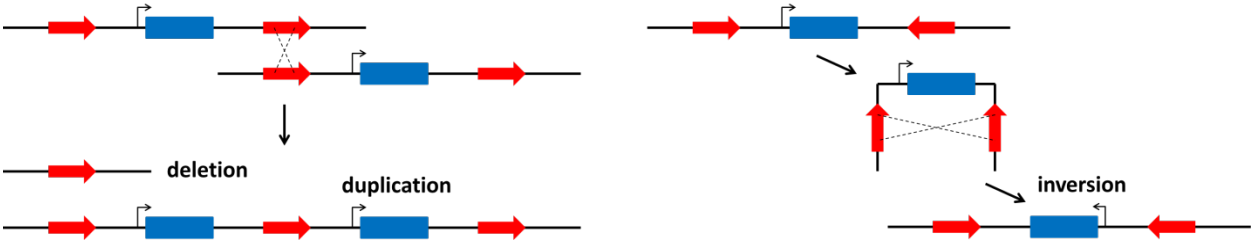


Figure 1.2. Duplication, Deletion, and Inversion via NAHR. (Left) Interspersed segmental duplications (red arrows) in direct orientation can undergo unequal crossover during meiosis, leading to the duplication and deletion of intervening sequence and genes therein. An interchromosomal event involving two of four chromatids is diagrammed (the two chromatids not involved in unequal crossover are not shown). (Right) Recombination between inversely oriented interspersed duplications results in inversion. An intrachromatidal event is diagrammed (the three chromatids not involved in unequal crossover are not shown).

1.3 The Paradox of Interspersed Duplications

Given their propensity to mediate disease-associated rearrangements, interspersed segmental duplications should be disfavored by evolution. Yet the genomes of humans and African great apes are peppered with these very duplications [4]. One intriguing hypothesis put forward to explain this discrepancy between expected and observed levels of interspersed duplications is that they serve as nurseries for the birth of novel advantageous genes [43]. Consistent with this idea, human duplications are enriched for transcriptional activity compared to random human genomic sequence [25, 69, 70]. Annotated genes within duplicated regions frequently exhibit strong signals of positive selection [71-76].

Finally, interspersed duplications have potential to promote domain accretion, the formation of multidomain proteins via the juxtaposition of sequences encoding different domains [3, 4, 71, 77, 78] (**Figure 1.3**). For all these reasons, interspersed segmental duplications are good candidate progenitors for new genes and possible associated evolutionary adaptations.

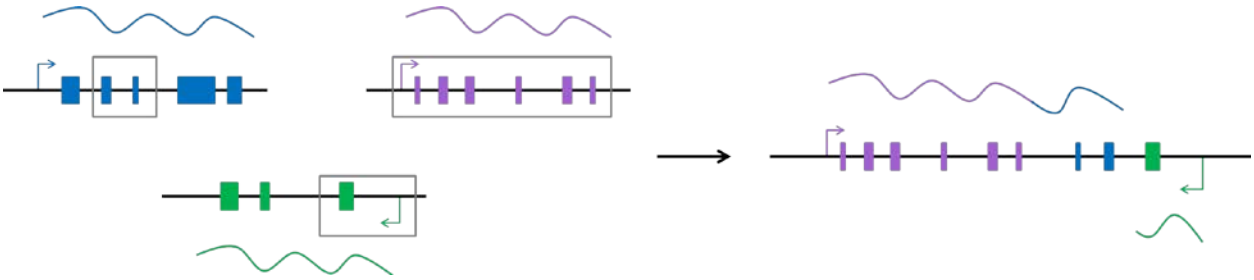


Figure 1.3. Novel Gene Formation via Interspersed Segmental Duplication. Schematic highlights how duplication events can bring together sequences from distinct loci and create novel genes [71, 77, 78]. In this case, three duplications events including sequences in gray boxes (left) juxtaposed exonic sequences from three genes. The duplicate locus (right) produces a truncated isoform of one gene (green) and a fusion transcript combining sequences from two genes (blue and purple).

1.4 Human-Specific Duplicate Genes

Intense interest in human uniqueness has motivated a systematic effort to compile all known information about human-specific differences from other great apes (http://carta.anthropogeny.org/content/about-moca). We have long recognized hallmark traits distinguishing our species from other primates [79], including language [80], losses of penile spines [81, 82] and body hair [83, 84], bipedalism [83, 85], and diet-related adaptations [86] such as increased tooth enamel thickness [87-89] and a gracilized jaw [88]. Most notably, human brain development differs drastically from our closest evolutionary relatives. Exclusively human properties of brain development include expansion of the cerebral cortex [90-92], neoteny of synaptic maturation [93-95], and divergent dendritic spine morphology [96, 97].

In stark contrast, with a few possible exceptions [79, 98-116], we have yet to discover the genetic foundations underlying our most distinctive phenotypes. Primarily due to the low quality of reference genomes for chimpanzee [74], bonobo [117], gorilla [118], and orangutan [119], a comprehensive catalog

of genomic differences between ape species remains elusive even today. Nevertheless, two key studies in 2004 and 2010 identified a potentially critical set of uniquely human genomic features: human-specific duplicate (HSD) genes [120, 121]. Defined as complete or partial copies of a gene originating via segmental duplication events specifically along the human lineage and rising to appreciable frequencies, such genes are absent from nonhuman primate genomes. About half of the HSD genes are single-copy in nonhuman apes (>23 gene families, Figure 1.4), while the other half are multicopy in at least one nonhuman ape species (>30 gene families) [121].

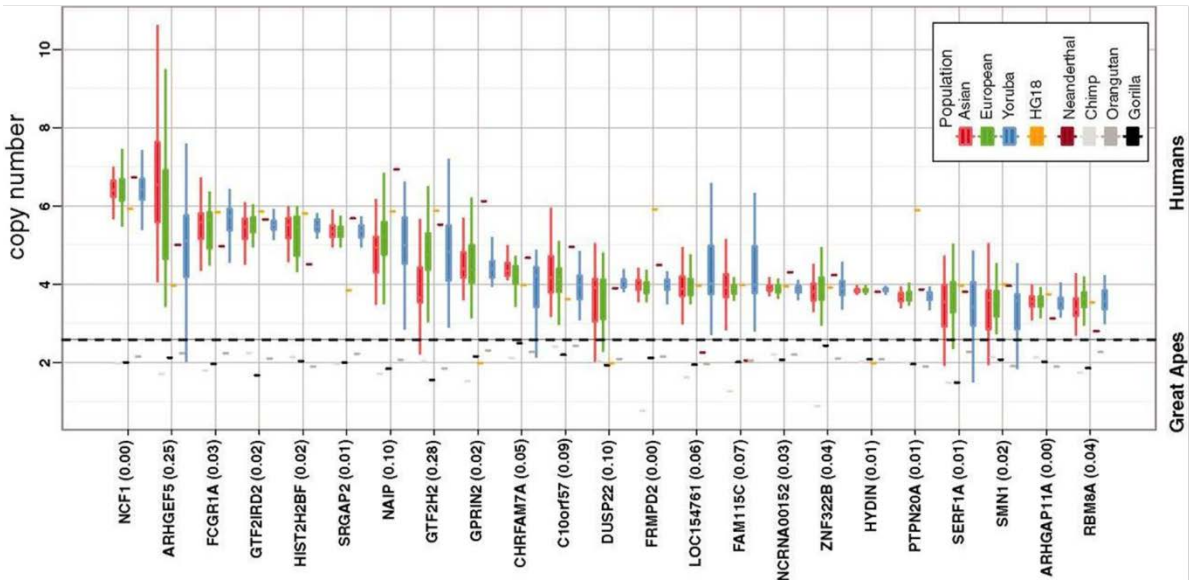


Figure 1.4. HSD Genes That Are Single-Copy in Nonhuman Apes. Genes in this category fall into 23 gene families arranged along the horizontal axis. Boxplots indicate estimated aggregate copy number for each gene family in human populations (red, green, and blue; 159 individuals total), the human reference genome NCBI36 (yellow), and nonhuman primates (gray; 1 individual per species). Modified from [121]. Since this study, a few additional genes in this category have been identified, including *BOLA2*.

Many of the challenges surrounding the identification of HSD genes make them especially difficult to investigate and characterize. Even the basic task of establishing accurate sequences for all family members is complicated by several factors. HSD genes are commonly contained in regions of the genome not properly represented in the human reference sequence, regions where gaps and misassemblies abound [98, 121-123]. Many HSD genes are copy number polymorphic, included in structural variation

not well captured in studies genotyping a panel of single nucleotide variants [34, 120, 124, 125]. Due to their young evolutionary age, HSD genes often exhibit high (>99%) sequence identity with other family members, making it challenging to resolve allelic from paralogous variation [98, 123].

Efforts to understand the function of HSD genes lag behind genomic studies, as general difficulties surrounding assessing function pose additional challenges beyond the obstacles above. Furthermore, common strategies to test function such as knockout experiments in model organisms such as mouse or *Drosophila* are not directly applicable to studies of HSD genes. Thus, research into the function of HSD genes has been mostly restricted to gene ontology categorical enrichment analyses [120, 121, 126]. Considering what is known about the functions of ancestral paralogs, intriguingly, HSD genes are overrepresented for functions related to neuronal cell death ($p=0.00057$) and neurological disease ($p=0.046$). Furthermore, many HSD genes have ancestral paralogs with some connection to brain development, including *SRGAP2* [127], *GPRIN2* [128], and *HYDIN* [129]. These results, together with the general evolutionary potential of duplicated genes detailed above, suggests an exciting hypothesis: HSD genes contributed to human brain evolution.

1.5 Research Goals

My major goal in this dissertation is to characterize HSD genes—specifically, to establish their genomic sequences, to chronicle their evolutionary histories, to explore their genetic diversity, to evaluate their functional potentials, and to determine their roles in disease. I generated high-quality sequence data for human and nonhuman primate genomes and used it to resolve errors in reference assemblies. I leveraged this sequence data together with phylogenetic analyses to elucidate evolutionary histories. Taking advantage of paralogous sequence variants, I developed new experimental and computational methods for assaying genetic variation in duplicated genes. I performed RT-PCR, cDNA sequencing, and RNA-seq analysis to shed light on expression patterns of HSD genes, reveal maintained open reading frames, and investigate potential fusion transcripts. Finally, I screened large cohorts of patients with autism spectrum disorder and/or intellectual disability for mutations that disrupt HSD genes. Ultimately,

this work explores the potential role of gene duplication in human origins, forging a deeper understanding of the tight relationships between gene duplication, disease, and evolution.

1.6 Chapter Organization

In this dissertation, I detail the evolution, diversity, functional potential, and disease significance of HSD genes. Chapter two describes the emergence of the *SRGAP2* gene family via three independent duplication events ~1-3.4 million years ago, resulting in four *SRGAP2* paralogs distributed across >80 Mbp of chromosome 1. This chapter highlights the use of an effectively haploid hydatidiform mole clone library to resolve sequences of HSD genes having >99% identity and to fix nearly 1 Mbp of euchromatic sequence within the human reference assembly. Chapter two also demonstrates that only one duplicate, *SRGAP2C*, is likely functional because others are copy number polymorphic and expressed at comparatively low levels. Finally, this chapter illustrates a new mechanism for functional divergence of duplicate genes—incomplete duplication of the ancestral gene yielding a dominant negative copy at birth.

In chapter three, I detail a method leveraging molecular inversion probes (MIPs) and paralogous sequence variants to genotype paralog-specific copy number and sequence content of duplicated genes. I show that this approach is more rapid, more accurate, and more scalable than strategies based on PCR, hybridization, or whole-genome sequencing. Importantly, because of these properties, the MIP method makes accessible over 900 duplicated genes which have been excluded from studies of human variation and disease due to their repetitive nature. Chapter three also reports the discovery of novel sites of interlocus gene conversion and refinement of unequal crossover breakpoints. Surprisingly, I found signatures of gene conversion between *SRGAP2* paralogs on different arms of chromosome 1, separated by >80 Mbp.

Chapter four presents the remarkable degree of evolutionary structural rearrangement between human, chimpanzee, and orangutan genomes at the chromosome 16p11.2 locus. Comparative high-quality sequence data from these species reveal how this region nearly doubled in size via a series of segmental duplications and is the most active known genomic locus for inversions, with six independent inversion

events occurring within the last ~15 million years. This chapter also delineates how segmental duplications occurring along the human lineage after divergence from Neanderthal and Denisova gave rise to the *BOLA2* gene family. Remarkably, the *BOLA2B* paralog has nearly fixed in the human species despite having arisen only ~282 kya and having created an unstable architecture mediating recurrent microdeletions and microduplications associated with disease. I identify correlations between *BOLA2* copy number and expression at the mRNA and protein levels, as well as a novel fusion transcript between the *BOLA2* and *SMG1P* genes maintaining an open reading frame across the fusion junction.

Chapter five reflects on lessons learned from my studies of HSD genes, presents preliminary data on *SRGAP2C* mutations in intellectual disability patients and variable chromosome 1q21 rearrangement breakpoints, and offers a perspective on the future of human evolutionary genomics. Detailed methodological details, supporting figures, and supporting tables for chapters two through four are provided in appendices A-C and a CD included as pocket material.

2. Evolution of Human-Specific Neural *SRGAP2* Genes by Incomplete Segmental Duplication

This chapter has been published: Dennis, MY*, Nuttle X*, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, Curry CJ, Shafer S, Shaffer LG, de Jong PJ, Wilson RK, Eichler EE. *Cell* **149**, 912-922 (2012).

*These authors contributed equally to this work. I assembled the *SRGAP2* contigs, designed the FISH experiment to determine the likely evolutionary order of *SRGAP2* duplication events, performed breakpoint sequence analyses, performed phylogenetic analysis of the chromosome 1p12 region distal to *SRGAP2C*, discovered *SRGAP2D* with Tina A. Graves, performed *SRGAP2B* Hardy-Weinberg equilibrium analysis, conducted RT-PCR, cloning, and capillary sequencing experiments, developed and performed paralog-specific qPCR assays, and wrote the paper with Megan Y. Dennis and Evan E. Eichler.

2.1 Summary

Gene duplication is an important source of phenotypic change and adaptive evolution. We leverage a haploid hydatidiform mole to identify highly identical sequences missing from the reference genome, confirming that the cortical development gene Slit-Robo Rho GTPase-activating protein 2 (*SRGAP2*) duplicated three times exclusively in humans. We show that the promoter and first nine exons of *SRGAP2* duplicated from 1q32.1 (*SRGAP2A*) to 1q21.1 (*SRGAP2B*) ~3.4 million years ago (mya). Two larger duplications later copied *SRGAP2B* to chromosome 1p12 (*SRGAP2C*) and to proximal 1q21.1 (*SRGAP2D*) ~2.4 and ~1 mya, respectively. Sequence and expression analyses show that *SRGAP2C* is the most likely duplicate to encode a functional protein and is among the most fixed human-specific duplicate genes. Our data suggest a mechanism where incomplete duplication created a novel gene function—antagonizing parental *SRGAP2* function—immediately ‘‘at birth’’ 2–3 mya, which is a time corresponding to the transition from *Australopithecus* to *Homo* and the beginning of neocortex expansion.

2.2 Introduction

Several genes have been implicated as being important in specifying unique aspects of evolution along the human lineage. These include genes involved with the development of language (*FOXP2*) [109], changes in the musculature of the jaw (*MYH16*) [101], and limb and digit specializations (*HACNS1*) [110]. Despite these intriguing candidates, the bulk of the morphological and behavioral

adaptations unique to the human lineage remains genetically unexplained. Not all genes, however, have been amenable to standard genetic analyses. This is particularly true for genes embedded within recently duplicated sequences [25], which are frequently missing or misassembled from the reference genome [122]. Genes residing in these complex regions are important to consider for three reasons: (1) duplicated genes have been recognized as a primary source of evolutionary innovation [1, 18]; (2) the human and great-ape lineages have experienced a surge of genomic duplications over the last 10 million years [40]; and (3) human-specific duplications are significantly enriched in genes important in neurodevelopmental processes [120, 121].

Among these human-specific duplicated genes, *SRGAP2* was recently shown to be important in cortical development [127, 130]. The gene encodes a highly conserved protein expressed early in development when it acts as a regulator of neuronal migration and differentiation by inducing filopodia formation, branching of neurons, and neurite outgrowth. Analysis of the human reference genome revealed that *SRGAP2* was misassembled and that most of its duplicate copies were not yet sequenced or characterized. We developed an approach by using genomic material devoid of allelic variation (from a complete hydatidiform mole [131]) to completely sequence and characterize the missing loci corresponding to this human-specific gene family. These data allowed us to reconstruct the complex evolutionary history of this gene family since humans diverged from nonhuman primates (~6 million years ago [mya]; [132]), understand the potential of these loci to generate functional transcripts, and assay the extent of human genetic variation. We put forward a model for gene evolution in which incomplete segmental duplication creates derivative copies that antagonize the ancestral function.

2.3 Results

2.3.1 Genome Sequencing

We confirmed that *SRGAP2* was specifically duplicated in the human lineage by fluorescent in situ hybridization (FISH) by using a probe corresponding to the human *SRGAP2* (spanning exon 3, Table S1). We identified three map locations on chromosome 1 (1q32.1, 1q21.1, and 1p12), as compared to a

single chromosomal signal at 1q32.1 among other ape species (**Figure 2.1A**). An analysis of the segmental duplication content of 11 additional mammalian genomes (see Extended Experimental Procedures) showed no evidence of recent duplication in any lineage other than human and established 1q32.1 as the ancestral copy. FISH analysis of cell lines derived from humans of diverse ethnicity consistently showed a pattern of three distinct signals on each chromosome 1 corresponding to paralogs that were all incompletely sequenced in the human reference genome (GRCh37/hg19).

We reasoned that the recent nature of the duplications resulted in high-identity duplications with little genetic variation. As a result, allelic and paralogous copies became difficult to disentangle during genome assembly [133]. To resolve the different genomic copies, we constructed a large-insert bacterial artificial chromosome (BAC) library from DNA derived from a complete hydatidiform mole (CHORI-17). Because a complete hydatidiform mole originates from the fertilization of an enucleated human oocyte with a single spermatozoon [131, 134], the corresponding DNA represents a haploid, as opposed to a diploid, equivalent of the human genome (**Figure 2.1B**). We leveraged the absence of allelic variation to unambiguously distinguish *SRGAP2* copies despite their high sequence identity. We selected clones with homology to *SRGAP2* and subjected them to high-quality capillary-based sequencing, requiring >99.9% sequence identity of the overlap between sequenced inserts for assembly into the same contig.

We generated three sequence contigs corresponding to *SRGAP2* paralogs at 1q32.1 (562,704 bp; *SRGAP2A*), 1q21.1 (441,682 bp; *SRGAP2B*), and 1p12 (603,678 bp; *SRGAP2C*) (**Figure 2.1C**), generating over 1.6 Mbp of high-quality finished sequence. During the assembly process, we identified a single BAC clone (CH17-248H7) that harbored sequence for a *SRGAP2* paralog (exons 7–9) but did not share >99.9% identity with any of the three contigs, suggesting that a fourth *SRGAP2* duplicate existed (*SRGAP2D*). Upon this discovery, we repeated our FISH analysis using a probe mapping across exon 1 of *SRGAP2* and discovered four distinct signals on chromosome 1, with *SRGAP2D* mapping proximally to *SRGAP2B* on chromosome 1q21.1 (Figure S1 available online, Table S1). The absence of this signal from

the initial FISH assay (**Figure 2.1A**) suggested that a genomic region containing exon 3 was deleted from *SRGAP2D*.

The new local assemblies resolved the sequence and structure of three copies, adding 379,665 bp of new sequence completely absent from the human reference, including 40,233 bp within the ancestral *SRGAP2A* (**Figure 2.1C**). Additionally, we discovered 559,693 bp of sequence mapped incorrectly in orientation or chromosomal location within the human reference. Combined, we added or corrected more than 0.4% of the human chromosome 1 euchromatic sequence [135]. All finished sequence data, as well as the new human genome assemblies, have been deposited into GenBank and will be integrated into subsequent human genome reference assemblies (see Extended Experimental Procedures for accession numbers).

Comparisons between the three sequence contigs revealed large, interspersed segmental duplications of high-sequence identity (99%–99.5%) that were incomplete with respect to the ancestral locus (**Table 2.1**). We determined that the original duplication event (258,245 bp) encompassed the promoter, other cis regulatory elements, and the first nine exons of the 22-exon ancestral *SRGAP2A* (**Figure 2.2A**). Clusters of Alu repeat elements mapped precisely at the boundaries of this duplicated segment (Figure S2), confirming previous observations that Alu repeats are strongly associated with primate genomic duplications [136, 137]. A larger, secondary duplication event (>515 kbp) was shared between the *SRGAP2B* (1q21.1) and *SRGAP2C* (1p12) loci and included the entirety of the original duplication, although the *SRGAP2B* locus was subjected to subsequent larger deletions (102.6 and 49.0 kbp) upstream of the gene (Figure S2). Using multicolor FISH assays, we determined that the ancestral *SRGAP2A* paralog at 1q32.1 is transcribed toward the telomere, whereas the duplicate paralogs *SRGAP2B* and *SRGAP2C* are oriented such that gene transcription would proceed toward the centromere (Figure S1).

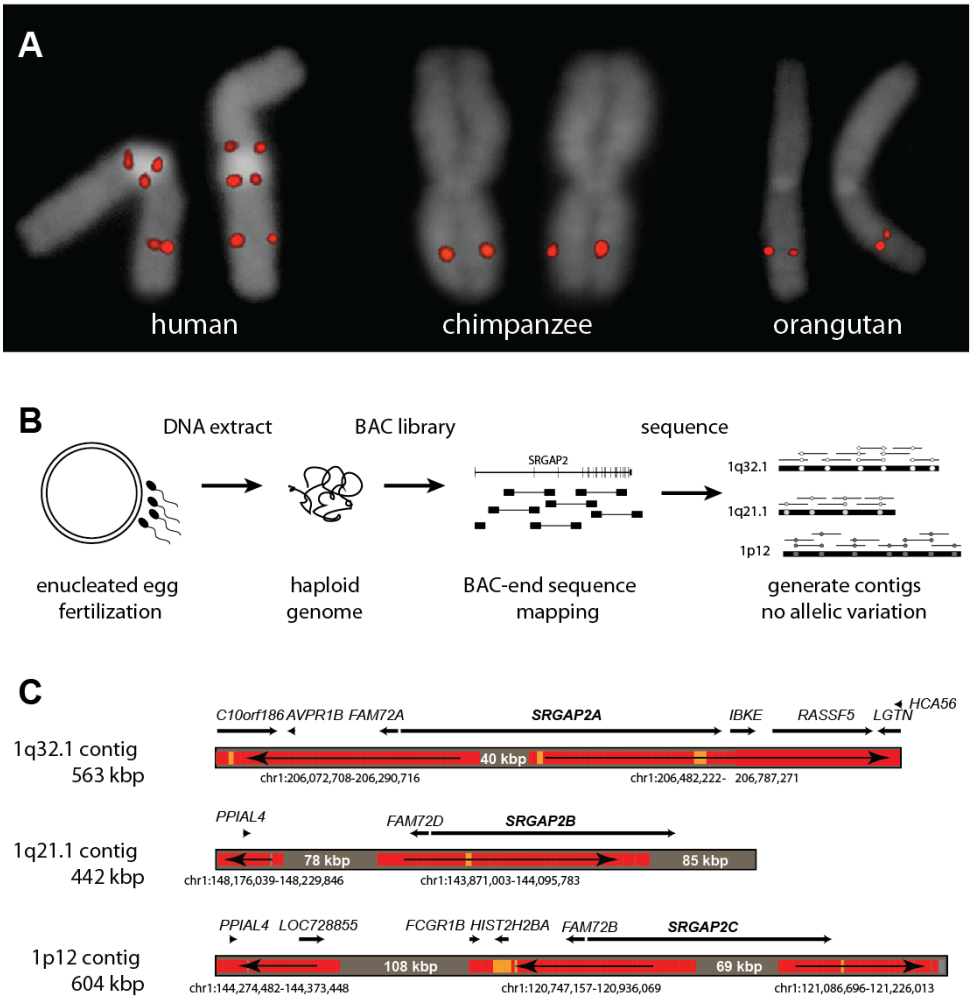


Figure 1. Genomic Characterization and Sequence Resolution of *SRGAP2* Loci. (A) FISH analysis shows three distinct copies of *SRGAP2* on metaphase human chromosome 1, compared to a single copy in chimpanzee and orangutan (see Figure 2A for location of FISH probe; Figure S1 and Table S1 for details of additional FISH assays). (B) *SRGAP2* genomic loci were sequenced and assembled using a BAC library (CHORI-17) created from human haploid genomic source material (complete hydatidiform mole). The absence of allelic variation allowed paralogous sequences to be resolved with high confidence based on near-perfect sequence identity overlap (>99.9%). (C) Regions highly identical to the reference genome (GRCh37/hg19) are colored in red (identity = 99.8%–100%) and orange (99.6%–99.8%), whereas regions completely absent from the current assembly are shaded gray (with region sizes indicated). Arrows show the orientation of the reference genome sequence with respect to the contigs (e.g., a left directional arrow indicates the reverse strand). Overall, this indicates that even the ancestral (*SRGAP2A*) gene locus was missing sequence data, misassembled, and incorrectly orientated over 400 kbp of the current high-quality reference assembly. Genomic coordinates correspond to the representative human reference region with corresponding genes within these regions mapped along each contig.

Table 2.1. Percent Sequence Divergence of *SRGAP2* Paralogs.				
	SRGAP2A	*SRGAP2B*	*SRGAP2C*	*SRGAP2D*
SRGAP2A	-	0.015	0.016	0.069
SRGAP2B	0.525	-	0.014	0.038
SRGAP2C	0.584	0.451	-	0.065
SRGAP2D	0.452	0.136	0.400	-

Kimura two-parameter model of genetic distance computed as base substitutions per site (left diagonal) and standard error (right diagonal). Pairwise distances are computed across 244,200 sites representing the complete shared genomic region between *SRGAP2* paralogs. Values for *SRGAP2D* represent pairwise distances computed across 9,541 sites. As a reference, the genetic distance between *SRGAP2A* and its chimpanzee ortholog locus is 0.852 ± 0.019, whereas that of chimpanzee to human paralogs *SRGAP2B* and *SRGAP2C* (0.901 ± 0.019 and 0.960 ± 0.020) are consistent with the accelerated mutation rate for these chromosomal regions.

2.3.2 Evolutionary History of *SRGAP2*

To reconstruct the evolution of the duplication events, we generated a multiple-sequence alignment for a 244.2 kbp region that is shared among the three contigs by using orthologous sequence from chimpanzee (build GGSC 2.1.3/panTro3) and orangutan (build WUGSC 2.0.2/ponAbe2) as outgroups (**Figure 2.2B**). Phylogenetic analysis provides strong support (>99%) for distinct duplication events occurring at different time points during human evolution. Notably, we find that the duplicated sequences have evolved much more rapidly (Tajima's relative rate test; p = 0.00001–0.0249) than the ancestral 1q32.1 locus (p = 0.5345). Mutation rates are known to vary significantly depending on chromosomal location and context [74]. Based on analysis of unique orthologous sequence adjacent to the *SRGAP2C* duplicate region, we determined that the distal 1p12 region shows a 20%–46% higher substitution rate when compared to 1q32.1. If we adjust for this difference, calibrating to the estimated 1q32.1 substitution rate, we predict that the initial duplication occurred ~3.4 mya and that the secondary event occurred ~2.4 mya. We note that estimates of molecular divergence between the paralogs are robust (e.g., 0.451 ± 0.014% substitutions per site between the *SRGAP2B* and *SRGAP2C* loci), owing to the large number of substitutions discovered in the high-quality sequence used in these comparisons (Table 1). Some uncertainty in our estimates comes from our correction factor for differing substitution rates, but most uncertainty arises from ambiguity in the evolutionary timing of the divergence of chimpanzee and human (estimated at ~6 mya) [132]. If we take into account previously reported human and chimpanzee

divergence times ranging from ~5–7 mya, based on fossil records [138-140] as well as recent genetic analyses [132], we estimate that the initial duplication occurred 2.8–3.9 mya, followed by the secondary duplication at 2.0–2.8 mya. We also performed phylogenetic analysis of the 9,541 bp region shared among the *SRGAP2A–C* paralogs and the incompletely sequenced *SRGAP2D* and determined that this copy was derived from the *SRGAP2B* locus ~1 mya (0.4–1.3 mya assuming a 6 mya divergence time for human and chimpanzee). Using comparative FISH analysis and probes mapping outside of the original duplication (**Figure 2.2C**), we determined the likely order of events: the ancestral *SRGAP2A* region duplicated first to 1q21.1 (*SRGAP2B*), and later the 1q21.1 copy duplicated to chromosome 1p12 (*SRGAP2C*) and within 1q21.1 (*SRGAP2D*).

Based on the gene structure of the ancestral *SRGAP2A*, sequence analysis predicts that *SRGAP2B* and *SRGAP2C* would produce transcripts maintaining an open-reading frame (ORF). These two duplicate copies, however, are predicted to produce a truncated form of *SRGAP2*, carrying nearly the entire F-BAR domain that lacks the final 49 amino acids (**Figure 2.2A**) [127]. The ancestral *SRGAP2* protein sequence is highly constrained based on our analysis of ten mammalian lineages. We find only a single amino acid change between human and mouse and no changes among nonhuman primates within the first nine exons of the *SRGAP2* orthologs. This is in stark contrast to the duplicate copies, which diverged from ancestral *SRGAP2A* less than 4 mya but have accumulated as many as seven amino acid replacements (five for *SRGAP2C* and two for *SRGAP2B*), compared to one synonymous change.

We used a likelihood ratio test [141] to evaluate differences in selective pressures acting on *SRGAP2* and found that the best model of selection allows an increased nonsynonymous (dN) to synonymous substitution (dS) ratio of the *SRGAP2* duplicate paralogs while maintaining purifying selection in the remaining lineages (compared with the fixed dN/dS model, $p = 1.32 \times 10^{-11}$, Table S2). This difference is consistent with an increased substitution rate of the 1q21.1 and 1p12 chromosomal regions and a relaxation of selective pressure on the duplicate copies. Overall, this mechanism provides a means for rapid evolutionary change of an otherwise constrained developmental gene [8].

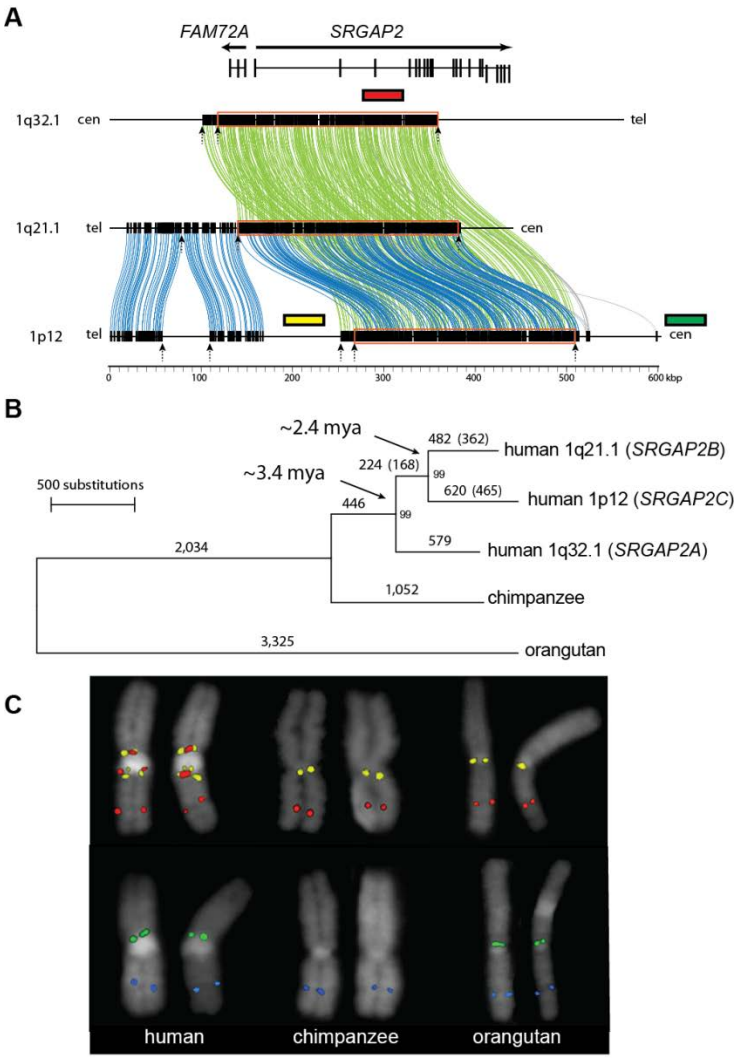


Figure 2.2. Evolutionary Characterization of *SRGAP2* Duplications. (A) A depiction of the gene structure of *SRGAP2* with respect to the three assembled contigs. Homologous segments are shown using Miropeats [142] where green lines indicate nearly identical segments (s = 1,000) shared between *SRGAP2A* and the duplicate *SRGAP2* paralogs, and the blue lines delineate the larger (>515 kbp) extent of homology between *SRGAP2B* and *SRGAP2C*. The 244.2 kbp genomic region shared among all three contigs is highlighted (red box) with clusters of Alu repeats at the breakpoints (arrows). Also see Figure S2 for a detailed representation of Alu elements and segmental duplications across duplicated regions. (B) An unrooted neighbor-joining tree was constructed based on a 244.2 kbp multiple sequence alignment of the three loci. Both 1p12 and 1q21.1 branches show accelerated rates of substitution ($p = 0.00001$ and $p = 0.0249$; Tajima's relative rate test). The actual (no parentheses) and adjusted (parentheses) number of substitutions for locus-specific acceleration is indicated above each branch along with the bootstrap support at each node. We estimate the timing assuming chimpanzee and human diverged 6 mya. Also see Table S2 for molecular evolution of the shared *SRGAP2* coding regions. (C) FISH experiments on metaphase human chromosome 1, as well as the orthologous chimpanzee and orangutan chromosomes, were performed to discern the order of duplication events. Locations of probes with respect to the contigs are shown in (A). A probe (yellow) targeting the sequence adjacent to the original *SRGAP2* duplicate region hybridizes to 1q21.1 in chimpanzee and orangutan, suggesting that the original *SRGAP2* duplicate paralog maps to the region homologous with nonhuman primate 1q21.1. A probe (green) targeting the unique sequence on the p arm of chromosome 1 proximal to *SRGAP2C* hybridizes to the chromosome 1p arm in orangutan, refuting the possibility that *SRGAP2C* moved to the p arm via a simple pericentromeric inversion [30] and distinguishing the p arm from the genomic region at 1q21.1 where the original *SRGAP2* duplicate paralog maps. A probe (blue) was used to distinguish the chromosome 1q arm.

2.3.3 *SRGAP2* mRNA Expression and Paralog Gene Structure

We assayed for expression of *SRGAP2* paralogs by designing specific reverse-transcriptase PCR (RT-PCR) assays that distinguish the duplicate paralogs from the ancestral copy based on the presence of a duplicate-specific 3' untranslated region (UTR) present in a previously sequenced cDNA mapping to the *SRGAP2C* locus (GenBank accession BC112927). A total of 96 transcripts were sequenced from RNA derived from the SH-SY5Y neuronal cell line, pooled fetal brain, a single fetal brain, and a single adult brain (**Figure 2.3A** and Table S3). Comparing genomic and cDNA sequences, we assigned the transcripts to their respective copies and identified the exon/intron structure, alternative splice forms, as well as fixed and polymorphic paralog-specific variants (PSVs) (**Figure 2.3B**). We found that all *SRGAP2* paralogs are transcribed, though at different relative proportions. We identified transcripts containing exons 1 through 9 that map specifically to *SRGAP2C* (n = 47) and *SRGAP2B* or *SRGAP2D* (n = 4). Using capillary sequencing of these transcripts and focusing our analysis on two fixed PSVs, we show that relative expression of the *SRGAP2B/D* transcript is markedly low (14%–25% and 30%–72% of *SRGAP2C* transcript abundance in fetal and adult brain, respectively) (Figure S3). The most abundant duplicate transcript is expressed from *SRGAP2C* and predicts an ORF that would encode a truncated *SRGAP2* protein (458 amino acids), including a partial F-BAR domain [127] and seven unique residues at the carboxyl terminus.

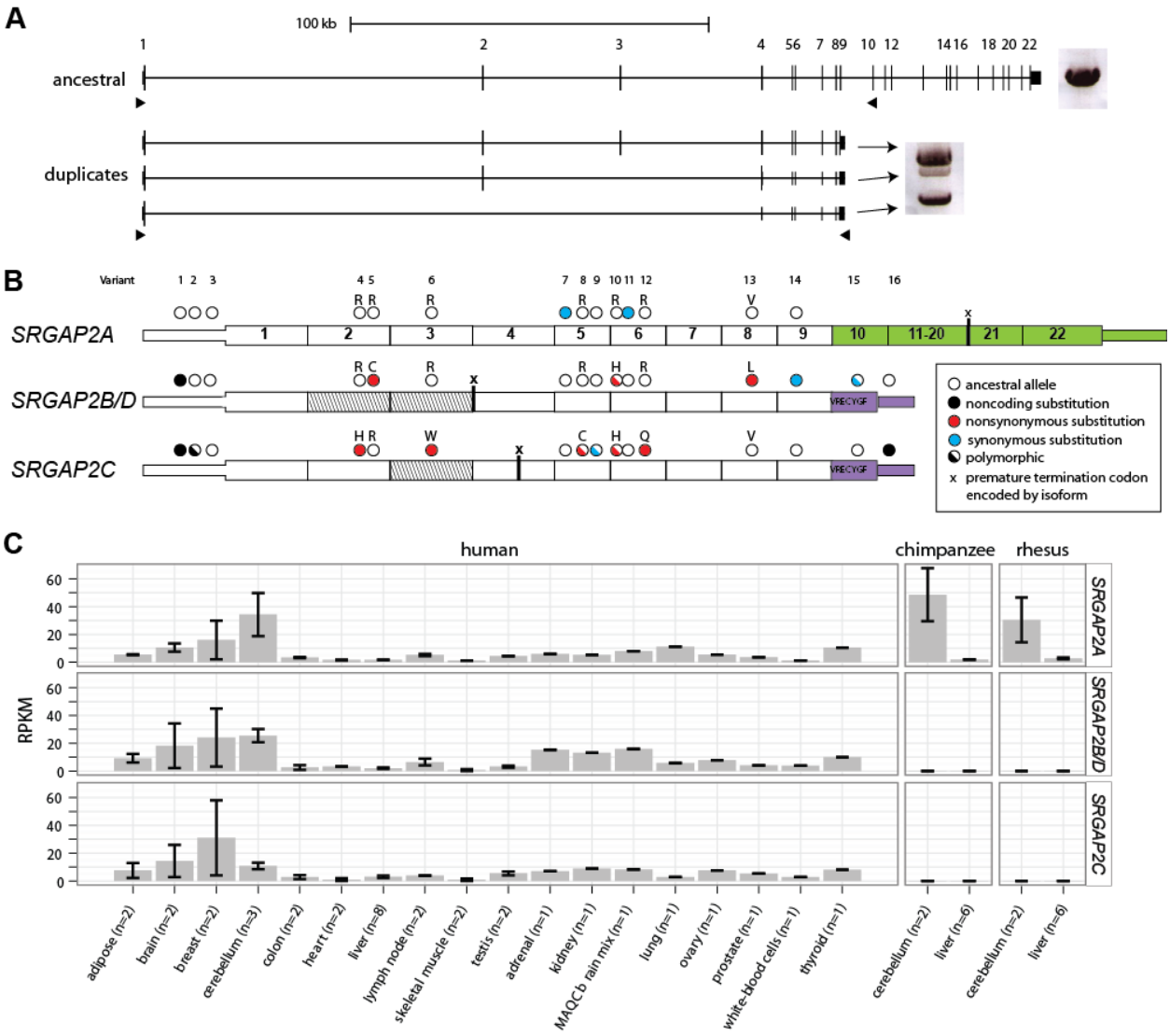


Figure 2.3. Paralog-Specific *SRGAP2* Gene Expression. (A) Long-range RT-PCR products from pooled fetal brain RNA are shown next to the gene models. A single band was amplified from the ancestral paralog, whereas three bands were amplified from duplicate paralogs by using primers (black triangles) designed to target alternative isoforms. Ninety-six cDNA transcripts were cloned and sequenced. (B) Fixed paralog-specific variants were used to assign transcripts to respective genomic loci, allowing both polymorphic and fixed putative amino acid changes to be deduced. Exonic sequence specific to the ancestral copy (*SRGAP2A*; green) and the duplicate loci (*SRGAP2B/C/D*; purple) are shown. The locations of stop codons encoded by isoforms missing exons are represented with an ‘‘x.’’ Exons missing from transcripts are indicated (diagonal lines) and likely correspond to the genomic deletion within *SRGAP2D* in the case of the exon 2 and 3 deleted isoform. (C) Paralog-specific expression profiling was performed by using RNA-Seq data mapped to unique sequence identifiers. The mean RPKM of each *SRGAP2* paralog is shown for a variety of primate tissue types, with error bars representing ±SEM. The specificity of next-generation sequence data and the determination of single base-pair differences between the copies were necessary to tease apart the expression profiles of these virtually identical copies. Chimpanzee and macaque RNA-Seq data affirm the specificity of this assay. Also see Figure S3 and Table S3 for additional expression results.

We also observed numerous transcripts and putative splice isoforms that are unlikely to encode functional proteins. The most abundant of these map to *SRGAP2B/D* (n = 31) missing exons 2 and 3 and result in a transcript that would encode a premature truncated protein (23 amino acids). These transcripts are consistent with our genomic sequence analysis, indicating that *SRGAP2D* has acquired a 115 kbp deletion including exons 2 and 3 (described later). Moreover, our analysis suggests that this transcript may be subjected to nonsense-mediated decay.

Using diagnostic PSVs to distinguish copies, we interrogated the expression of specific *SRGAP2* paralogs in various human and nonhuman primate tissues using RT-PCR (Figure S3) and RNA-Seq data (**Figure 2.3C**). The tissue profile reveals that the paralogs show similar broad patterns of expression, including expression in the developing human fetal brain concurrently with *SRGAP2A*. We observe higher expression in multiple regions of the human cortex and cerebellum when compared to other tissues including lung, kidney, and testis. As expected, we did not detect expression of the duplicate copies in any of the nonhuman-primate-derived tissues.

2.3.4 *SRGAP2* Copy Number Variation

Because *SRGAP2* has been shown to play an important role in brain development, we initially focused on the ancestral *SRGAP2A* gene by examining a large cohort of pediatric cases with developmental delay (1,602 individuals tested using a quantitative PCR [qPCR]) assay specifically targeting *SRGAP2A* and 15,767 individuals reported by Cooper et al. [143]) for potential copy number variation. We identified six large (>1 Mbp) copy number variants (CNVs), including three deletions of the ancestral 1q32.1 region (**Table 2.2**), with no similar large CNVs observed among 10,123 controls. Because the CNVs are large and encompass multiple candidate genes, this observation does not prove pathogenicity of dosage imbalance of *SRGAP2A*. We note, however, that in one patient the proximal breakpoint maps within the first intron of *SRGAP2A*, potentially disrupting the gene (Figure S4 and Table S4). The patient is a ten-year-old child with a history of seizures, attention deficit disorder, and learning disabilities. An MRI of this patient also indicates several brain malformations, including hypoplasia of the

posterior body of the corpus callosum. Recently, a de novo-balanced translocation t(1;9)(q32;q13) breaking within intron six of *SRGAP2A* was reported in a five-year-old girl who was diagnosed with West syndrome and exhibited epileptic seizures, intellectual disability, cortical atrophy, and a thin corpus callosum [144]. Although much more work needs to be done, the neurological phenotypes observed in these two cases are consistent with neuronal migration deficits implicated in forms of developmental delay and epileptic encephalopathies [144].

Table 2.2. *SRGAP2A* and *SRGAP2C* Copy Number Variation Genotyping of Cases and Controls

Genotype Method	Size Resolution	Cohort[a]	Total Genotyped	Deletions	Duplications
SRGAP2A[b]					
Custom array CGH platforms	>50 kbp	intellectual disability (Signature Genomics) [143]	15,767	3	3
SNP arrays	>50 kbp	controls [143]	8,329	none	none
qPCR[c]	n/a	intellectual disability	1,602	none	none
		controls (NIMH and ClinSeq)	1,794	none	none
Illumina sequencing	>100 kbp	controls (1000 Genomes Project)	661	none	none
SRGAP2C[d]					
qPCR[e]	n/a	intellectual disability	1,602	none	1
		controls (NIMH and ClinSeq[g])	1,794	none	1
Custom array CGH[f]	>300 kbp	idiopathic autism (SSC)	2,294	none	2
		familial autism (AGRE)	579	none	none
		controls (NIMH and ClinSeq[g])	580	none	none
Illumina sequencing	>100 kbp	controls (1000 Genomes Project)	661	none	none

All detected deletions and duplications of *SRGAP2A* and *SRGAP2C* were >1 Mbp and include additional genes. Data from the Cooper et al. study [143] could not be used to assess CNVs for *SRGAP2C*, as there was insufficient probe coverage on the microarrays used in those studies. See also Figure S4 and Table S4 for details of CNV breakpoints, phenotypes, and inheritance status.

[a]Abbreviations: SSC, Simons Simplex Collection [145]; AGRE, Autism Genetic Resource Exchange [146]; NIMH, National Institute of Mental Health (https://www.nimhgenetics.org/available_data/controls/); ClinSeq, Clinical Sequencing Pilot Project [147].

[b]Cases, n = 17,369; Controls, n = 10,784.

[c]The assay targeted intron 11 of *SRGAP2A*.

[d]Cases, n = 4,475; Controls, n = 2,662.

[e]Two assays were used targeting introns 6 and 7 of *SRGAP2C*, respectively.

[f]Using probes targeting the chromosome 1p11.2 region proximal to *SRGAP2C*, we identified duplications and determined that a subset of them extended into *SRGAP2C* by using qPCR assays. Notably, all duplications of *SRGAP2C* identified from the qPCR assay alone extended into the 1p11.2 proximal region and would have been detected using this same method.

[g]ClinSeq controls (n = 373) were screened both with array CGH and qPCR assays.

We next focused on assessing copy number variation of each *SRGAP2* paralog in the human population. This is particularly challenging because most recently duplicated genes are typically highly copy number polymorphic [121, 148], and experimental assays for accurately predicting copy number are problematic. For this purpose, we took advantage of diagnostic singly unique nucleotide (SUN) identifiers (n = 3,535) determined using our high-quality sequence of the three loci (see above). We mapped genome-sequencing data from 661 human individuals corresponding to 14 populations (1000 Genomes Project) and estimated the diploid copy number for each paralog by measuring read depth to these SUNs (**Figure 2.4A**) [121].

We find that both the ancestral *SRGAP2A* and the derived *SRGAP2C* copy are fixed at diploid copy number two across all humans assayed. In contrast, the *SRGAP2B* and *SRGAP2D* copies varied from 0–4 copies among the individuals tested (**Figures 2.4B–2.4C**). Importantly, we identified three individuals that are homozygously deleted for *SRGAP2B*. Notably, we also identified normal individuals that were homozygously deleted for *SRGAP2D*, the granddaughter copy with an acquired internal deletion of exons 2 and 3 (see Figure S5 for characterization of this internal deletion). We prepared cDNA from lymphoblastoid cells corresponding to one of these *SRGAP2B*-deletion homozygotes and observed no full-length *SRGAP2B* transcript by RT-PCR, which is in contrast to samples carrying the paralog (Figure S3). Because the frequency of homozygotes is consistent with Hardy-Weinberg Equilibrium expectation and these individuals are representatives of the sample populations, the discovery of *SRGAP2B*-homozygous deletions in a ‘‘normal’’ population argues against a critical functional role of this copy in brain development. We additionally applied our method to 34 nonhuman primates and the Denisova and Neanderthal genomes [149, 150] and found that, consistent with our sequence-based estimations of the timing of the duplication events, *SRGAP2B*, *SRGAP2C*, and *SRGAP2D* copies are absent from all assayed nonhuman great apes yet are present in both the Neanderthal and Denisova genomes. We conclude that no new *SRGAP2* duplications have occurred since *Homo sapiens* and *Homo neanderthalensis* diverged about 1 mya.

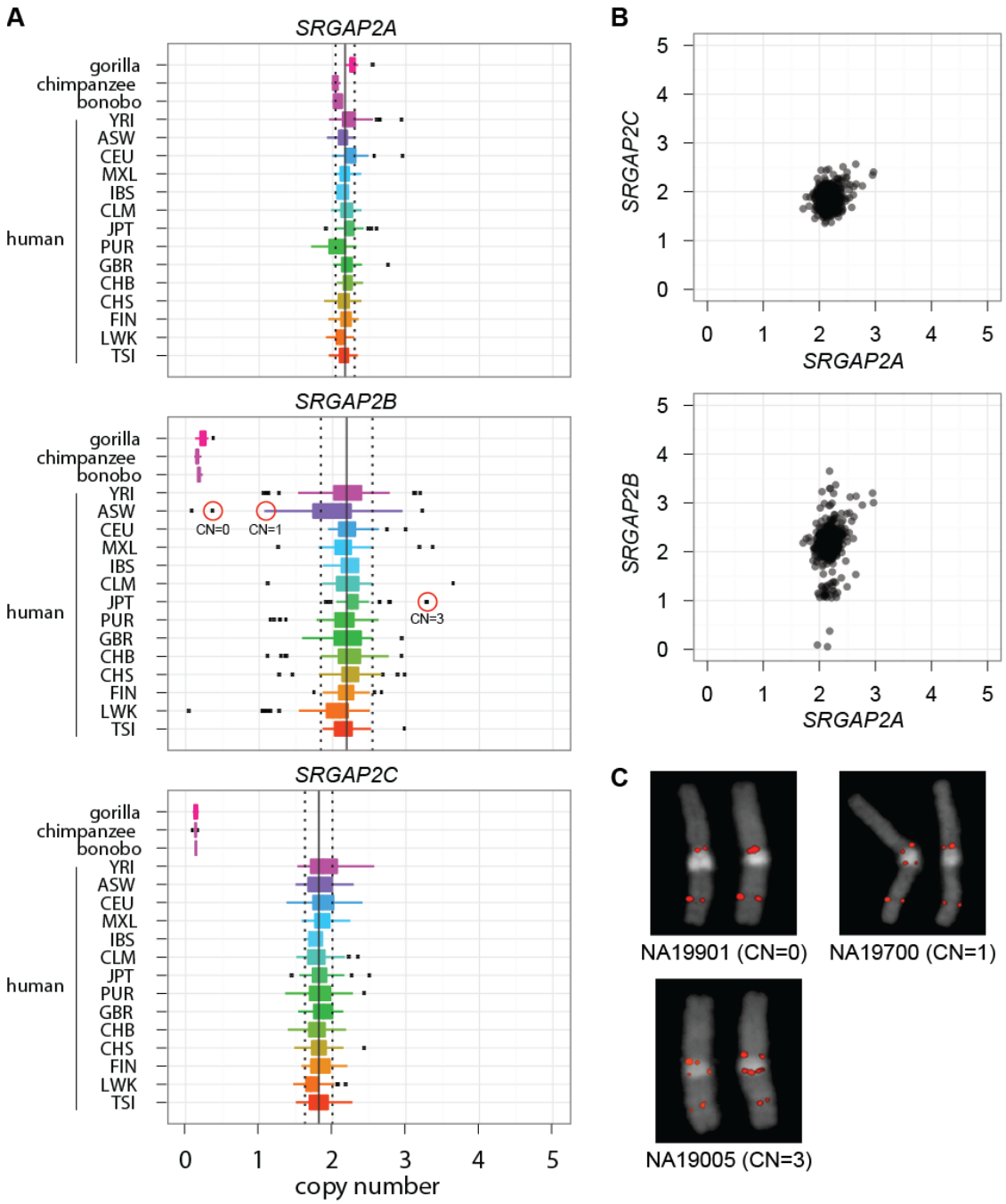


Figure 2.4. *SRGAP2* Copy Number Diversity in Human Populations. (A) Diploid copy number estimates of *SRGAP2* paralogs for 661 sequenced human genomes from 14 distinct populations (1000 Genomes Project, Appendix A) and from nonhuman primates are graphically represented as boxplots (the box contains the 25th to 75th percentile of the distribution, and the black dots represent outliers). The solid vertical and dashed lines represent the median copy number estimate and ±SD, respectively, of each paralog across all populations. (B) *SRGAP2A* and *SRGAP2C* paralogs clearly are fixed at a copy number of two, while *SRGAP2B* is polymorphic showing four distinct copy number states. Note, we also detect polymorphism for *SRGAP2D* and have identified individuals homozygously deleted for this paralog. (C) FISH validation of three HapMap individuals genotyped for *SRGAP2B* (circled in red in part [A]). All samples falling at the lower and upper tails of copy number distributions for all three paralogs were experimentally genotyped by using a paralog-specific qPCR assay; in all cases, *SRGAP2A* and *SRGAP2C* were validated as diploid copy number two. Also refer to Figure S5.

Although it is common to observe a functional progenitor duplicated gene fixed in copy number, the discovery that a gene as recently evolved as *SRGAP2C* is fixed at a diploid copy number state is striking. When compared to the 23 genes duplicated specifically in the human lineage, we previously found that *SRGAP2* is among the six least copy number polymorphic gene families under a naive analysis that does not distinguish paralogs [121]. When we extend this analysis to human-specific duplicates for which complete sequence is available and limit our analysis solely to those genes (n = 23), we find that *SRGAP2C* is the least copy number variable gene duplicate. Using qPCR assays that specifically assess copy number variation of *SRGAP2C*, we investigated this experimentally and found one individual harboring an ~1 Mbp duplication containing numerous genes in an additional set of 1,794 controls (**Table 2.2** and Figure S4). Applying this same assay to patients with intellectual disability and/or autism spectrum disorder (n = 4,475), we identified three additional individuals carrying large duplications of this locus. Strikingly, in our cumulative analysis of 7,137 individuals (cases and controls), we detected no deletions of *SRGAP2C*. In total, our combined analyses indicate that both *SRGAP2A* and *SRGAP2C* copies are nearly fixed at a copy number of two in all human populations assayed, with rare deletions and duplications observed only in cases with intellectual disability for *SRGAP2A* ($p = 0.055$, Fisher's exact test) and rare duplications observed at a frequency of ~0.06% for *SRGAP2C*.

2.4 Discussion

SRGAP2 has been highly conserved over mammalian evolution, and human is the only lineage wherein gene duplications have occurred. Our analysis indicates that the duplications spread across 80 Mbp of chromosome 1 at a time corresponding to the transition from *Australopithecus* to *Homo* (**Figure 2.5**). This included an initial large interspersed duplication (258 kbp) from chromosome 1q32.1 to 1q21.1, creating *SRGAP2B* ~3.4 mya. The initial duplication was followed by larger (>515 kbp), secondary duplications of the 1q21.1 locus, creating *SRGAP2C* and *SRGAP2D* (~2.4 and 1 mya, respectively). Consistent with these timing estimates, archaic *Homo* species, including Neanderthal and Denisova, carry these *SRGAP2* paralogs (Figure S5). It is intriguing that the general timing of the potentially functional

copies, *SRGAP2B* and *SRGAP2C*, corresponds to the emergence of the genus *Homo* from *Australopithecus* (2–3 mya). This period of human evolution has been associated with the expansion of the neocortex and the use of stone tools, as well as dramatic changes in behavior and culture [79].

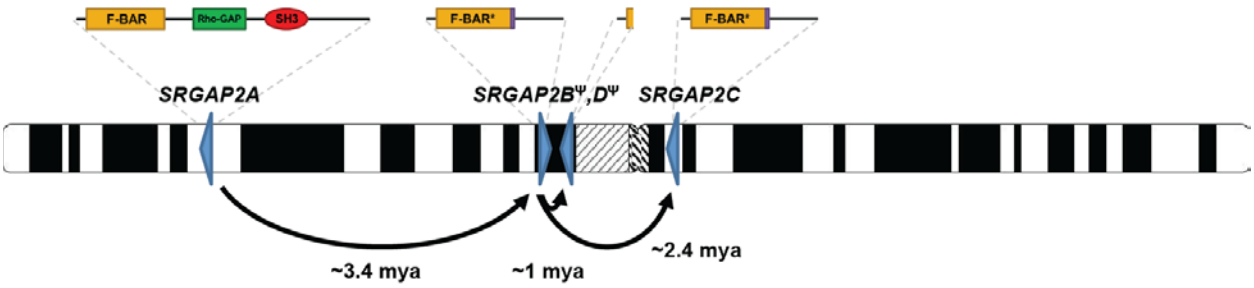


Figure 2.5. Model for *SRGAP2* Evolution. Schematic depicts location and orientation (blue triangles) of *SRGAP2* paralogs on human chromosome 1 with putative protein products indicated above each based on cDNA sequencing. Asterisks indicate a 49 amino acid truncation of the F-BAR domain. Arrows trace the evolutionary history of *SRGAP2* duplication events. Copy number polymorphism and expression analyses suggest both paralogs at 1q21.1 (*SRGAP2B* and *SRGAP2D*) are pseudogenes, whereas the 1q32.1 (*SRGAP2A*) and 1p12 (*SRGAP2C*) paralogs are likely to encode functional proteins.

Our analysis provides insight into one mechanism by which gene duplicates evolve. We find that the initial genomic duplication of *SRGAP2* was incomplete, encompassing the promoter and first nine exons of a 22 exon gene. Because *SRGAP2* has been shown to homodimerize via its F-BAR domain [127], we propose that incomplete segmental duplication of the gene ~3.4 mya created an antagonistic functional state. In fact, functional evidence suggests that these partial *SRGAP2* copies produce protein with a nearly complete F-BAR domain but are missing other functional domains. These copies also heterodimerize with the full-length *SRGAP2*, creating a de facto dominant negative interaction equivalent to a knockdown of the ancestral copy [113]. The large size of the segmental duplication included the putative *cis* regulatory machinery of this gene and ensured that the duplicate genes would be developmentally coexpressed with the parental copy. Experimental analyses indicate [113, 127] that if the segmental duplication had been slightly larger (i.e., included exon 10), such antagonism would not be possible.

The incomplete nature of the segmental duplication was, therefore, ideal to establish this new function by virtue of its structure, which arose at the time of its ‘‘birth.’’ This model of gene duplication that involves an ‘‘instantaneous’’ dominant negative function at birth stands in stark contrast to the favored model that involves duplication of a complete gene followed by the gradual accumulation of adaptive mutational events leading toward subfunctionalization or neofunctionalization [18]. We suggest that *SRGAP2C* ultimately assumed the antagonistic function of the *SRGAP2B* duplicate, which shows evidence of pseudogenization in contemporary humans. Although all four *SRGAP2* paralogs show evidence of transcription, it is unlikely that the two copies at 1q21.1 are now functional for several reasons. *SRGAP2B* has a markedly reduced expression in human brain compared to *SRGAP2C*. Likewise, the transcripts produced by *SRGAP2D* lack two internal exons, leading to a premature termination codon. Therefore, this copy is unlikely to produce a functional protein. Both *SRGAP2B* and *SRGAP2D* are highly copy number polymorphic, with normal individuals identified that completely lack these paralogs. This argues that if there is a phenotypic consequence to their complete deletion, it is likely to be relatively minor.

In stark contrast, both the *SRGAP2A* (progenitor) and *SRGAP2C* (granddaughter) paralogs are nearly fixed at a diploid state based on our analysis of 28,153 and 7,137 human DNA samples, respectively. If we assume that the original *SRGAP2B* function was acquired by *SRGAP2C*, there is a possibility that both paralogs were functional at some point during human evolution. It is interesting that the comparison of the >515 kbp of duplicated sequence shared between *SRGAP2B* and *SRGAP2C* indicates that *SRGAP2B* has been subjected to large upstream deletions (103 kbp and 49 kbp in size), whereas *SRGAP2C* has not. Thus, the genomic instability of the *SRGAP2B* locus and its reduced expression in the contemporary human brain imply that the 1q21.1 locus may have been a suboptimal environment for gene transcription. The duplication event that yielded *SRGAP2C* ~2.4 mya may have provided a means of escape, transporting this truncated gene to a much more stable genomic environment for robust, long-term expression. One cannot, of course, definitively exclude the possibility that *SRGAP2B* and *SRGAP2D* transcripts may still confer some function [113], perhaps via transcript

regulation, but the finding of apparently normal individuals completely missing these duplicate copies would suggest that they are not critical for normal development.

We have identified larger deletions of the ancestral locus, *SRGAP2A*, only among children with developmental delay. Although the deletion intervals are large and other genes contributing to the disease phenotype cannot be excluded at this time, the absence of structural variation in the normal population and the discovery of a de novo translocation [144], as well as a second patient with a duplication breakpoint mapping within *SRGAP2*, provide some evidence of its role in brain development. In this light, the fixation of the duplicated *SRGAP2C* is especially noteworthy. *SRGAP2C* was found to be the least copy number polymorphic of all human-specific duplicate genes, despite the fact that it is embedded in a complex region prone to nonallelic homologous recombination. Our data, thus, point to two functional *SRGAP2* copies at 1p12 and 1q32.1, consistent with experimental characterization [113]. Based on these data, we propose more systematic screening of these genes for mutations in children with developmental delay and brain malformations that include West Syndrome, agenesis of the corpus callosum, and epileptic encephalopathies. This will be particularly challenging because most commercial SNP microarrays have failed to include probes from these duplicated regions, and reads from next-generation sequencing platforms are typically too short to assign to a specific paralog [151]. Nevertheless, final proof of the functional significance of these genes will rest on the discovery of disruptive mutations associated with human phenotypes.

Finally, we emphasize that much of the genomic sequence corresponding to the ancestral and duplicate gene copies was missing or misassembled in the current human reference genome. In this study, we sequenced, corrected, and annotated ~0.4% of the euchromatin of chromosome 1 more than 6 years after the ‘‘finished’’ human genome was declared [152]. This was possible because the clone-based resource we developed using a complete hydatidiform mole essentially provides a haploid version of the human genome. Because this resource is devoid of allelic variation, we can rapidly distinguish even highly identical duplicate genes, thus providing a clear path forward for the characterization of other complex duplicated regions. It is worthwhile noting that we ensured the hydatidiform mole primary cell

line (CHM1hTERT) we used did not contain any large CNVs that could confound our analysis [134]. It is especially intriguing that *SRGAP2* is only one of several human-specific duplicate genes missing or incompletely assembled in the human genome [121]. A number of remaining genes (e.g., *GPRIN2*, *GTF2IRD2*, and *HYDIN*) in this category have been implicated in neurodevelopment, neurite outgrowth, and behavior [65, 128, 153]. Additionally, human-specific protein-coding genes derived de novo from noncoding DNA merit further exploration [154]. We propose that these uncharacterized human-specific genes constitute important pieces in the puzzle underlying the genetic basis of human brain evolution.

2.5 Experimental Procedures

2.5.1 Fluorescent In Situ Hybridization

Metaphase spreads were prepared from lymphoblastoid human cell lines (NA12878, NA19317, NA20334, NA19901, NA19700, and NA19005; Coriell Cell Repository, Camden, NJ), a chimpanzee cell line (Douglas, provided by Dr. Mariano Rocchi), and an orangutan cell line (PR01109, a.k.a. Susie; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using fosmid clones (Extended Experimental Procedures) as described previously [155].

2.5.2 Cloning Using a Complete Hydatidiform Mole Library

A large-insert BAC library (CHORI-17) was generated from a well-characterized complete hydatidiform mole primary cell culture (CHM1hTERT) using a modified protocol [156] (http://bacpac.chori.org/library.php?id=231). To ensure the quality of CHM1hTERT, a karyotype analysis and extensive SNP genotyping with 1,494 SNP markers [134] and array comparative genomic hybridization (CGH) using the NimbleGen 2.1 M whole-genome array were performed. We generated paired-end sequences (n = 169,022) by using Sanger dideoxy methods, and we mapped sequence reads to the human reference genome. This provided a haplotype-resolved tiling path of clones for selection and sequencing [157].

2.5.3 Sequencing and Assembly

We selected BAC clones with at least one sequenced end mapping to a *SRGAP2* region in the human reference genome and completely sequenced and assembled the insert (see Extended Experimental Procedures for detailed clone order, sequence assembly, and annotation). Inserts overlapping with >99.9% sequence identity were assembled into distinct contigs corresponding to *SRGAP2* loci at 1q32.1, 1q21.1, and 1p12.

2.5.4 Phylogenetic Analysis

We created a 244.2 kbp multiple sequence alignment from three completely sequenced *SRGAP2* genomic loci (ClustalW [158]) and constructed an unrooted phylogenetic tree (MEGA [159]) by using the neighbor-joining method [160] with the complete-deletion option. Genetic distances were computed with the Kimura two-parameter method [161] with standard error estimates (an interior branch test of phylogeny [162, 163]; n = 500 bootstrap replicates). For the incompletely sequenced *SRGAP2D* paralog and the 1p12 chromosomal distal region, we created phylogenetic trees by using a 9.5 kbp and 50 kbp multiple species alignment, respectively (see Extended Experimental Procedures for details). The orthologous *SRGAP2* exons were extracted from different mammalian reference genomes without segmental duplications and were used to test various models of selection using a maximum-likelihood framework (codemL; PAML statistical software package [141]).

2.5.5 *SRGAP2* Transcript Analysis

Total RNA was isolated using Trizol reagent (Invitrogen) and the RNeasy Mini Kit (QIAGEN) from SH-SY5Y neuronal cell line. Total RNA was analyzed from human fetal brain (collected from spontaneously aborted fetuses, 50–60 pooled samples, 20–33 weeks of development; ClonTech S2437) as well as a single fetal (R1244035, BioChain) and adult brain sample (M1234035, BioChain) (see Extended Experimental Procedures for details regarding RT-PCR, cDNA cloning, and sequencing). We also analyzed RNA-Seq data from 17 different human tissues (Illumina's Human BodyMap 2.0), seven human cell lines [164], and both chimpanzee and macaque cerebellum and liver tissues [165]. Briefly, RNA-Seq

data sets were mapped to the human reference genome (NCBI36/hg18) and our described *SRGAP2* contigs. Expression levels for specific paralogs were calculated in units of RPKM (reads per kilobase of exon model per million mapped reads) [166] with transcribed PSVs, which allowed RNA-Seq data to be unambiguously assigned to a specific paralog.

2.5.6 Paralog-Specific Copy Number Genotyping

CNVs in cases with intellectual disability and controls for *SRGAP2A* were identified from previously published array CGH data and SNP microarray data, respectively [143]. Copy number estimates of specific *SRGAP2* paralogs by using SUNs were determined using previously described methods [121]. Custom qPCR assays were performed in triplicate using variants specific to each *SRGAP2* paralogous locus (see Extended Experimental Procedures for a description of variant detection and primer sequences). Validations of deletions and duplications, as well as identification of CNVs in the autism cohorts and some controls, were performed by array CGH using custom microarrays (Agilent) and a HapMap individual (NA18507) as a reference.

2.6 Notes

Accession Numbers

The GenBank accession numbers for the sequences reported in this paper are listed in Table S5.

Supplemental Information

Supplemental Information includes Extended Experimental Procedures, five figures, and six tables and can be found in Appendix A.

Acknowledgements

We thank B. Coe for assistance in CNV analysis and the 1000 Genomes Project for access to sequence data of the *SRGAP2* loci. For DNA samples used in paralog-specific CNV screening and detailed phenotypic information of patients, we would like to thank C. Romano, M. Fichera, J. Gecz, B. de Vries, R. Bernier, the Simons Foundation, Autism Speaks, the National Institute of Mental Health, and the ClinSeq Project. We acknowledge C. Baker, L. Vives, and J. Huddleston for technical assistance, T. Brown for manuscript editing, and the laboratory of S. Fields for use of their Roche LC480. We also thank J. Akey, T. Marques-Bonet, A. Andres, S. Girirajan, and K. Meltz Steinberg for helpful discussion, as well as the laboratory of F. Polleux for comments and kindly sharing human RNA samples for expression studies. The BAC clones from the complete hydatidiform mole were derived from a cell line created by U. Surti. M.Y.D. is supported by U.S. National Institutes of Health (NIH) Ruth L. Kirchstein

National Research Service Award (NRSA) Fellowship (1F32HD071698-01). X.N. is supported by an NIH NRSA Genome Training Grant to the University of Washington (2T32HG000035-16). P.H.S. is a Howard Hughes Medical Institute International Student Research Fellow. This work was supported by NIH Grants HG002385 and GM058815. E.E.E. is an investigator of the Howard Hughes Medical Institute. J.A.R. and L.G.S. are employees of Signature Genomic Laboratories, a subsidiary of PerkinElmer, Inc. E.E.E. is on the scientific advisory boards for Pacific Biosciences, Inc. and SynapDx Corp.

3. Rapid and Accurate Large-Scale Genotyping of Duplicated Genes and Discovery of Interlocus Gene Conversions

This chapter has been published: Nuttle X, Huddeston J, O'Roak BJ, Antonacci F, Fichera M, Romano C, Shendure J, Eichler EE. *Nat. Methods* **10**, 903-909 (2013).

I designed the study with Jay Shendure and Evan E. Eichler, designed molecular inversion probes with Brian J. O'Roak, performed capture experiments, wrote analysis software, analyzed data, and wrote the paper with Evan E. Eichler.

3.1 Summary

Over 900 genes have been annotated within duplicated regions of the human genome, yet their functions and potential roles in disease remain largely unknown. One major obstacle has been the inability to accurately and comprehensively assay genetic variation for these genes in a high-throughput manner. We developed a sequencing-based method for rapid and high-throughput genotyping of duplicated genes using molecular inversion probes designed to target unique paralogous sequence variants. We applied this method to genotype all members of two gene families, *SRGAP2* and *RH*, among a diversity panel of 1,056 humans. The approach could accurately distinguish copy number in paralogs having up to ~99.6% sequence identity, identify small gene-disruptive deletions, detect single-nucleotide variants, define breakpoints of unequal crossover and discover regions of interlocus gene conversion. The ability to rapidly and accurately genotype multiple gene families in thousands of individuals at low cost enables the development of genome-wide gene conversion maps and 'unlocks' many previously inaccessible duplicated genes for association with human traits.

3.2 Introduction

Duplicated genes are important contributors to genetic variation [121, 167-169], evolutionary adaptation [1, 4, 73, 76] and human disease [44, 53, 170, 171]. Despite this, most individual duplicated genes remain poorly characterized at the genetic level [98] because of high sequence identity [98, 123], extensive copy-number polymorphism [121, 167-169], missing sequencing data [98] and low correlation with flanking single-nucleotide polymorphisms [125, 167, 172]. As a result, these genes and regions have

often been excluded from genetic analyses [151, 173], or contradictory associations with disease have been reported [174, 175].

Several different technologies have been applied to assay copy number for such genes [176]. Both quantitative real-time PCR (qPCR) and the paralog ratio test [177], which uses PCR product specificity to distinguish copies, are labor intensive, requiring the design and testing of multiple primers. Multiplex ligation-dependent probe amplification (MLPA) [178] and multiplex amplification and probe hybridization (MAPH) [179] allow for copy-number analysis at up to 50 loci simultaneously, but they cannot be applied to genotype many gene families at high spatial resolution in a single reaction. Array comparative genomic hybridization (CGH) lacks paralog specificity and can access only a fraction of duplicated genes, typically where the number of duplicated copies is low [167, 172]. Finally, mapping whole-genome sequencing (WGS) data to singly unique nucleotide (SUN) identifiers that tag a particular paralog and analyzing the read depth [121, 180] has yielded genome-wide paralog-specific copy-number estimates. However, the sensitivity of this approach depends on genome sequencing coverage, and sequencing remains a costly proposition that cannot be applied to thousands of samples in a laboratory setting.

Here we used molecular inversion probes (MIPs), short oligonucleotides designed to capture targeted genomic regions [181-184], together with massively parallel DNA sequencing for genotyping duplicated genes. We evaluated this method by examining *SRGAP2* and *RH* genetic variation in 1,056 individuals and explored its potential application to the discovery of interlocus gene-conversion events in humans. The method scaled well to thousands of samples and yielded accurate, paralog-specific sequence and copy-number genotypes at a low cost.

3.3 Results

3.3.1 Genotyping Strategy

Our approach leverages SUN variants, fixed paralogous sequence variants that uniquely tag a specific paralog and distinguish it from all other copies [121]. We systematically designed MIPs to

hybridize to sequences that are identical between paralogs flanking SUNs across the length of the duplicated segment (**Figure 3.1, i,ii** and Appendix B) and additional MIP assays targeting exons in the paralogs to assay coding-sequence variation. The probability of an individual MIP capturing sequence from a particular paralog is a function of its copy number relative to the copy number of related paralogs (**Figure 3.1, iii**). Massively parallel sequencing of amplified capture products allows simultaneous quantification of sequences derived from each paralog (**Figure 3.1, iv**) and detection of sequence-level genetic variation. We selected two gene families to demonstrate the proof of principle of our approach and to assess its power to discover novel genetic variation in duplicated regions: *SRGAP2* [98], a highly identical (>99%) human-specific gene family, and *RH*, a clinically relevant blood-antigen gene family that has been extensively characterized for common copy-number polymorphism [185], rearrangement breakpoints [186] and interlocus gene conversion [187] in the human population.

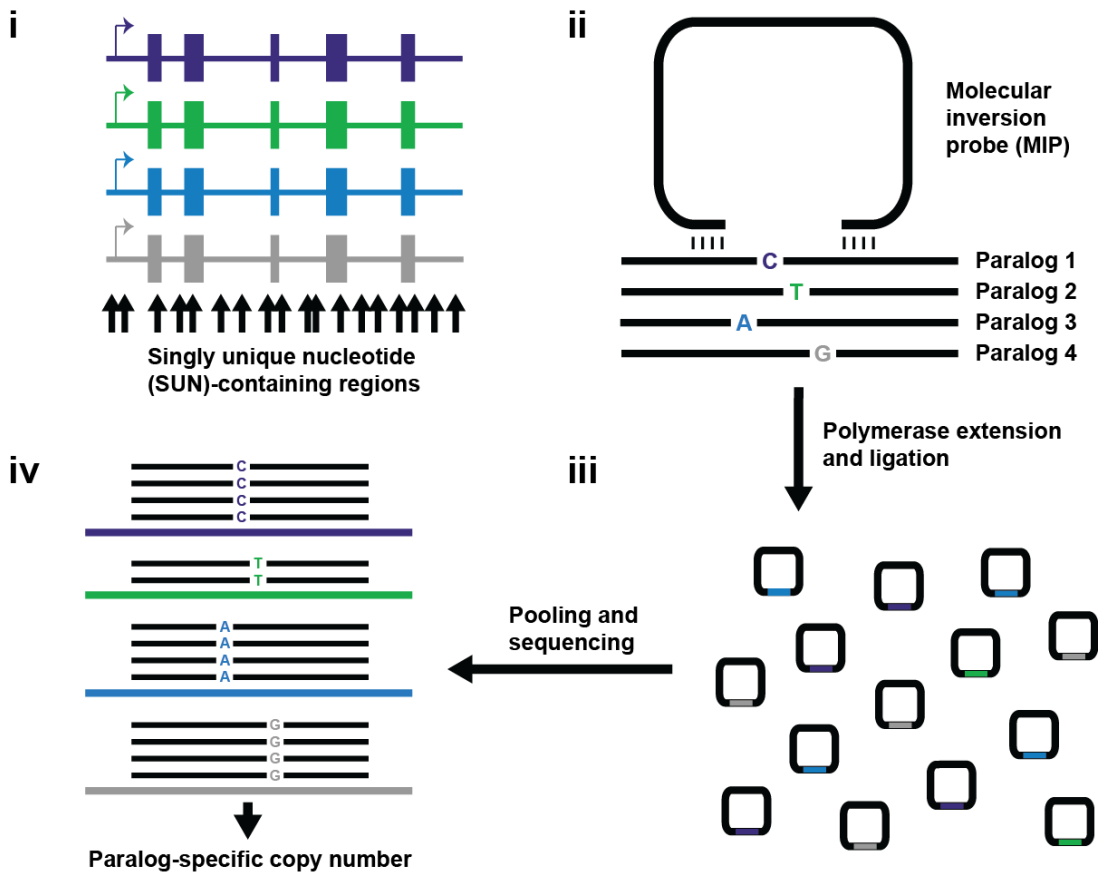


Figure 3.1. MIP Copy-Number Genotyping Assay for Duplicated Genes. (i) 112-nucleotide (nt) regions (black arrows) containing sequence variants that uniquely distinguish one paralog (potential SUNs) are identified through alignment of genomic sequence. (ii) 70-nt MIPs used for copy-number genotyping have 16- to 24-nt hybridization arms complementary to sequence flanking SUN-containing regions. Several such MIPs are designed, collectively spanning the spatial extent of duplicated genic sequence. (iii) DNA polymerase extension and ligation incorporates SUN-containing sequences into covalently closed circular molecules, which are then barcoded, pooled and sequenced. (iv) Reads are mapped to reference sequences for each paralog, and paralog-specific read counts for each MIP are quantified. A genotyping program infers paralog-specific copy number from these counts. The schematic shows counts consistent with a deletion of paralog 2 (green).

3.3.2 Copy-Number and Sequence Genotyping

For *SRGAP2*, we designed a total of 142 MIPs targeted to sites corresponding to potential SUNs (Supplementary Tables 1 and 2 and Appendix B) that could reliably differentiate *SRGAP2* paralogs. Forty of these MIP targets harbor nucleotide differences that distinguish all four *SRGAP2* paralogs from one another, 28 distinguish two *SRGAP2* paralogs from the other two paralogs, and the remaining 74 distinguish a single *SRGAP2* paralog from the remaining three. We initially used these MIPs to genotype 48 individuals for which orthogonal *SRGAP2* copy-number data were generated or were available from WGS data (Appendix B), array CGH and/or FISH. All captured sequences from a given DNA sample were barcoded, pooled with those from other samples and sequenced using HiSeq or MiSeq (Illumina) to an approximate coverage of 350 reads per MIP per individual. For each individual, paralog-specific read counts served as a proxy for copy number for each *SRGAP2* gene. We developed a maximum-likelihood approach using paralog-specific read-count data to generate *SRGAP2* paralog-specific copy-number calls across the spatial extent of duplicated *SRGAP2* sequence (Appendix B). Incorporating data from all MIPs overwhelms noisy signals from poorly performing individual MIPs. Plotting MIP data for 90 high-performing copy-number MIPs (Appendix B and Supplementary Figs. 1 and 2) alongside FISH data for three representative individuals highlights the precision with which MIP genotyping detected known duplications and deletions of *SRGAP2B* and *SRGAP2D* (**Figure 3.2**).

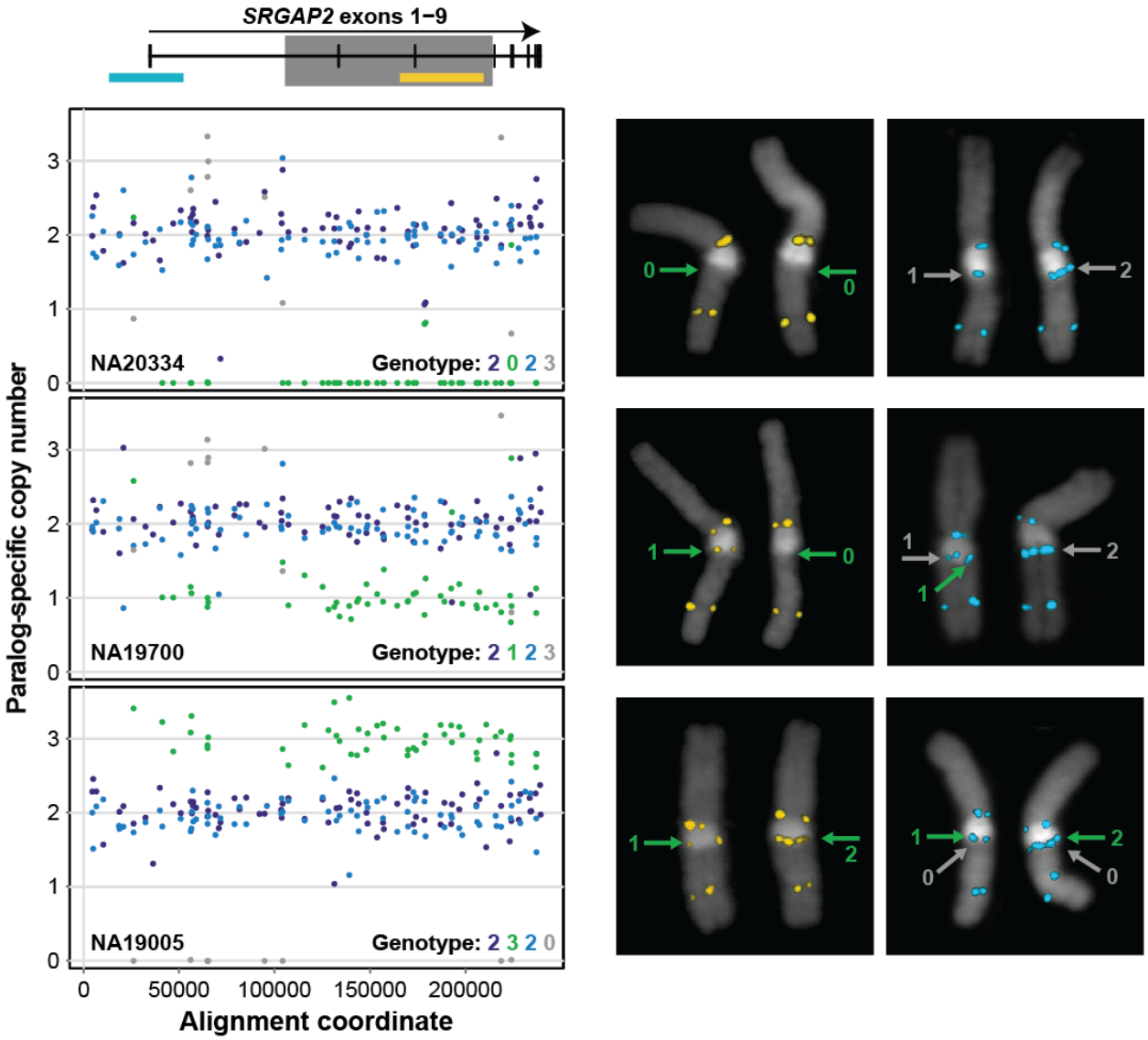


Figure 3.2 Accuracy of Paralog-Specific Copy-Number Genotyping. MIPs (142) and FISH were used for genotyping *SRGAP2* copy number in the HapMap individuals NA20334, NA19700 and NA19005. Exon locations are plotted relative to the FISH probes (cyan and yellow rectangles) and MIP data below. The gray box indicates the region deleted in *SRGAP2D*. Paralog-specific copy-number estimates are shown for 90 high-performing MIPs across ~240 kbp of aligned *SRGAP2* genomic sequence. Each point indicates a paralog-specific copy-number estimate (purple, *SRGAP2A*; green, *SRGAP2B*; blue, *SRGAP2C*; gray, *SRGAP2D*), calculated as the product of the paralog-specific read-count frequency for a particular MIP and the aggregate estimated *SRGAP2* copy number at the corresponding locus. Shown are homozygous and heterozygous deletions and a duplication of *SRGAP2B* as well as duplications and a homozygous deletion of *SRGAP2D*. Right, FISH data validate the MIP-based paralog-specific copy-number genotypes for these individuals. Colored numbers indicate copy number of *SRGAP2B* or *SRGAP2D* for the adjacent chromosome. FISH data for NA20334 and NA19700 are consistent with either two or three diploid copies of the *SRGAP2D* paralog.

We found that 97.2% (35 of 36) of copy-number calls were concordant with FISH, 8 of 8 were consistent with array CGH data, and 91.5% (150 of 164) agreed with estimates made from WGS data (Supplementary Table 3). All inconsistencies involved genotyping results for the *SRGAP2D* pseudogene. This paralog is the shortest and most recently duplicated segment having ~99.6% identity to *SRGAP2B*. Low WGS coverage together with the paucity of *SRGAP2D* SUNs likely confounded sequencing-based copy-number estimates for *SRGAP2D*. To explore this possibility, we generated aggregate *SRGAP2* copy-number estimates from WGS data. These aggregate estimates are more accurate than corresponding paralog-specific estimates [121] because all reads mapping to *SRGAP2* (rather than just those mapping to SUN identifiers) inform this analysis. Notably, 13 of 14 aggregate *SRGAP2* copy-number estimates were consistent with MIP-based paralog-specific estimates rather than with corresponding WGS-based paralog-specific estimates in cases when these results disagreed. We extended our analysis to include 1,056 HapMap individuals, 73 of which we genotyped more than once using MIPs to examine the reproducibility of our approach. We found 99.5% (390 of 392) of replicate *SRGAP2* paralog-specific copy-number genotypes were concordant with initial MIP-based genotypes (Supplementary Table 4).

Our data allowed us to estimate allele frequencies for *SRGAP2* duplications and deletions in nine human populations (Supplementary Table 5). As expected from a previous analysis of WGS data [98], *SRGAP2B* and *SRGAP2D* showed evidence of complete loss or gain, ranging in copy number from 0 to 4 in the human population. In contrast, complete duplication or deletion of *SRGAP2A* or *SRGAP2C* was not observed, a result consistent with the notion that these two paralogs are functional copies. Our analysis of *SRGAP2B* and *SRGAP2D* copy-number variation suggests population stratification. Deletion of *SRGAP2B*, for example, is more common in populations of African descent than deletions of *SRGAP2D*, which segregate at higher frequencies in several out-of-Africa populations.

Unlike most other copy-number genotyping assays, MIPs also provide information on the sequence content of targeted regions [183, 184, 188]. We reasoned that in some cases, linkage of discovered single-nucleotide variants (SNVs) to a nearby paralog-distinguishing SUN would allow inference of the paralog of origin. We evaluated whether our method could accurately genotype such

SNVs by comparing MIP sequence data (Appendix B) with fosmid clone end-sequence data [157] and WGS data for NA18507, an individual previously sequenced to high coverage [189]. The WGS data validated 93.8% (15 of 16) of our genotype calls (Supplementary Table 6), including a heterozygous nonsynonymous variant. Fosmid end-sequence data including a putative variant site were available in only three cases, but each validated the SNV identified from MIP data. Thus, our method can successfully detect SNVs within highly identical duplicated sequence and in some cases accurately assign them to specific paralogs.

3.3.3 Internal *SRGAP2* Deletion and Duplication Discovery

We applied our MIP-based method to two individuals having array CGH profiles showing complex structural variation in *SRGAP2* [98]. MIP genotyping correctly identified large *SRGAP2C* and *SRGAP2A* events discovered via array CGH and resolved the internal deletions as specifically affecting *SRGAP2C*, removing exon 2 and inducing a frameshift (**Figure 3.3**). MIP-based genotyping of 1,056 HapMap individuals indicated that this deletion is segregating at low frequency (<3%) exclusively in populations with some European ancestry. In addition to this *SRGAP2C* deletion, we identified seven other additional internal deletion and duplication events in HapMap individuals ranging in size from 1.5 kilobase pairs (kbp) to 144 kbp and assigned them to specific *SRGAP2* paralogs (Supplementary Table 5 and Supplementary Fig. 3). These structural variants included three distinct exon-overlapping events in *SRGAP2B* and an intronic duplication in *SRGAP2A*.

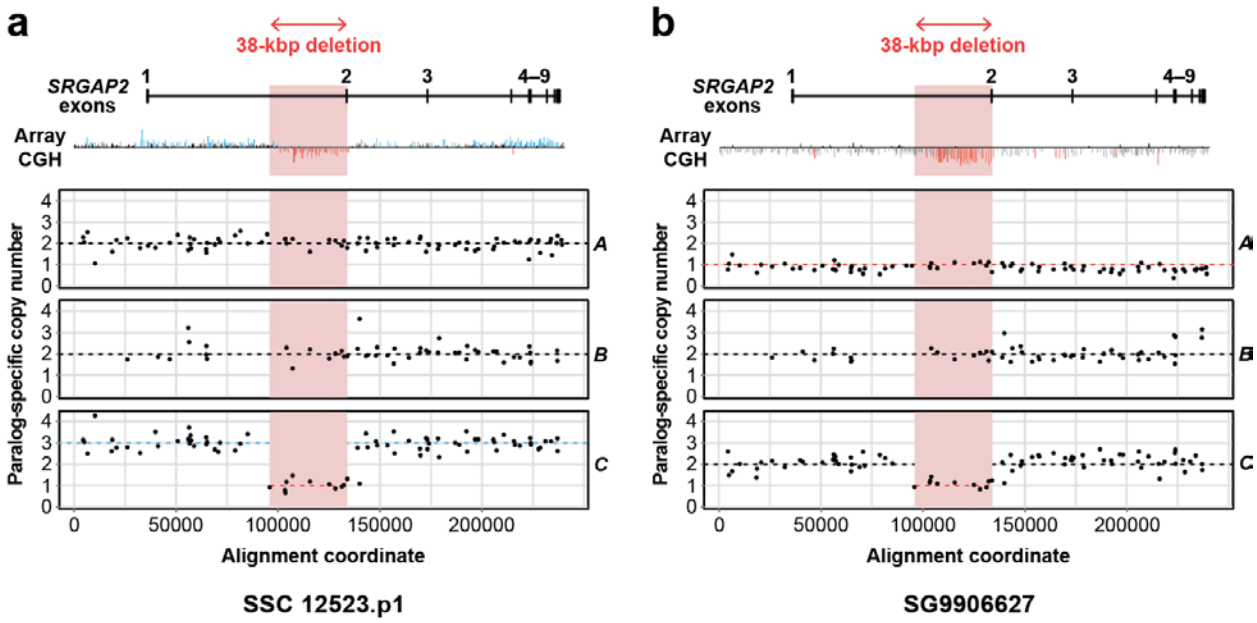


Figure 3.3. Resolution of Complex Structural Variation in *SRGAP2*. (a) The array CGH profile for *SRGAP2* loci predicts a gain and an interstitial loss for a patient with autism but cannot distinguish which paralogs the variation affects. The MIP copy-number assay predicts two copies for *A*, *B* and *D* (not shown) but duplication of a copy of *C* having an ~38-kbp internal deletion containing exon 2. Dashed lines indicate paralog-specific copy-number calls from the automated caller. (b) Similar analysis of a patient with developmental delay shows that the individual is diploid for *B* and *D* (not shown) but has lost a copy of *A* (the ancestral locus) and carries the internal deletion for *C*.

3.3.4 *RH* Gene Conversion, Copy Number and Breakpoint Resolution

To assess the applicability of our method to assaying interlocus gene conversion and resolving breakpoints associated with nonallelic homologous recombination (NAHR), we applied our MIP genotyping method to *RHD* and *RHCE*—sites of known gene conversion and unequal crossover with clinical relevance for Rh antigen presentation. We reasoned that these two forms of mutation would generate characteristic sequence signatures with respect to SUN copy number. In the case of gene conversion, we would expect to observe a reciprocal copy-number shift at a pocket of homology with no difference in copy number of flanking regions. Gains would correspond to donors and losses to acceptors of gene conversion, allowing inference of the directionality of the event. In contrast, at a site of unequal crossover, a reciprocal SUN copy-number transition should be observed around the NAHR breakpoint.

We designed 39 MIPs targeting *RH* paralogs and flanking regions (**Figure 3.4a**) and included them in the same capture reactions as *SRGAP2* MIPs, which allowed us to simultaneously genotype the

same individuals described above for *RH*. Searching for reciprocal copy-number shifts, we observed seven distinct putative *RH* gene-conversion events, ranging in length from 1,709 base pairs (bp) to ~39 kbp (Supplementary Table 5 and Supplementary Fig. 4). Although we denote these events as gene conversions, other mutational mechanisms [190-192] may be responsible for the signatures we observed. Four events involved a transfer of genetic information from *RHCE* to *RHD*, four corresponded to polymorphic variants reported in the Blood Group Antigen Gene Mutation Database (dbRBC at the US National Center for Biotechnology Information), and four were supported by at least one observed instance of transmission from parent to child. The most common involved sequence transfer from *RHD* to *RHCE* at a known gene-conversion site including *RHD* exon 2 [193] and was confirmed by whole-genome and fosmid clone sequencing data [121] from an individual predicted from MIP data to be homozygous for this event (**Figure 3.4b**).

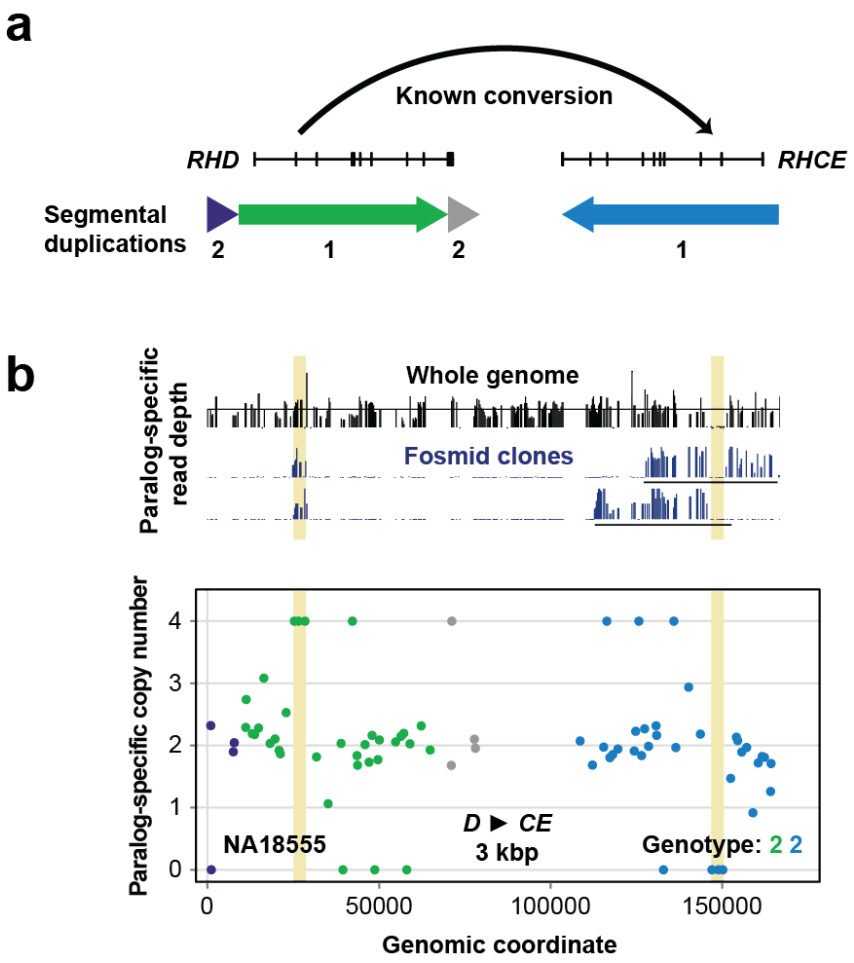


Figure 3.4 Detection of Gene Conversion at the *RH* Locus. (a) *RHD* and *RHCE* lie within an ~60-kbp segmental duplication (green and blue arrows) and frequently undergo interlocus gene-conversion events. (b) MIPs (39) were used for genotyping paralog-specific *RH* copy number in the HapMap individual NA18555. MIP data (bottom) are plotted relative to locations of *RH* exons and associated segmental duplications in a. Colors correspond to segmental duplications shown in a. A homozygous gene conversion from *RHD* to *RHCE* spanning at least ~3 kbp at a known conversion site including exon 2 is highlighted in yellow. We validated this homozygous conversion by mapping whole-genome and fosmid clone short-read sequence data from this individual to SUN identifiers and examining paralog-specific read depth.

Using our copy-number genotyping strategy, we identified known deletions and duplications in *RHD* associated with unequal crossover between flanking segmental duplications (**Figure 3.5a**). We found 97.6% (80 of 82) of our *RH* paralog-specific copy-number estimates agreed with those from WGS data (Supplementary Table 3). Reproducibility was lower for *RH* copy-number genotyping than for *SRGAP2* genotyping (Supplementary Table 4), as only 91.8% (180 of 196) of replicate MIP-based *RH* paralog-specific copy-number genotypes were concordant with initial genotypes. We calculated

logarithm-of-odds confidence scores for each of these genotypes (Appendix B) and observed that discordancies' scores fell at the low end of the score distribution (Supplementary Table 7 and Supplementary Fig. 5), a result suggesting that potential errors can be readily distinguished from high-confidence genotype calls. To attempt to refine NAHR-associated breakpoints, we looked for instances of a reciprocal paralog-specific copy-number transition within the segmental duplications flanking *RHD*. This approach allowed us to narrow breakpoint locations to within ~6-kbp windows (**Figure 3.5b**), regions previously found to contain *RHD* deletion breakpoints as determined by Sanger sequencing of spanning PCR products [186].

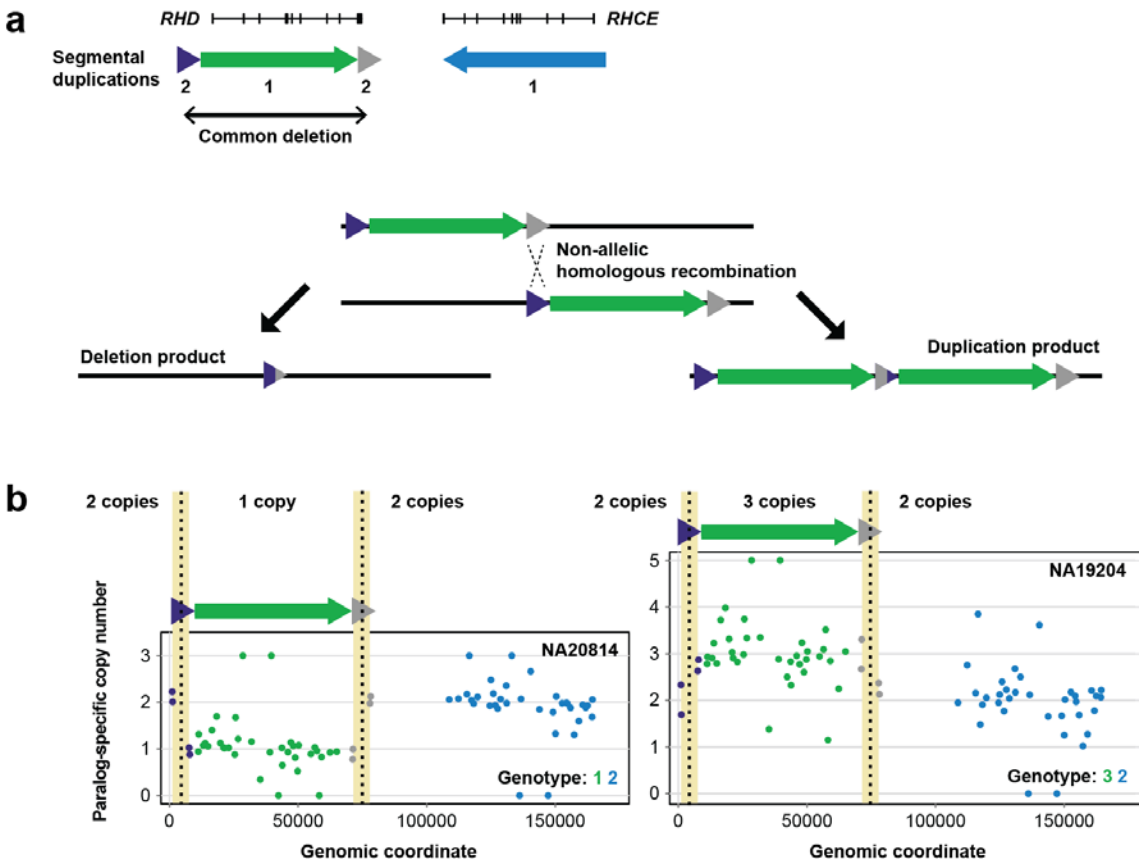


Figure 3.5 Resolution of Nonallelic Homologous Recombination (NAHR)-Associated *RHD* Deletion and Duplication Breakpoints. (a) An ~9-kbp segmental duplication (purple and gray triangles) flanks *RHD*. NAHR between these flanking sequences results in deletion and duplication of *RHD*. (b) Data from 39 *RH* MIPs for HapMap individuals NA20814 and NA19204 reveal copy-number variation at *RHD*. Note the signatures of NAHR in the four MIP data points corresponding to the *RHD*-flanking segmental duplications. These data refine the NAHR-associated breakpoints to ~6-kbp homologous genomic regions (highlighted in yellow) where *RHD* deletion breakpoints have been previously reported.

3.3.5 Discovery of Interlocus Gene Conversions in *SRGAP2*

Given the >99% sequence identity between *SRGAP2* paralogs, we reasoned that interlocus gene conversion or other mechanisms of nonreciprocal sequence transfer may have occurred at these loci and left signatures detectable using our MIP genotyping method. Analysis of the 1,056 HapMap individuals revealed ten such events ranging in size from 416 bp to 23 kbp (Supplementary Table 5 and Supplementary Fig. 6), collectively involving all four *SRGAP2* paralogs (**Figure 3.6a**). All paralogs except *SRGAP2C* were observed as putative gene-conversion donors. Unlike *RHD/CE*, these putative conversion events appear to have occurred over large genetic distances. For example, two distinct nonreciprocal exchanges of genetic information occur across the centromere between *SRGAP2A* and *SRGAP2C*—paralogs over 80 megabase pairs (Mbp) apart on chromosome 1 ([98]; **Figure 3.6b** and Supplementary Fig. 6).

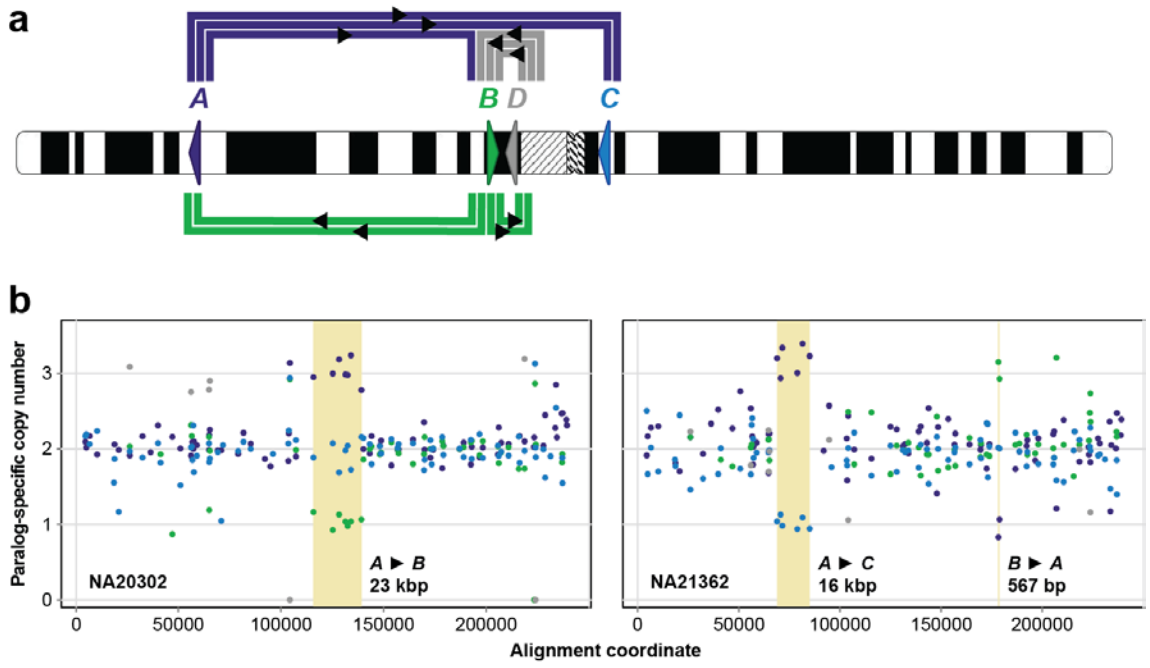


Figure 3.6 Extensive Interlocus Gene Conversion Between *SRGAP2* Paralogs. (a) Schematic denoting location and orientation (triangles) of *SRGAP2* paralogs on human chromosome 1. Thick colored lines connect *SRGAP2* paralogs exhibiting signatures of interlocus gene conversion in the MIP data. Line colors correspond to conversion donors, and each line corresponds to a distinct conversion event. (b) Examples of different interlocus gene-conversion events (highlighted in yellow). Reported sizes indicate the minimum length of the conversion event based on the MIP data, assuming a single conversion event underlies the conversion signature. All events shown, except for the *B*-to-*A* conversion revealed by two MIPs, were detected by the automated caller.

To corroborate these findings, we examined inheritance for putative gene-conversion events detected in members of HapMap trios. We observed at least one instance of transmission from parent to child for six distinct putative gene conversions, and no such events were inferred as *de novo*. We also validated one putative conversion using paralog-specific qPCR and array CGH. MIP data suggested a complete *SRGAP2D* duplication and a gene conversion resulting in replacement of *SRGAP2C* sequence with paralogous sequence in a patient with intellectual disability (Supplementary Fig. 7). If the MIP genotyping were accurate, results from *SRGAP2C*-specific qPCR using primers in the putative conversion region would be expected to signal a loss in *SRGAP2C* copy number, but results from array CGH would be expected to signal a slight gain in aggregate *SRGAP2* copy number over *SRGAP2* sequence shared with *SRGAP2D*. Performing the qPCR and array CGH experiments yielded precisely these results, providing additional support for the accuracy of our method and its applicability to detect novel signatures of interlocus gene conversion.

3.4 Discussion

What would be required to obtain the same volume of genotype information for an arbitrary gene family comparable to *SRGAP2* or *RH* using existing approaches? WGS offers great potential given its comprehensive nature [121, 180], but it remains prohibitively expensive for genotyping projects of even moderate size, especially given that accuracy demands high coverage. More scalable available targeted methods, on the other hand, provide limited genotyping power. PCR-based strategies for copy-number genotyping query at most a few sites per reaction because PCR multiplexes poorly [194, 195]. MLPA and MAPH allow for the simultaneous analysis of up to 50 loci, but even this greater scale of multiplexing cannot match the ability of our method to assay many gene families each at high spatial resolution. None of the targeted methods above provides exonic sequence information, and none has been successfully applied in large-scale studies of gene conversion. As our analyses demonstrate, genetic variation in duplicated genes exhibits considerable complexity. Any method for genotyping such genes should be developed with this consideration in mind.

Although we focused on *SRGAP2* and *RH*, our method will be useful for studying other duplicated genes that have proven difficult to genotype accurately, including *CCL3L1* [174, 175], beta-defensins [196, 197] and *C4* [198] (Appendix B). We provide programs to obtain genotypes with confidence scores from MIP-sequence data and to assist in the identification of informative sites from aligned sequences (https://github.com/xnuttle/mips_cnv_typer/). We also provide a complete list of ~3.8 million SUNs based on the current human reference genome (GRCh37) for use with other duplicated regions and gene families (Appendix B and http://eichlerlab.gs.washington.edu/mips_cnv_typer/). Although higher copy number and more polymorphic gene families will pose additional challenges, generating high-coverage sequence data precisely over the most informative sites promises to significantly improve our understanding of genetic variation of these complex regions of the genome.

Successful application of our method to a particular gene family of interest depends on several factors, including availability of accurate sequence, the number of paralogs, their sequence identity, their GC content [183, 184], their copy-number ranges and their sizes. First, optimal MIP design requires high-quality reference sequences for all family members, so gene families lacking complete sequence characterization (Appendix B and Supplementary Table 8) will be at least partially inaccessible using MIP-based genotyping. Second, some genetic variation must distinguish different paralogs from one another—our method cannot determine copy number when copies are identical at the genomic level (<1% of all paralogous sequences). Third, gene families with high numbers of paralogs, or with paralogs at high copy numbers showing a range of copy-number variation, pose several challenges for MIP-based genotyping. In general, as the number of distinct paralogs increases, fewer potential target regions will contain SUNs allowing discrimination of all paralogs; thus, more MIPs will need to be designed for copy-number genotyping. Furthermore, paralog-specific read-count frequencies become more difficult to confidently distinguish as the aggregate copy number for a gene family increases. This particular issue could be mitigated somewhat via the use of single-molecule MIPs to quantify individual capture events [199]. Accurate sequence genotyping also becomes more difficult as the aggregate copy number increases and the number of possible assignments of sequences to paralog copies grows.

Our method will facilitate efforts to map NAHR-associated structural variation breakpoints, which often occur in complex regions of segmental duplication. Identifying SUNs that discriminate the high-identity paralogs followed by MIP genotyping will provide sequence-level precision to determine the effect of such rearrangements on the genes embedded in such complex regions [200]. We anticipate MIP-based genotyping will also be very valuable for studies of interlocus gene conversion, providing an experimental platform for surveying the most highly identical paralogs where this mechanism frequently operates [201]. In this study, we provide evidence of conversion-like events between paralogs separated by more than 80 Mbp—a somewhat surprising finding given that conversion is thought to occur most frequently between high-identity segments in close proximity [202-204]. Most notably, our MIP-based method will encourage the inclusion of many previously intractable duplicated genes in future genetic analyses of human phenotypes. With accurate, scalable genotyping, we will be well positioned to assess the impacts of hundreds of these genes on human traits and disease.

Associated software, documentation and an example data set are freely available via GitHub at https://github.com/xnuttle/mips_cnv_typer/.

3.5 Notes

Methods

Methods and any associated references are available in Appendix B.

Accession Codes

US National Center for Biotechnology Information Sequence Read Archive: SRP027257.

Acknowledgements

We thank J. Kitzman for early ideas and enthusiasm for the project; P. Sudmant, E. Karakoc, F. Hormozdiari, B. Dumont and O. Penn for thoughtful discussion; L. Vives, K. Mohajeri and C. Lee for technical assistance; and T. Brown for assistance with manuscript preparation. X.N. is supported by a US National Science Foundation Graduate Research Fellowship under grant no. DGE-1256082. This work was supported by US National Institutes of Health grants HG004120 and HG002385 to E.E.E. E.E.E. is supported by the Howard Hughes Medical Institute.

Author Contributions

X.N., J.S. and E.E.E. designed the study. X.N. and B.J.O. designed the MIPs. X.N. performed capture experiments, wrote analysis software and analyzed data. F.A. performed FISH experiments. J.H. contributed to the analysis software, prepared it for public access and identified SUNs from the reference genome. M.F. and C.R. contributed to sample collection. X.N. and E.E.E. wrote the paper, with input and approval from all coauthors.

Competing Financial Interests

The authors declare competing financial interests: details are available in the online version of the paper.

4. Emergence of a *Homo sapiens*-Specific Gene Family and the Evolution of Autism Risk at Chromosome 16p11.2

This chapter has been submitted for publication: Nuttle X*, Giannuzzi G*, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, Benner C, Marchetto MC, Denman L, Harshman L, Barker C, Raja A, Penewit K, Janke N, Tang WJ, Ventura M, Antonacci F, Akey JM, Amemiya CT, Gage FH, Reymond A, Eichler EE. (2015).

*These authors contributed equally to this work. I performed experiments to sequence large-insert clones (together with several others), assembled the chromosome 16p11.2 contigs, performed segmental duplication analysis and structural comparisons between haplotypes, developed the evolutionary model, developed and performed paralog-specific copy number experiments and analyses, developed and performed RT-PCR experiments, discovered the *BOLA2-SMG1P* fusion transcripts, quantified susceptibility to recurrent microdeletion/microduplication of the autism critical region between species and different human haplotypes, refined rearrangement breakpoints, and wrote the paper with Evan E. Eichler.

4.1 Summary

Recurrent deletions and duplications at chromosome 16p11.2 account for ~1% of autism cases and are mediated by a complex set of segmental duplications. We reconstructed the evolutionary history of the locus by complete high-quality sequencing of orangutan, chimpanzee, and multiple human haplotypes. Using whole-genome sequencing data from 2,359 humans, 86 great apes, a Neanderthal and a Denisovan, we quantify the extent of copy number variation in 16p11.2 breakpoint regions in humans and identified *BOLA2* as a gene duplicated exclusively in *Homo sapiens*. We show that the duplication of *BOLA2* has led to a novel human-specific in-frame fusion transcript and that *BOLA2* copy number correlates with both RNA expression ($r = 0.36$) and protein level ($r = 0.65$), with the greatest expression difference between human and chimpanzee stem cells. We estimate that the *BOLA2*-containing segment duplicated ~282 thousand years ago (kya), one of the latest among a series of genomic changes that dramatically restructured the region during hominid evolution. All humans examined carry one or more copies of the duplication, which nearly fixed early in the human lineage—a pattern unlikely to have arisen so rapidly in the absence of selection ($p < 0.012$). Analyses of 151 patients carrying a 16p11.2 rearrangement showed that >96% of breakpoints occur within this *Homo sapiens*-specific duplication. We propose that predisposition to recurrent rearrangements associated with autism is linked to the emergence of a novel *BOLA2* gene family at the root of the *Homo sapiens* lineage ~300 kya.

4.2 Introduction

Recurrent deletions and duplications of an ~550 kbp region on human chromosome 16p11.2 are one of the most common risk factors for autism spectrum disorder [59, 60]. Collectively, these rearrangements account for ~1% of simplex autism cases but also associate with other conditions such as reduced verbal IQ, schizophrenia, and extremes of body mass index and head circumference [62, 67, 68, 205]. Despite widespread interest, the relationship between dosage imbalance of this region, neurodevelopment, and phenotypic variability is not well understood. Previous genetic studies of 16p11.2 in humans and model organisms have primarily focused on the 27 genes within unique sequence [206-210]. In contrast, the directly oriented duplication blocks that flank the region have not been well characterized.

These large segmental duplications (>500 kbp, Fig. S1) are important because they have been reported as sites of normal copy number variation [211], they promote the recurrent microdeletion and microduplication associated with disease [212] (Fig. S3), and they contain at least three duplicated genes potentially relevant to aspects of chromosome 16p11.2 rearrangement phenotypes. In this study, we reconstruct the evolutionary history of this rapidly evolving region of the genome and delineate the pattern of normal human genetic variation at the sequence level. Our data highlight how the expansion of the segmental duplications led to the emergence of *Homo sapiens*-specific duplicated genes and susceptibility to rearrangements associated with autism.

4.3 Results

4.3.1 Evolution and Structural Diversity of Chromosome 16p11.2

To reconstruct the evolutionary history of the 16p11.2 region, we generated complete, reference-quality genome sequences over the region by single-molecule, real-time (SMRT) sequencing of 85 large-insert clones [213, 214] (Table S1), including 46 clones from two nonhuman primate genomes. Duplicated regions are notoriously difficult to sequence and assemble and are often misassembled or missing from nonhuman primate reference genomes [122]; thus, the generation of new and reliable

references was a critical first step. We successfully generated three sequence contigs spanning ~2 Mbp across the 16p11.2 locus (breakpoints BP1-BP5 and flanking unique sequences [205]): one for orangutan and two for different chimpanzee structural haplotypes (**Figure 4.1a**). To understand the pattern of human genetic diversity, we also constructed four new human references, including sequence from breakpoints BP3-BP5 and flanking unique sequences, corresponding to three distinct human structural haplotypes (**Figure 4.1b**).

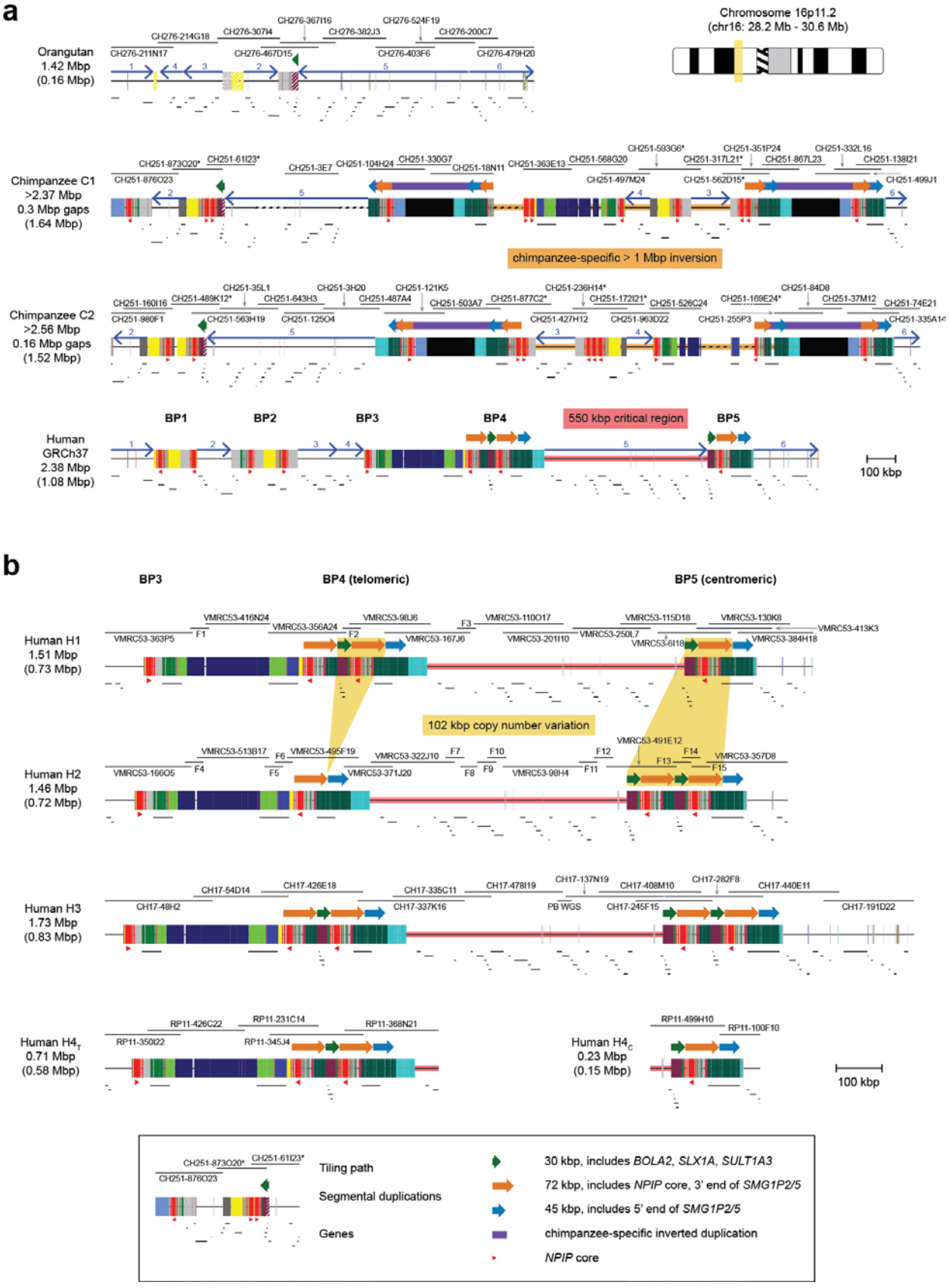

a
Chromosome 16p11.2
(chr16: 28.2 Mb - 30.6 Mb)
Orangutan
1.42 Mbp
(0.16 Mbp)
Chimpanzee C1
>2.37 Mbp
0.3 Mbp gaps
(1.64 Mbp)
chimpanzee-specific > 1 Mbp inversion
Chimpanzee C2
>2.56 Mbp
0.16 Mbp gaps
(1.52 Mbp)
BP1
BP2
BP3
BP4
550 kbp critical region
BP5
Human
GRCh37
2.38 Mbp
(1.08 Mbp)
100 kbp
b
BP3
BP4 (telomeric)
BP5 (centromeric)
Human H1
1.51 Mbp
(0.73 Mbp)
102 kbp copy number variation
Human H2
1.46 Mbp
(0.72 Mbp)
Human H3
1.73 Mbp
(0.83 Mbp)
Human H4T
0.71 Mbp
(0.58 Mbp)
Human H4C
0.23 Mbp
(0.15 Mbp)
100 kbp
Tiling path
Segmental duplications
Genes
30 kbp, includes BOLA2, SLX1A, SULT1A3
72 kbp, includes NPIP core, 3' end of SMG1P2/5
45 kbp, includes 5' end of SMG1P2/5
chimpanzee-specific inverted duplication
NPIP core

Figure 4.1. Comparative Sequence Analysis of Chromosome 16p11.2 Among Apes. a) Schematic depicts the genomic organization of chromosome 16p11.2 for one orangutan and two chimpanzee haplotypes along with the human reference haplotype (GRCh37 chr16:28195661-30573128) (see ideogram for approximate chromosomal location). Breakpoint regions associated with human disease are enumerated (BP1-BP5) [205] with the ~550 kbp autism critical region between BP4 and BP5 highlighted (pink), as well as a >1 Mbp chimpanzee-specific inversion polymorphism (orange). Tiling paths of sequenced clones are indicated above each haplotype, with chimpanzee clones that could not be fully resolved marked with asterisks. Colored boxes and thick arrows indicate the extent and orientation of segmental duplications (with different colors denoting duplicons from different ancestral genomic loci, and hashed boxes indicating sequence duplicated in humans but not in the species represented). Thin numbered arrows show orientations of gene-rich regions of unique sequence. Numbers (left) indicate the size of each orthologous haplotype, with the number of segmentally duplicated base pairs shown in parentheses. Note that for chimpanzee, these sizes are lower bounds due to gaps in the contigs (dotted line sections) and the contigs not reaching unique sequence beyond BP1. b) Schematic depicts distinct human structural haplotypes over the 16p11.2 critical region and flanking sequences (three complete haplotypes extending from unique sequence distal to BP3 to unique sequence proximal to BP5 and one partial haplotype including BP3-BP4 and BP5 sequence contigs). High-quality sequence for each haplotype was generated by sequencing a total of 40 BACs and 15 fosmids from three different human genomic libraries. Regions of copy number variation (highlighted in yellow along the first two haplotypes) occur on both sides of the critical region and involve the same 102 kbp unit in direct orientation, including a 30 kbp block containing *BOLA2* and two other genes and a 72 kbp block harboring a partial segmental duplication of *SMG1* (*SMG1P*). Expansion and contraction of this cassette underlie hundreds of kbp of structural diversity between human haplotypes.

Comparing these reference sequences with the mouse genome GRCm38, we observed conserved gene order synteny between mouse and orangutan (Fig. S7), establishing the orangutan configuration as the most likely ancestral ape configuration. In contrast, both the human and chimpanzee have been dramatically and independently restructured, resulting in marked changes in genomic size, content and synteny between humans and chimpanzees. In both humans and chimpanzees, the euchromatic regions have nearly doubled in length, primarily by the accumulation of segmental duplications (~1 Mbp and >1.5 Mbp of lineage-specific duplications in each lineage, respectively) (**Figure 4.1a**). We identified three inversions between human and orangutan, spanning 47 genes and affecting >1 Mbp of sequence (Fig. S4). One of these inversions (BP1-BP2) is polymorphic among humans [215, 216]. Similarly, we infer five inversions between orangutan and chimpanzee that reordered the same 47 genes (Fig. S5 and Fig. S6). An ~1 Mbp region flanked by the largest chimpanzee-specific segmental duplications associates with a heterozygous inversion in the sequenced chimpanzee. A simulation based on known inversion breakpoints that have occurred during the evolution of the human and chimpanzee lineages indicates that this 2 Mbp region represents the most active inversion hotspot in our genome—a pattern unlikely to have occurred randomly ($p < 1 \times 10^{-6}$, Figs. S55-S56).

All inversion breakpoints map to segmental duplications, often in close proximity to an ~20 kbp LCR16a (low copy repeat 16a) core duplicon—a repeat element associated with the expansion of duplications along chromosome 16 and the emergence of a positively selected gene family (*NPIP*) on the human-African great-ape lineage [72]. Notably, *NPIP* is not present in this region of the orangutan genome but is present 8-11 times in human and chimpanzee (**Figure 4.1a**), suggesting a strong link between its presence and the observed evolutionary instability. In total, the data reveal unprecedented restructuring of chromosome 16p11.2 in both the chimpanzee and human lineages where gene order has been significantly reshuffled by at least six inversions and the region has expanded by more than 1 Mbp as a result of segmental duplication. This lineage-specific accumulation predisposes each great ape to different hotspots of nonallelic homologous recombination (NAHR) and rearrangement (**Figure 4.1a**). Importantly, only within the human lineage do large (>100 kbp) segmental duplications exist in a direct orientation flanking the autism critical region, implying that susceptibility to microdeletion and microduplication at this locus arose specifically within the human lineage.

Compared to chimpanzee, sequenced human haplotypes show far less structural variation. Structural differences between human haplotypes are largely restricted to a 102 kbp duplication block composed of two different segmental duplications originating from chromosome 16—a 72 kbp segment duplicated from chromosome 16p12.1 carrying a portion of the *SMG1* serine-threonine kinase gene (*SMG1P*) and a 30 kbp segment carrying three intact genes, *BOLA2*, *SLX1* and *SULT1A3* (**Figure 4.1b** and Fig. S2). Sequence analysis of human haplotypes indicates that *BOLA2* is a human-specific duplication with paralogs located at BP4 and BP5. Although each complete human haplotype we sequenced contains at least one duplicate copy of *BOLA2*, neither BP4 nor BP5 is fixed for copy number. For example, the H2 haplotype lacks *BOLA2* at BP4, and the H1 haplotype lacks a duplicate *BOLA2* copy at BP5 (**Figure 4.1b**). Human haplotypes differ, thus, by integral changes in the copy number of this 102 kbp segment within both the proximal and distal breakpoint regions (Fig. S3), a finding confirmed by fluorescence *in situ* hybridization (FISH) (Table S2 and Fig. S25).

We constructed a detailed model for the evolution of the region based on comparisons of genomic architecture between species, phylogenetic analyses, and patterns of sequence divergence (**Figure 4.2a** and Figs. S6-S17). The results reveal a rapid and complex series of duplicative transposition, large-scale inversion and interlocus gene conversion events during human-ape evolution (Table S3). More than one dozen large-scale structural changes, including six duplicative transpositions (>830 kbp) from elsewhere on chromosome 16, are required to reconcile the organization of human and chimpanzee chromosome 16p11.2. Conserved gene-order synteny analysis clearly places *BOLA2* ancestrally at BP5 (*BOLA2A*, Fig. S7). The identical junction structure between the *SMG1P* and *BOLA2B* genes at BP4 and tandem duplications at BP5 argues that the 102 kbp cassette originated from an initial tandem duplication event at BP5, followed by a 95 kbp duplicative transposition including this junction to BP4, likely by replication-based template switching (Figs. S12-S13, and Table S4). In order to estimate the timing of the *BOLA2* duplication, we generated a multiple sequence alignment over an ~21 kbp region including the duplicated genes using sequences from our contigs and a gorilla clone containing orthologous sequence (CH277-206A9). Assuming a human-chimpanzee divergence time of 6 million years ago (mya) [132] and a constant substitution rate (Table S6), we estimate that *BOLA2* duplicated across the critical region ~282 kya (95% confidence interval: 361-209 kya, **Figure 4.2b** and Fig. S14), around the time when contemporary *Homo sapiens* emerged as a species [217].

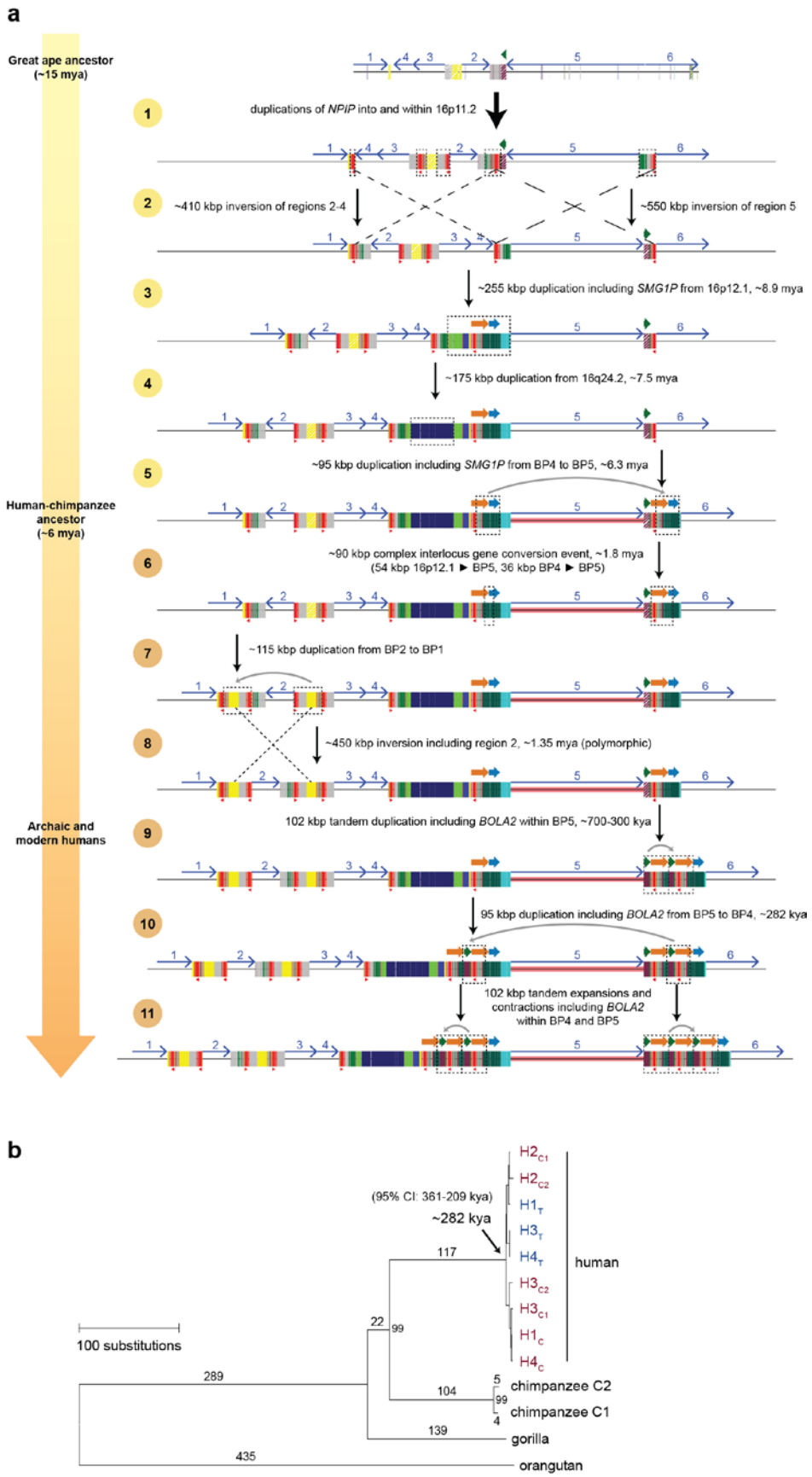

a
Great ape ancestor
(~15 mya)
1
duplications of NPIP into and within 16p11.2
2
~410 kbp inversion of regions 2-4
~550 kbp inversion of region 5
3
~255 kbp duplication including SMG1P from 16p12.1, ~8.9 mya
4
~175 kbp duplication from 16q24.2, ~7.5 mya
5
~95 kbp duplication including SMG1P from BP4 to BP5, ~6.3 mya
Human-chimpanzee
ancestor
(~6 mya)
6
~90 kbp complex interlocus gene conversion event, ~1.8 mya
(54 kbp 16p12.1 ► BP5, 36 kbp BP4 ► BP5)
7
~115 kbp duplication from BP2 to BP1
8
~450 kbp inversion including region 2, ~1.35 mya (polymorphic)
Archaic and
modern humans
9
102 kbp tandem duplication including BOLA2 within BP5, ~700-300 kya
10
95 kbp duplication including BOLA2 from BP5 to BP4, ~282 kya
11
102 kbp tandem expansions and
contractions including BOLA2
within BP4 and BP5
b
(95% CI: 361-209 kya)
~282 kya
117
human
22
99
100 substitutions
289
104
5
99
4
chimpanzee C2
chimpanzee C1
139
gorilla
435
orangutan

Figure 4.2. Dynamic Evolution of Chromosome 16p11.2. a) A model for the evolution of the chromosome 16p11.2 BP1-BP5 region during great ape evolution. The schematic depicts structural changes over time leading to the present-day human architecture (see Supplementary Information and Fig. S6 for details and changes on the chimpanzee lineage). The orangutan structure (top) is largely devoid of segmental duplications and deemed to represent the ape ancestral organization because it is conserved with mouse. Subsequent steps were inferred based on phylogenetic reconstruction, origins of the duplicated sequences, and the most parsimonious path with respect to changes in gene order (inversions). b) A phylogenetic tree representing the last interspersed segmental duplication from BP5 to BP4 in humans (step 10). The unrooted neighbor-joining tree was constructed from a 21,102 bp multiple sequence alignment including allelic and paralogous copies of the *BOLA2*-containing segmental duplications. Human taxon labels denote the haplotypes and locations of different copies (telomeric, T, blue; centromeric, C, red, with C1 closer to the critical region than C2 for haplotypes having two centromeric *BOLA2* copies). The number of substitutions (above each branch) and bootstrap support (at nodes) are indicated. Timing estimates assume a human and chimpanzee divergence time of 6 mya.

4.3.2 Human Copy Number Variation and the Rapid Non-Neutral Expansion of *BOLA2*

We examined copy number diversity of the three intact duplicated genes mapping to the 102 kbp cassette: *BOLA2*, *SLX1*, and *SULT1A3*. We applied a read-depth method [121] to genotype aggregate diploid copy number for each of these genes using massively parallel short-read whole-genome sequence (WGS) data from 2,359 humans [218, 219], 86 nonhuman primates [220], 3 archaic humans [221, 222], 1 Neanderthal [223], and 1 Denisovan specimen [224]. We find that *BOLA2* is duplicated in all *Homo sapiens* examined, including archaic representatives of Neolithic and Mesolithic populations, as well as the oldest sequenced archaic human, Ust'-Ishim, estimated to have lived 45,000 years ago [221]. In sharp contrast, *BOLA2* appears to be single copy (i.e., diploid copy number = 2) among nonhuman primates as well as archaic hominids Neanderthal and Denisova (**Figure 4.3a** and Table S7). Neither *SLX1* nor *SULT1A3* show a pattern of duplication restricted to humans (Fig. S20), although we cannot rule out the possibility that *SLX1* is also a *Homo sapiens*-specific duplication, as its small size reduces precision for genotyping copy number. As Neanderthal and Denisova diverged from the *Homo sapiens* lineage approximately 700 kya [223], this copy number analysis is consistent with the phylogenetic estimate of the duplication occurring ~282 kya.

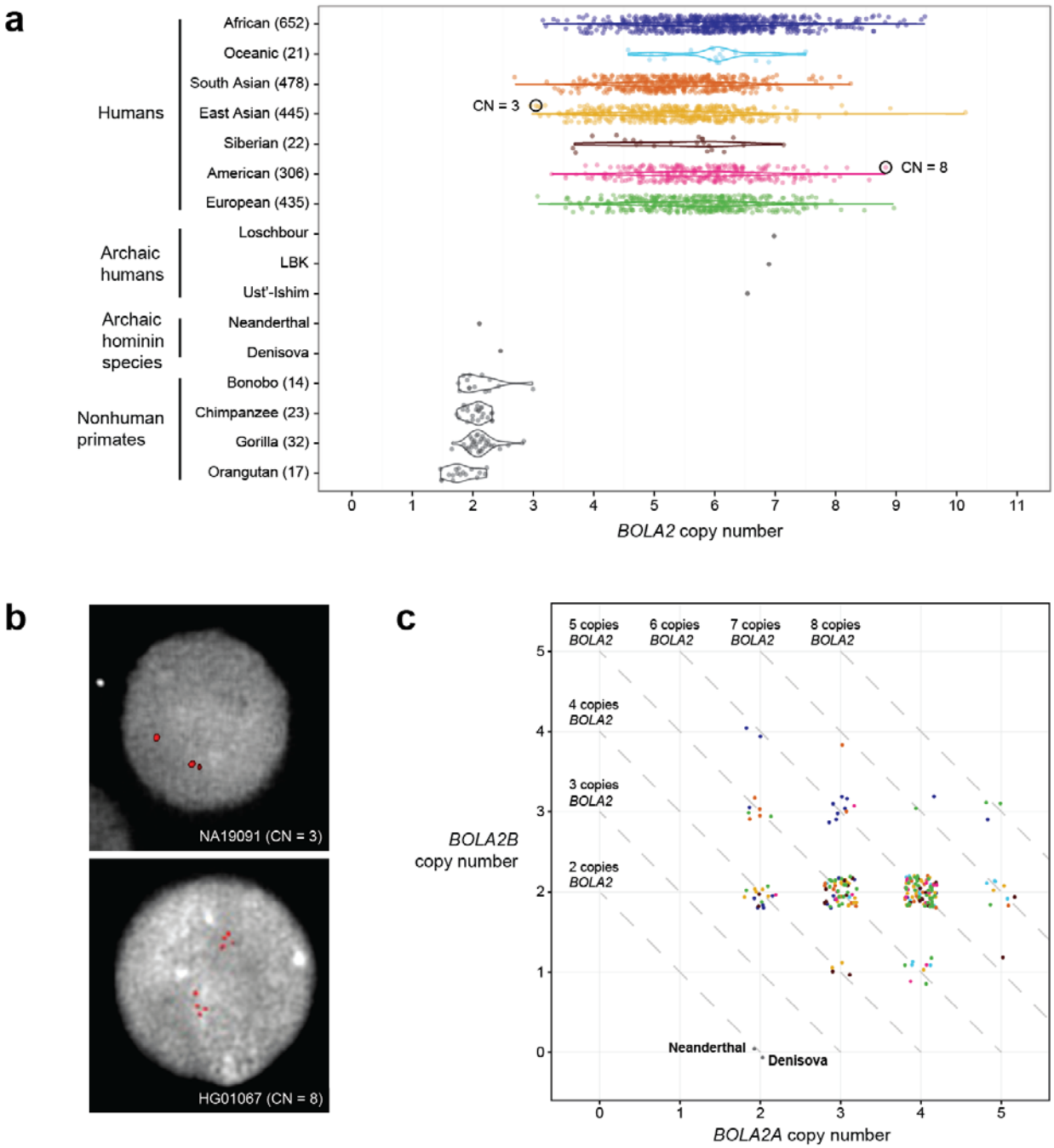


Figure 4.3. *Homo sapiens*-Specific *BOLA2* Duplication and Copy Number Diversity. a) Diploid copy number estimates (points) for *BOLA2* based on sequence read depth are shown for 2,359 humans from seven populations, three archaic humans, a Neanderthal, a Denisova, and 86 nonhuman primates, with violin plots overlaid. b) FISH experiments using a probe targeting *BOLA2* and MIP experiments were performed to assess the accuracy of *BOLA2* copy number genotypes for a subset of individuals across the range of predicted copy numbers and all outliers. Interphase nuclei are shown from individuals having the lowest (three copies) and highest (eight copies) validated aggregate *BOLA2* copy numbers. c) Paralog-specific *BOLA2* copy number genotypes (points, jittered around their integer values for clarity) were inferred from WGS read depth over informative markers for 222 individuals sequenced to high coverage. Colors correspond to different populations as in panel a.

Different human populations show similar distributions in *BOLA2* copy number, with South Asian and East Asian populations having slightly lower means and medians (5 copies) compared to other populations (6 copies, Table S8). Orthogonal FISH analysis and molecular inversion probe (MIP) experiments [225] (Table S2 and Table S9) confirm a range of *BOLA2* copy number in humans from 3 to 8 (**Figure 4.3b**, Fig. S19, and Fig. S26). We developed both computational and experimental methods to genotype *BOLA2* copy number more accurately and to distinguish telomeric (*BOLA2B*) and centromeric (*BOLA2A*) copies (Fig. S21 and Fig. S25). Analysis of a diversity panel of 236 deeply sequenced genomes [219] reveals that all humans have at least one copy of *BOLA2B* (range = 1-4; mean and median = 2 copies) and at least two copies of ancestral *BOLA2A* (range = 2-5 copies; mean and median = 4 copies, **Figure 4.3c** and Fig. S22a). *BOLA2A* shows greater copy number variability (s.d. = 0.77) than *BOLA2B* (s.d. = 0.46), consistent with higher identity tandem arrays at BP5 promoting more NAHR (Fig. S24). Applying our paralog-specific copy number analysis to archaic genomes, we confirm *BOLA2* is duplicated exclusively in *Homo sapiens*, with Neanderthal and Denisova having two copies of *BOLA2A* (**Figure 4.3c** and Fig. S23). Based on a combined analysis of 2,833 individuals, we identified only seven humans with an aggregate *BOLA2* copy number of 3—all with a single copy of *BOLA2B* and two copies of *BOLA2A* (Table S7, Fig. S18, and Fig. S22). Despite the potential for recombination between human haplotypes lacking *BOLA2B* and haplotypes having a single copy of *BOLA2A* and the presence of rare haplotypes with only a single copy of *BOLA2*, no human was identified with an aggregate diploid copy number of 2—the ancestral state of the *Homo* genus.

In light of its recent origin and its potential to promote disease-causing rearrangement, we considered it remarkable that 99.8% of humans carry four or more copies of this segment. Ancient humans such as Ust'-Ishim as well as some of the oldest branches of modern humans (e.g., San and Biaka pygmy [226-229]) typically carry five or six copies, indicating that it spread rapidly early in human history. Because time is required for a new mutation to rise in frequency, we tested whether the near fixation of *BOLA2B* in humans was likely to occur under a neutral model of evolution. We modeled various evolutionary scenarios and assessed the probability of *BOLA2B* being absent from a single

Neanderthal and a single Denisovan but present on 8 of 8 San haplotypes and 108 of 110 Yoruban haplotypes, the observed genotype data for these populations. We initially performed simulations incorporating a realistic model of human demographic history (**Figure 4.4a**) assuming neutral evolution [230-232], both with and without conditioning on the 282 kya age estimate for *BOLA2B*. The observed genotypes were improbable ($p < 0.012$) given the distribution of simulated genotypes under the latter scenario (Fig. S27) and were never observed in simulated data when conditioning on the point estimate for *BOLA2B* age (**Figure 4.4b**). Varying the age parameter yielded the observed genotypes rarely—at most 10 of 1 million simulations matched the observed data, and in these cases *BOLA2B* was at least 1.5 million years old (Fig. S28). These results suggest that the *BOLA2B* duplication unlikely rose to high frequency under neutrality and that this inference is robust to uncertainty in the age of the duplication. Accordingly, we next implemented a model incorporating the 282 kya age estimate but varying the selection coefficient (s) as an input parameter. We calculated probabilities of the observed *BOLA2B* genotype data from simulations spanning a range of values for the selection coefficient, yielding a maximum likelihood estimate of $s = 0.0016$ (**Figure 4.4c**). This scenario suggests that *BOLA2B* or the 102 kbp duplication cassette wherein it resides possibly evolved in our ancestors under positive selection.

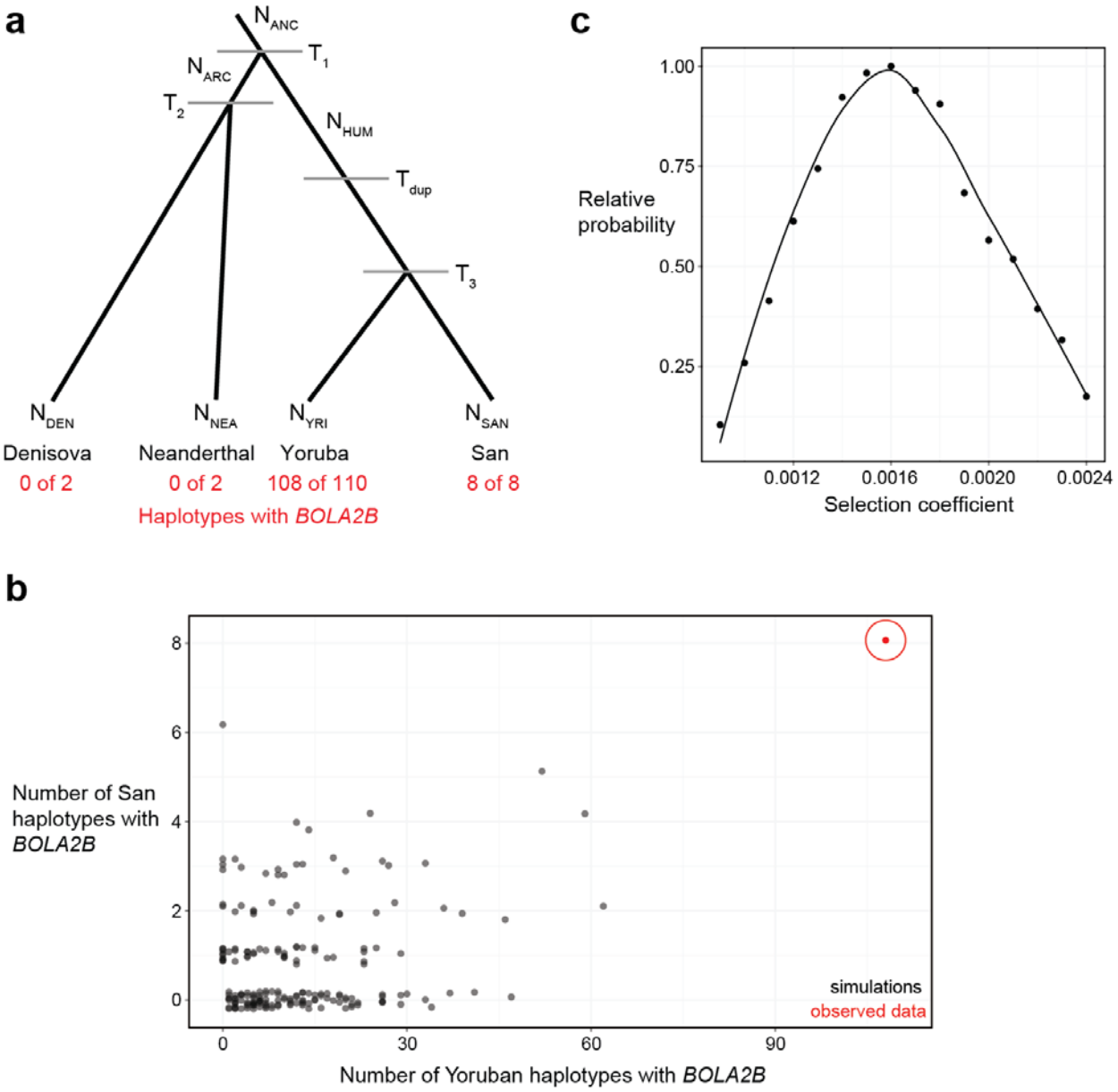


Figure 4.4. Population Genetic Modeling of the *BOLA2B* Duplication. a) Demographic model (adapted from [230]) used to simulate *BOLA2B* evolution under different scenarios. b) Simulation results (n = 1,000,000) assuming the duplication that formed *BOLA2B* occurred once, 282 kya, along the modern human ancestral lineage and evolved under neutrality compared to the observed genotype frequencies of *BOLA2B* in 8 San and 110 Yoruban haplotypes. Nearly all (999,807) simulations resulted in *BOLA2B* being lost from both populations; results from the remaining 193 simulations (black) are shown alongside the observed data (red, circled). Under this simple neutral model incorporating *BOLA2B* age, the observed *BOLA2B* frequency is never approached. c) Simulation was repeated exploring a range of selection coefficients from 0.0009 to 0.0024 (increments of 0.0001) and the relative probability of the observed data under each scenario was calculated as the number of simulations yielding the observed *BOLA2B* genotypes relative to the maximum number of such simulations for any single selection coefficient considered. The maximum likelihood estimate for the selection coefficient was $s = 0.0016$. Smoothed line is LOESS regression curve.

4.3.3 *BOLA2* Expression and Discovery of a Novel *Homo sapiens*-Specific Fusion Transcript

The most direct consequence of the *Homo sapiens* duplication would be to increase dosage of *BOLA2* gene products. Human BOLA2A and BOLA2B show no amino acid (aa) differences and are both predicted to encode a small 86 aa protein (Fig. S38) that evolved under negative constraint ($\omega = 0.1899$, $p = 0.0069$) during primate evolution (Figs. S39-S41). Because humans show extensive copy number variation, we assessed whether copy number correlated with mRNA and protein levels. We find a significant correlation ($r = 0.36$, $p = 2.09 \times 10^{-12}$) between *BOLA2* copy number and expression at the RNA level based on analysis of 366 lymphoblastoid cell lines (LCLs) [233] (**Figure 4.5a**, Figs. S34-S36, and Tables S10-S11). Conditioning on aggregate copy number, the number of distal versus proximal copies had no effect on expression levels, indicating that copies from both BP4 and BP5 contribute equally to the overall expression level. This observation is consistent with the fact that the duplication was likely large enough to carry most of the predicted *BOLA2* regulatory machinery and, thus, equal transcriptional potential. We next tested whether this correlation with copy number extended to BOLA2 protein levels. After ensuring the specificity of three different commercial antibodies via transfection experiments (Fig. S37), we analyzed whole protein lysates from 34 LCLs [233] and identified a consistent ~10 kDa protein by Western blot analysis (Fig. S42). Similar to RNA levels, BOLA2 protein levels correlated with copy number (**Figure 4.5b**, Fig. S43, and Tables S12-S13, $r = 0.64$, $p = 4.34 \times 10^{-5}$). Immunofluorescence and Western blot analyses suggest that the protein localizes in the cell cortex (Fig. S44) similar to its interacting partner glutaredoxin 3 [234, 235] and is likely secreted outside of the cell using a non-classical secretory pathway as previously described [236].

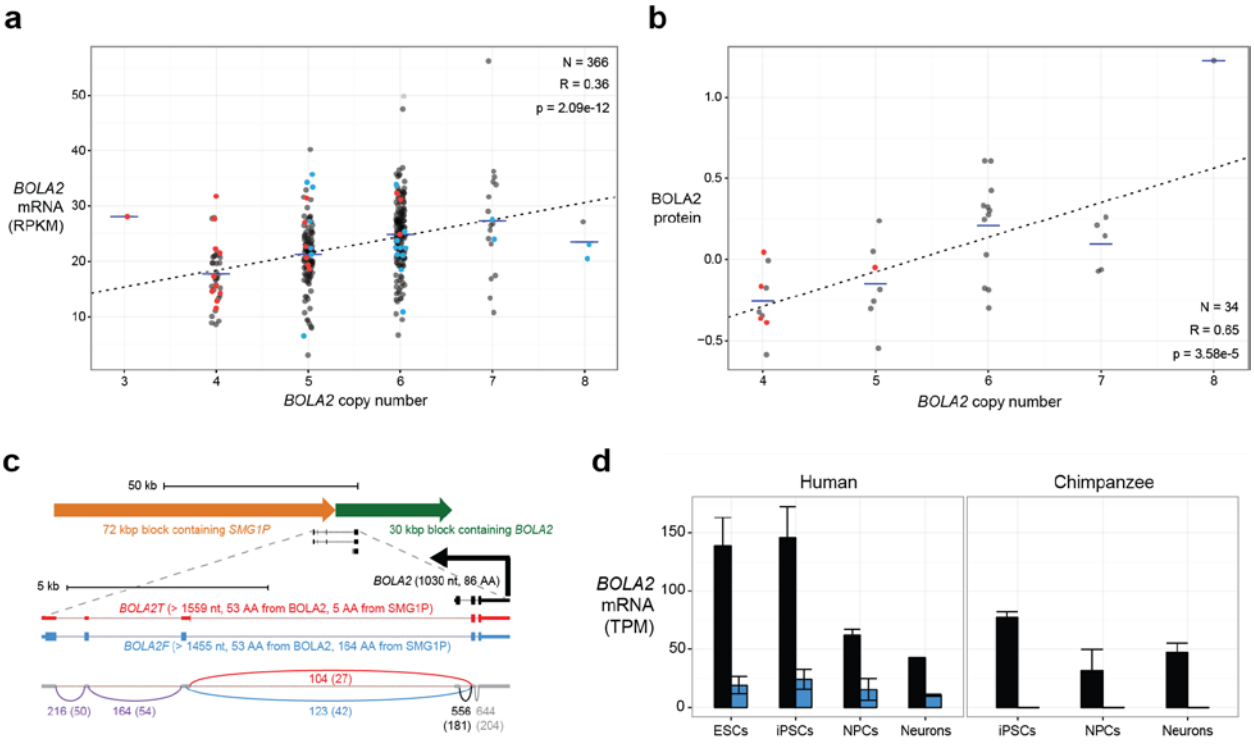


Figure 4.5. *BOLA2* Expression Analyses. a) Normalized *BOLA2* mRNA expression was quantified using RNA-seq data from 366 lymphoblastoid cell lines from individuals genotyped for *BOLA2* paralog-specific copy number. Points indicate expression levels and copy number (jittered) for each cell line, and horizontal lines show the mean expression level for each copy number. These data reveal a modest yet significant correlation between *BOLA2* copy number and RNA expression. Least squares regression line is shown. Point colors indicate *BOLA2B* copy number (red = 1 copy, black = 2 copies, blue = 3 copies). Conditioning on aggregate *BOLA2* copy number showed that groups with the same total copy number but different combinations of paralog-specific copy number do not exhibit differential expression, consistent with both *BOLA2A* and *BOLA2B* being expressed at the RNA level. b) Plot layout is the same as in panel a, but data show BOLA2 protein expression quantified by Western blot densitometry on protein lysates from 34 lymphoblastoid cell lines from individuals genotyped for paralog-specific *BOLA2* copy number. Protein expression appears to correlate more strongly than RNA expression with *BOLA2* copy number. Though the sample size is small, no evidence suggests differential protein expression of distinct *BOLA2* paralogs. c) *BOLA2* gene models, predicted protein products, and support from RNA-seq data from human iPSCs. RT-PCR, cloning, and capillary sequencing experiments identified three *BOLA2* isoforms: the canonical isoform (*BOLA2*, black) encoding an 86 aa protein and two fusion isoforms consisting of the first two exons from canonical *BOLA2* fused with three exons from duplicated *SMG1* sequence. One of the fusion isoforms (*BOLA2F*, blue) maintains the *BOLA2* ORF well beyond the fusion junction and is predicted to encode a 217 aa protein deriving primarily from SMG1P, while a third isoform (*BOLA2T*, red) contains a premature stop codon within the first *SMG1P*-derived exon. RNA-seq reads from two human iPSCs (two independent clones each) were analyzed to validate the accuracy of the three gene models and provide some insight into the relative expression levels of the different isoforms. Numbers next to curved lines indicate mean counts of reads supporting each exon-exon junction, with standard errors in parentheses. d) RNA-seq quantification of *BOLA2* mRNA expression through *in vitro* differentiation of primate iPSCs into neurons. Data from two human and two chimpanzee cell lines (two independent clones each, except for neurons) reveal significantly higher levels of *BOLA2* transcripts in human iPSCs than in chimpanzee iPSCs and that *BOLA2* RNA levels decrease through neuronal differentiation. Bar heights indicate mean expression levels for each species and differentiation stage in transcripts per million (TPM), with error bars showing standard errors. Bar colors correspond to different *BOLA2* isoforms as in panel c. *BOLA2* expression in human ESCs (two cell lines) is consistent with data from human iPSCs, suggesting the iPSC data accurately reflect *BOLA2* expression at early stages in development.

During our transcript analysis, we identified several additional *BOLA2* splice isoforms in addition to the canonical three-exon transcript (Fig. S29). One of the most abundant mRNAs indicated an alternate gene structure composed of same transcription start site and first two exons as the canonical *BOLA2* isoform as well as three novel 3' exons. These three additional exons were derived from an older segmental duplication containing part of the *SMG1* gene (*SMG1P*) that became juxtaposed adjacent to *BOLA2* as a result of the original 102 kbp segmental duplication (**Figure 4.5c**). To validate these transcripts and their protein-coding potential, we performed RT-PCR on a panel of tissues (Figs. S30-S31) and cloned and sequenced products from the brain. The results revealed the presence of a fusion open reading frame (ORF) isoform in addition to the canonical transcript. This 217 aa ORF fusion transcript is formed by the juxtaposition of two exons from *BOLA2* (53 aa) and three exons from *SMG1P* (164 aa). RNA-seq analyses indicate that both transcript forms are ubiquitously co-expressed in a wide variety of tissues and developmental stages (**Figure 4.5c**, Fig. S32, and Figs. S45-S46). Importantly, since the ancestral *BOLA2* at BP5 lacked the *SMG1P* duplication in the correct orientation, the origin of this fusion product must have coincided with the juxtaposition of *BOLA2* and *SMG1P* by the tandem 102 kbp segmental duplication ~700-300 kya at BP5. We conclude that this fusion product is *Homo sapiens*-specific. Because the duplication from BP5 to BP4 included the junction between *BOLA2* and *SMG1P*, potential exists for transcription of the fusion transcript from duplicate *BOLA2* copies at BP4 or BP5.

BOLA2 was previously identified as one of the top 50 genes differentially expressed between humans and nonhuman apes in induced pluripotent stem cells (iPSCs) [237], implying that this gene may be particularly relevant early in development. Based on our characterization of the different *BOLA2* isoforms, we revisited this observation by quantifying *BOLA2* mRNA levels by RNA-seq in human and chimpanzee iPSCs. We also performed RNA-seq on human and chimpanzee neural progenitor cells (NPCs) and eight-week-old neurons. Remarkably, we found the greatest differences in canonical *BOLA2* expression at the iPSC state (2-fold) and to a lesser extent in NPCs (1.5-fold) (**Figure 4.5d**, Fig. S45, and Table S14). There was no significant difference in *BOLA2* expression between human and chimpanzee eight-week-old neurons. Quantification of *BOLA2* expression in two human embryonic stem cell (ESC)

lines revealed transcript levels comparable to iPSCs, confirming that *BOLA2* is expressed at higher levels in human stem cells compared to nonhuman primate stem cells (**Figure 4.5d**, Fig. S45, and Table S14). The expression of the fusion *BOLA2-SMG1P* transcript was also detected exclusively in human cells, as expected (Fig. S33 and Figs. S45-S46). We found that its expression remains relatively constant across iPSCs, NPCs, and neurons (**Figure 4.5d**). Together, these results highlight both quantitative and qualitative differences between human and chimpanzee *BOLA2* expression and suggest these differences are greatest early during embryonic development.

4.3.4 Susceptibility to 16p11.2 Rearrangements

Our evolutionary reconstruction implies that directly oriented segmental duplication blocks flanking the 16p11.2 critical region were already present in the genomes of our ancestors prior to the *Homo sapiens*-specific expansion of *BOLA2* (**Figure 4.2a** and Fig. S47). The duplication of *BOLA2* across the critical region, however, increased by threefold the size of high-identity sequence blocks (Fig. S12), likely increasing the propensity of this locus to microdelete and microduplicate via NAHR (Fig. S3). To test this, we systematically refined breakpoint locations [238] in autism and developmental delay patients carrying either the chromosome 16p11.2 microduplication or microdeletion event [239]. First, we generated and analyzed WGS data from six families (19 genomes), each including at least one proband having a *de novo* microdeletion (Table S18). We calculated normalized read depth at all positions corresponding to unique 30-mer sequences in the reference genome GRCh37 [240]. In every case, the normalized read depth at mapping locations in between the *Homo sapiens*-specific duplicated sequences in probands was about half that observed in unaffected parents, while the normalized read depth at positions outside the *Homo sapiens*-specific duplicated sequences on both sides of the critical region was equivalent in probands and parents (**Figure 4.6a** and Fig. S49). These observations refine the 16p11.2 microdeletion breakpoints in these six families to the ~95 kbp interval corresponding to the *Homo sapiens*-specific duplication including *BOLA2*.

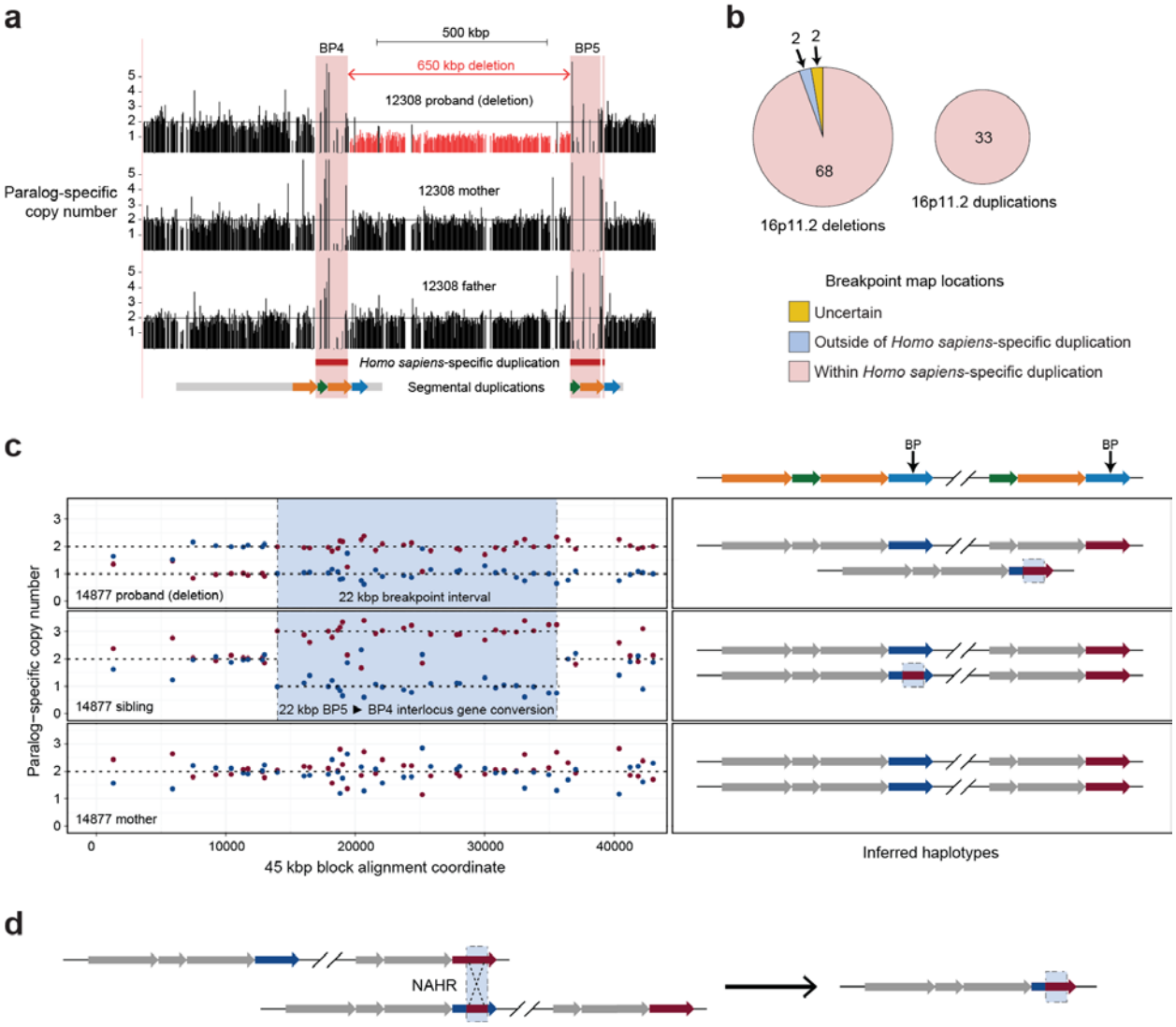


Figure 4.6. Refinement of 16p11.2 Rearrangement Breakpoints. a) Results of whole-genome sequencing of a family with a *de novo* 16p11.2 microdeletion in a child with autism. Normalized read depth at unique 30-mer positions in the human reference genome GRCh37 is depicted for the proband, her mother, and her father, respectively. Read-depth signatures reveal a deletion in the proband extending between but not beyond the *Homo sapiens*-specific duplicated sequences (highlighted in pink). b) Summary of results across 105 independent microdeletion and microduplication events from 152 individuals, based primarily on MIP sequencing of informative sites. ~96% of breakpoints map to the *Homo sapiens*-specific segmental duplication. c) Data from an atypical patient where the breakpoints are inferred to map outside of the *Homo sapiens*-specific segmental duplication. The plots show paralog-specific copy number for a 16p11.2 microdeletion proband, his sibling, and his mother over a 45 kbp duplication block shared between BP4 and BP5. Paralog-specific copy number was estimated using a MIP assay targeting 54 informative markers over this region. Adjacent schematics indicate the inferred haplotypes for each individual based on this data with approximate breakpoint locations shown (arrows). The results demarcate the location of the unequal crossover interval based on the reciprocal copy number transition between the BP5 (red) and BP4 (blue) segmental duplications. In this case, the breakpoints clearly map to a 22 kbp region outside of the typical hotspot. Analysis of the sibling suggests that this region was the site of an interlocus gene conversion event from BP5 to BP4, and data from the mother imply that chromosomes having this event were present in the paternal germline. DNA from the father was not available for testing. d) It is likely that the high degree of sequence identity between converted BP4 sequence and BP5 sequence promoted NAHR within the 22 kbp conversion interval during paternal meiosis, leading to the microdeletion observed in the patient.

We next leveraged our paralog-specific *BOLA2* copy number genotyping analyses [225] to validate this result in the six families and to localize breakpoints for an additional 145 rearrangement carriers (99 additional independent rearrangement events, Table S18). Both WGS- and MIP-based approaches detect signatures of reciprocal marker-specific copy number transitions corresponding to unequal crossover breakpoints [225, 238] (Fig. S50). Marker-specific read count frequency data confirm the location of breakpoints in the six families (Fig. S51). Most strikingly, they reveal that, with the exception of two individuals having breakpoints mapping within a 45 kbp duplication block (in one case clearly within a region homogenized by interlocus gene conversion), all resolved breakpoints occur within the *Homo sapiens*-specific duplication (**Figure 4.6b-d**, Fig. S2, Figs. S52-S53, and Table S18). In two cases, it was impossible to discriminate between breakpoints occurring outside the ~95 kbp interval or an interlocus gene conversion event yielding the same expected signal (Fig. S54). Overall, we observe that at least 96% (101 of 105) of rearrangement breakpoints map within the *Homo sapiens*-specific duplication including *BOLA2*, strongly implicating this recent evolutionary event in rendering the 16p11.2 locus susceptible to recurrent rearrangements in humans.

4.4 Discussion

The selective disadvantage of segmental duplications has been extensively studied [241]. More than one-third of the copy number variant burden associated with developmental delay in the human population [242] is directly linked with the expansion of segmental duplications in the ancestral lineage of humans and great apes [40, 120]. More recently, there has been compelling evidence that the selective disadvantage of human segmental duplication architecture may be offset by the advantage provided by newly minted human-specific genes that emerged as a result of duplication—the "core duplicon hypothesis" [43, 78]. *SRGAP2C*, for example, is a human-specific duplicate gene associated with neuronal migration and increases in dendrite density [98, 113]. Similarly, *ARHGAP11B* is a human-specific duplicate expressed highly in basal radial glial cells and implicated in human cortical neuronal expansion and gyrification of the brain [243]. In the latter example, the "birth" of the human-specific

duplicate is also associated with increased genomic instability predisposing our lineage to recurrent rearrangements associated with developmental delay while simultaneously conferring a potential selective benefit [43].

Our findings regarding the evolution of human chromosome 16p11.2 and susceptibility to one of the most common genetic causes of autism are consistent with this evolutionary model. The level of restructuring at chromosome 16p11.2 over the last 15 million years of evolution is exceptional (**Figure 2a** and Fig. S6). The region shows the greatest density of evolutionary inversions in the primate genome. These changes occurred in concert with dramatic lineage-specific accumulation of segmental duplications and expansions of gene families. The level of genetic difference between humans and chimpanzees for chromosome 16p11.2 stands in sharp contrast to the oft-quoted 99% genetic identity between the species.

As part of this restructuring, a 102 kbp segment duplicated specifically in *Homo sapiens* after our divergence with ancient hominins, and we have shown that copy number varies extensively among modern humans but never returns to the ancestral diploid state. As humans diverged from Neanderthal and Denisova <700 kya [223, 224], the duplication was likely absent or segregating at low frequency at this point in time, consistent with our phylogenetic point estimate of its origin ~282 kya. This makes it the youngest interspersed segmental duplication to have risen to high frequency in the human species. The size of this duplication event also makes it the largest amount of derived sequence specific to the human lineage, dwarfing the 35,500 single-nucleotide variants and indels that have been previously reported [223]. The rapid rise and distribution of this duplicated segment at the root of *Homo sapiens* is unlikely to have occurred under a model of neutral evolution. Our simulations are consistent with modest positive selection ($s = 0.0016$). Given that the *BOLA2B* duplication predisposes to recurrent, disease-associated rearrangements, the actual benefit must exceed the deleterious effect of the architecture itself [244]. An alternative explanation is that the rise to high frequency was driven by recurrent duplicative transposition from BP4 to BP5 and potential rapid interlocus gene conversion such that these sites have been driven to near fixation, with at least one duplication on every human haplotype. Although also an exciting

possibility, there is no known mutational mechanism to account for such rapid and recurrent interlocus gene conversion in the human genome.

The high degree of sequence identity resulting from the *BOLA2B* duplication (Fig. S12, Table S5, and Table S15) renders human chromosome 16p11.2 particularly susceptible to NAHR. At least 96% of disease-associated rearrangement breakpoints map to the *Homo sapiens*-specific duplication block. This is consistent with the longest and most highly identical tracts of sequence identity driving unequal crossover events [245]. As a result of the duplication, there are more than 60 directly oriented tracts at least 500 bp in length with perfect identity in humans (Table S16 and Table S17). Despite harboring a more complex duplication architecture, chimpanzee genomes lack such tracts, and they are also absent in orangutan. In contrast, they were found in all human haplotypes, with the greatest number and longest tracts found in haplotypes that include the human-specific duplicate *BOLA2B* (Fig. S48, Table S16, and Table S17). The evolution of this susceptibility occurred in concert with the expansion of the *BOLA2* gene, which shows a commensurate increase in mRNA and protein expression. Although the expansion of *BOLA2* is circumstantial and may not be the source of the selection, it is intriguing that i) the duplication was also associated with the formation of a novel fusion transcript and ii) that the *BOLA2* gene shows the greatest difference in expression between humans and chimpanzees early in embryonic development. Biochemical data suggest that BOLA2 physically interacts as a heterodimeric complex with glutaredoxin 3 and has a function critical for iron metabolism and cellular signaling [234]. Iron is one of two selective brain minerals; it is an essential co-factor for many enzymatic reactions and its level is critical for immune response [246, 247], maintenance of pregnancy, and proper human development [248]. More efficient metabolism of iron relieves the metabolic constraint of the developing brain [249, 250]. Although the phenotypic consequences of increased BOLA2 concentration, the fusion transcript, and variation in expression in the human species await future experimental characterization, it is tempting to speculate that the expansion of this conserved gene may be related to enhanced intracellular iron homeostasis associated with changes in the ancestral diet of humans [251] and/or pathogen resistance [246, 247].

4.5 Methods

SMRT sequencing was used to generate high-quality sequence from bacterial artificial chromosome (BAC) clones obtained from genomic libraries [213]. Clone sequences were assembled using HGAP and error-corrected using Quiver [214]. Contig assembly was performed using Sequencher (Gene Codes Corporation, Ann Arbor, MI) and validated by FISH. Copy number genotyping of genes and segmental duplications was performed using a read-depth method [121] and WGS sequence data from humans [218, 219], nonhuman primates [220], and archaic genomes [221-224][26-29], as well as single-molecule molecular inversion probes (smMIPs) [199] targeted to paralogous sequence variants [225]. We estimated evolutionary timing of segmental duplication events based on comparative sequencing and phylogenetic analyses (neighbor-joining method) adjusting branch lengths for trees that failed the Tajima's relative rate test and assuming divergence times of 6 mya (human-chimpanzee) and 15 mya (human-orangutan). Evolutionary conservation analysis of *BOLA2* was performed by maximum likelihood (PAML). Likelihoods of *BOLA2B* fixation under different scenarios were assessed using the coalescent simulators *ms* [231] and *msms* [232], adapting a previously published demographic model [230]. *BOLA2* copy number estimates were correlated (Pearson's r) using RNA-seq quantifications (PEER-normalized RPKM) [233] and Western blot BOLA2 densities in human LCLs grown in complete RPMI medium and lysed in RIPA buffer. After SDS-PAGE and transfer to PVDF membrane, blots were incubated with an anti-BOLA2 antibody (Santa Cruz, CA) and band densities quantified using the Bio1D software. *BOLA2* CDS was cloned using the Gateway system (Invitrogen, Carlsbad, CA). HeLa cells were transfected with CMV-*BOLA2* CDS (both 10 and 17 kDa forms) and analyzed by Western blotting and immunofluorescence. LCL and HeLa cells probed with anti-BOLA2 antibody were observed using a Zeiss LSM710 confocal microscope. *BOLA2* gene models were established via RT-PCR, cloning, and capillary sequencing. RNA-seq data was generated from previously described ES and iPS cell lines [237], as well as iPS cell lines differentiated into NPCs and neurons. *BOLA2* mRNA expression was quantified in transcripts per million (TPM) with Kallisto [252] (version 0.42.1) using a custom catalog of transcripts including all human RefSeq transcripts with the three *BOLA2* isoforms. Breakpoints of chromosome

16p11.2 rearrangements were refined using Illumina whole-genome shotgun sequencing [238, 240] and smMIP analysis [225, 238] of patient DNA obtained from the Simons Variation in Individuals Project (Simons VIP) and Simons Simplex Collection (SSC). All procedures for clinical assessment and blood extraction were approved by the institutional review boards (IRBs) of participating institutions, and informed consent was obtained for participation in this research.

4.6 Notes

Accession Numbers

Upon acceptance for publication, clone sequences, haplotype contig sequences, and patient whole-genome shotgun sequence data will be made available at GenBank and dbGaP. RNA-seq data for NPCs and neurons will be made available at GEO. The paper will contain the specific accession numbers.

Acknowledgements

We are grateful to all of the families at the participating Simons VIP and SSC sites, as well as the Simons VIP Consortium. Approved researchers can obtain the Simons VIP dataset, the SSC dataset, and/or biospecimens by applying at https://base.sfari.org. We thank M. Chaisson for providing SMRT WGS data, B. Nelson and K. Munson for technical assistance, and T. Brown for assistance with manuscript preparation. This work was supported by the Paul G. Allen Foundation (grant #11631 to E.E.E.), grants from the Simons Foundation Autism Research Initiative (SFARI #303241 to E.E.E. and #274424 to A.R.), and grants from the U.S. National Institutes of Health (NIH grant 2R01HG002385 to E.E.E.) and the Swiss National Science Foundation (31003A_160203 and CRSII33-133044 to A.R.). X.N. was supported by a U.S. National Science Foundation Graduate Research Fellowship under grant #DGE-1256082. G.G. was awarded a Pro-Women scholarship from the Faculty of Biology and Medicine, University of Lausanne. M.H.D. is supported by U.S. National Institute of Mental Health grant 1F30MH105055-01 and by the Simons Foundation. E.E.E. is an investigator of the Howard Hughes Medical Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Financial Interests

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc., and is a consultant for the Kunming University of Science and Technology (KUST) as part of the 1000 China Talent Program.

5. Summary and Future Directions

5.1 Lessons from the Gaps

A fundamental goal of modern biological science is to elucidate the relationships between genotype and phenotype, to understand how genomes influence traits. This research in humans and our closest evolutionary relatives is stifled by incomplete pictures of corresponding genomes and their full spectrum of variation. Remarkably, even the human reference genome, one of the highest-quality finished assemblies, contains gaps in euchromatic sequence more than a decade after the Human Genome Project (HGP) was completed [253]. My investigation of *SRGAP2* exemplifies how these poorly characterized regions contain genes, including perhaps some of the most important genes for the evolution of our species. In the words of science writer Ed Yong [254], "the reference genome, supposedly the full catalogue of human DNA, may be missing some of the elements that most make us human."

Nonhuman ape genomes suffer from gaps and misassemblies to a much larger degree than the human reference. In sequencing these genomes, quality was sacrificed for reduced time and cost, as the HGP clone-by-clone sequencing strategy was eschewed in favor of a whole-genome shotgun approach [74, 117-119]. Regions harboring segmental duplications typically exhibit the lowest contiguity and accuracy due to copy number polymorphism and high sequence identity with paralogous loci [98, 122, 123]. These very regions, however, are crucibles for evolutionary change [4], accounting for more variant bases between human and chimpanzee (2.7%) than single nucleotide substitutions (1.2%) [34]. My examination of the evolution of chromosome 16p11.2 reveals drastic differences in size, gene order, and orientation between human, chimpanzee, and orangutan genomes that arose within the past ~15 million years. Thus, chromosome 16p11.2 dispels the conventional paradigm that humans and chimpanzees are virtually identical at the genetic level [255]—at least some genomic regions differ drastically between these species. Importantly, generating high-quality genome sequence data was a necessary first step for evolutionary reconstruction, comparative analyses, and assessment of human variation. Efforts to characterize other duplication-rich loci at the genomic level have identified (and will likely continue to

uncover) around one dozen to a few dozen additional regions divergent in structure among ape species [155, 240, 256].

This dissertation highlights basic deficiencies in our current knowledge of human and nonhuman primate genomes, shows how targeted clone-based sequencing can resolve them, and demonstrates that their resolution can yield valuable biological insights. Filling remaining gaps in the human genome and generating high-quality assemblies for chimpanzee, gorilla, and orangutan will undoubtedly prove crucial for deciphering the genetics of human evolution. In addition, better characterization of duplicate genes and their variation in copy number and sequence content is necessary. Despite demonstrated roles in diversity [121, 167-169], evolutionary adaptation [1, 4, 73, 76], and disease [44, 53, 170, 171], duplicate genes and their progenitors are often excluded from genetic analyses in large part because they are difficult to accurately genotype [151, 173]. My MIP assay renders many of these genes accessible for the first time, enabling exploration of their potential impacts on disease and human phenotypic variation. In the work presented here, I added ~380 kbp of sequence to the human reference genome sequence and discovered >800 kbp of copy number variation affecting HSD genes within the normal human population. Future efforts to understand the genetic underpinnings of complex disease and other human traits would greatly benefit from examining a more comprehensive picture of genetic variation than single nucleotide variants alone and including these duplicated regions routinely overlooked in most studies today.

5.2 Experiments of Nature

The preceding chapters testify to the power of probing experiments of nature—examining patterns of extant genetic diversity and their evolutionary origins and detailing disruptive mutations. Identification of healthy individuals homozygously lacking *SRGAP2B* and *SRGAP2D* provided strong evidence that these HSD genes are not critical for brain development. Conversely, the near-fixation of *SRGAP2C* and *BOLA2B* suggests their possible functional importance. The evolution of *BOLA2B* is especially striking. Its dramatic rise in frequency within the past 300 thousand years is inconsistent with a simple model of neutrality. Furthermore, data from microdeletion and microduplication patients

definitively implicate the *BOLA2B* duplication in predisposing the region to disease-associated rearrangements. One intriguing explanation for the high frequency of the *BOLA2B* duplication despite this deleterious effect and its young age is that it conferred some fitness benefit, perhaps related to the *Homo sapiens*-specific fusion transcript.

Ethical and technical considerations prohibit classic gene knockout experiments in humans. Although such experiments can be performed in human cells using *in vitro* systems, these systems do not always accurately reflect *in vivo* biology underlying human development. Nature's experiments provide an attractive alternative: occasionally, naturally occurring mutations will disrupt genes of interest [257]. Provided that these disruptions are not lethal, they have potential to yield insights into the functions of affected genes. Leveraging experiments of nature should prove especially powerful for studies of HSD genes. If an HSD gene has evolved an important function, it follows that damaging mutations will contribute to disease.

Implementing this logic, I developed the MIP genotyping method to screen large patient cohorts for loss-of-function mutations in *SRGAP2C*. This gene's expression in the brain and its effects on dendritic spine development when expressed ectopically in mouse [113] suggest such mutations would most likely manifest as neurodevelopmental disorders. Accordingly, I genotyped over 11,000 total patients with autism spectrum disorder, intellectual disability, epilepsy, or some combination of these conditions, as well as over 4,000 controls. I identified ten patients (all with intellectual disability and/or autism) having homozygous disruptions of *SRGAP2C*: nine with homozygous 38 kbp deletions including exon 2 (**Figure 5.1**), and one with a single base pair frameshifting deletion in exon 3. In contrast, no homozygous *SRGAP2C* losses were observed in controls, though the 38 kbp deletion was found heterozygously at low frequency (<3%) in populations having at least some European ancestry. Although the excess of complete *SRGAP2C* loss in cases versus controls is not statistically significant ($p = 0.13$), these data further support *SRGAP2C* as a candidate contributor to human brain evolution. Phenotype data is not available for all individuals with homozygous *SRGAP2C* loss, but several patients exhibit similar clinical presentations. For example, two have orbitofrontal cortex sizes more than two standard deviations

below the mean for their ages, two have hearing loss, two have psychosis, and three have head size abnormalities.

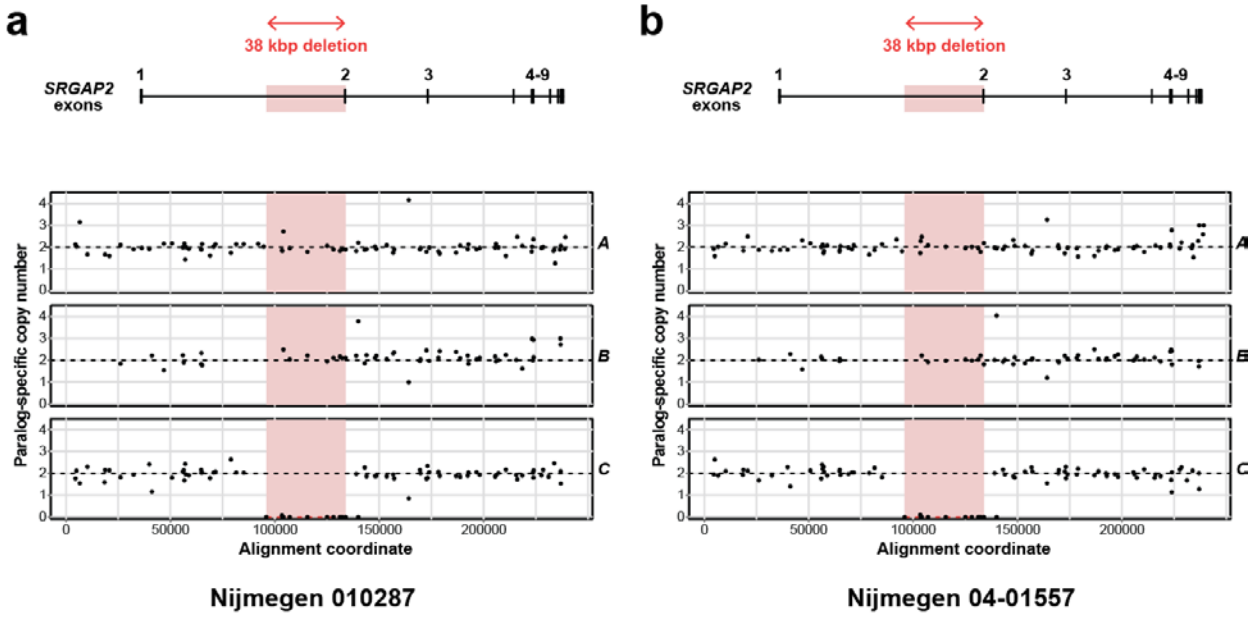


Figure 5.1. Homozygous *SRGAP2C* Exon 2 Deletions in Intellectual Disability Patients. Plots show data from the MIP copy-number assay, with the region homozygously deleted in *SRGAP2C* highlighted.

5.3 New Frontiers for Disease and Evolution

The evolutionary histories, genetic diversity profiles, and initial functional characterizations of two HSD gene families, *SRGAP2* and *BOLA2*, reveal their likely relevance for human evolution and disease. These families were prioritized based on known involvement of the ancestral paralog in neuronal development (*SRGAP2* [127]) and association with recurrent rearrangements constituting one of the most common known genetic contributors to autism (*BOLA2* [212]). Nevertheless, it is tempting to speculate that investigating other HSD genes not yet well characterized will yield a wealth of anthropogenic insights. Unlike chimpanzee-specific duplicate genes [41], four HSD genes have been implicated in neurodevelopment [120, 121]. Like *SRGAP2*, *GPRIN2* is an inducer of neurite outgrowth [128], and mutations in *HYDIN* in mouse have been linked to hydrocephalus [129], the accumulation of cerebrospinal fluid in the brain. Most intriguingly, the HSD gene *ARHGAP11B* has been reported to

promote neocortex expansion and induce gyrification, the formation of the hallmark folds of the cerebral cortex [243]. Thus, continued genetic characterization of HSD genes promises an exciting path forward for future studies of human origins.

An important parallel objective for studies of HSD genes should be thorough genomic characterization of the regions in which they reside. High-quality sequencing of chromosome 15q13.3 [240], chromosome 16p12.1 [155], and chromosome 17q21.31 [256] in multiple ape species have shown that structural evolutionary changes are not restricted to chromosome 16p11.2. Perhaps the most interesting region in this regard is chromosome 1q21, an ~12 Mbp locus containing members of four HSD gene families (**Figure 5.2**). Recurrent rearrangements in this region have been associated with a wide spectrum of disease phenotypes, including macrocephaly, microcephaly, mental retardation, and autism [58, 64, 65]. The HSD gene *HYDIN2* was reported to be affected by these rearrangements and has been proposed as a candidate for head size anomalies in microdeletion and microduplication patients [65]. Following up on this hypothesis, MIP genotyping indicated that *HYDIN2* is most commonly, but not always, included in the 1q21 microdeletion/microduplication interval (**Figure 5.3**). Identification of more patients with atypical rearrangement events (i.e., excluding *HYDIN2*) and comprehensive phenotyping of these individuals and those having typical rearrangements will be necessary to elucidate the potential role of *HYDIN2* dosage in disease.

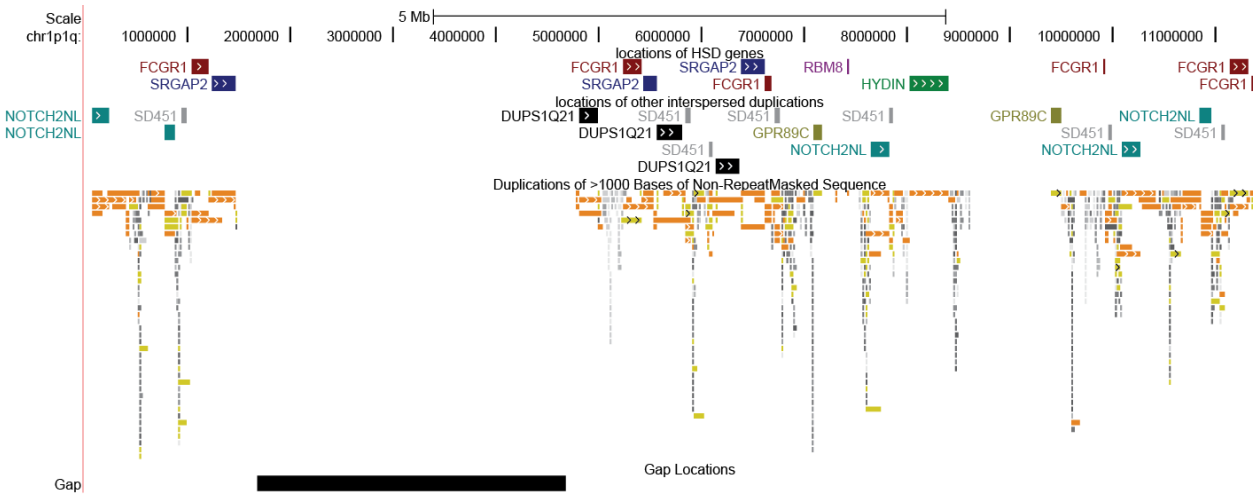


Figure 5.2. HSD Genes at the Chromosome 1q21 Locus. This region contains members of the *SRGAP2*, *HYDIN*, *RBM8*, and *FCGR1* HSD gene families, as well as several other segmental duplications, including many that are interspersed. The gap corresponds to the centromere, with sequence on the left from the p arm.

Perhaps the most important yet least well established component of future studies of HSD genes is functional characterization. Expression analyses have thus far been restricted to the RNA level, so targeted proteomic experiments customized to detect paralog-specific peptides will prove critical for determining whether or not particular HSD paralogs are expressed at the protein level as expected based on the presence long open reading frames. CRISPR/Cas gene editing technology promises to allow targeted disruption of HSD genes in human cell lines [258, 259], enabling the analysis of knockouts *in vitro* in cell culture systems. The generation of knockout mouse mutants lacking a working ancestral copy of particular HSD gene families should shed light on the functions of the ancestral human paralogs, and overexpression of human-specific transcripts in developing mouse embryos can serve to functionally model the human-specific duplicate paralogs [113, 243]. These experiments will be very challenging, as assessing function requires highly specialized assays depending on the particular function of interest, and developmental context is an important factor to account for, as several HSD genes examined thus far show restricted spatiotemporal expression patterns. Nevertheless, experimental functional characterization will spearhead continued investigations of HSD genes in the years and decades to come and will provide

a powerful complement to experiments of nature and more detailed assessment of the potential roles of HSD genes in human disease.

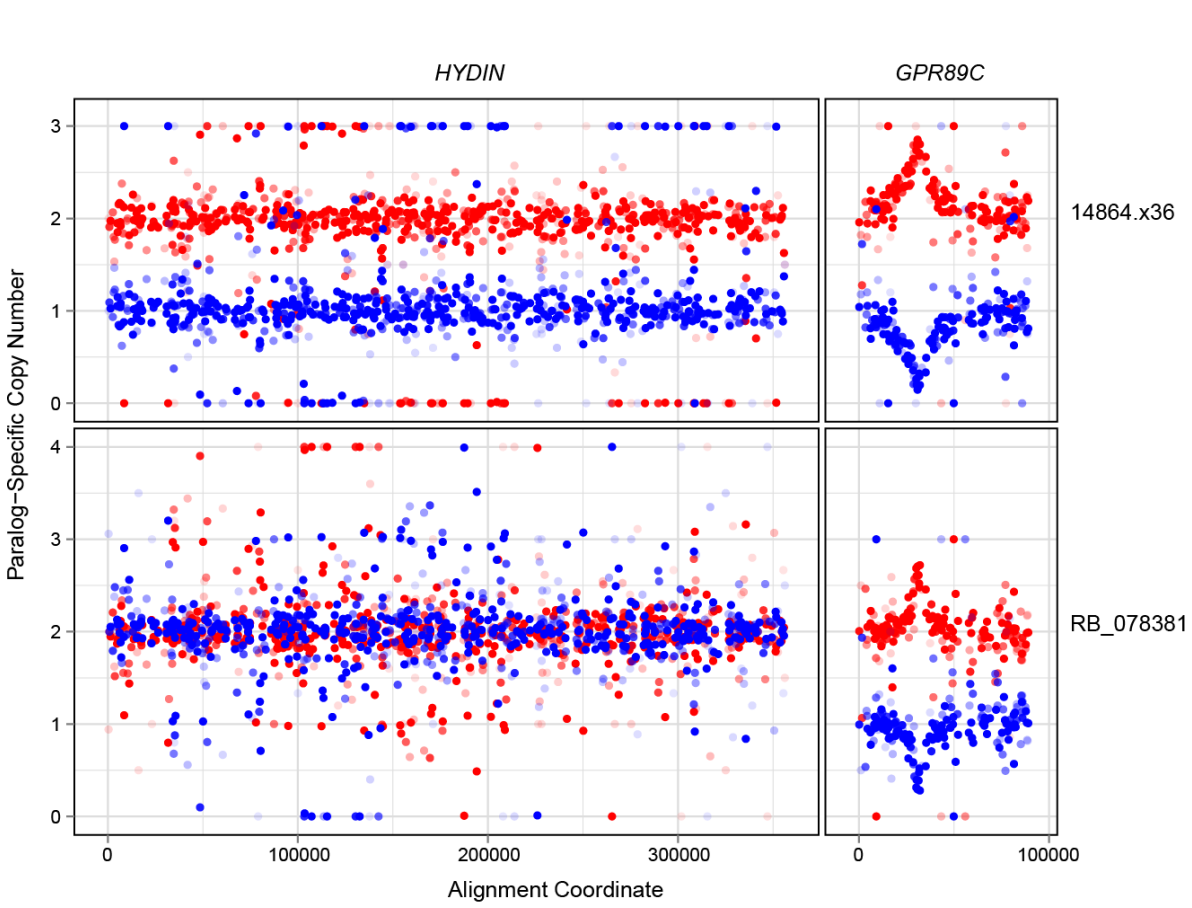


Figure 5.3. 1q21 Rearrangement Breakpoint Variability. a) Data from two microdeletion patients (top and bottom) from 717 MIPs targeting *HYDIN* paralogs and 184 MIPs targeting *GPR89C* paralogs. In contrast to the telomeric *GPR89C* paralog deleted in both cases (blue points, right), *HYDIN2* (blue points, left) is only deleted in one case (top). b) Minimal deletion intervals for both patients are shown above a diagram of segmental duplications in the chromosome 1 pericentromeric region, with locations of *HYDIN* and *GPR89C* paralogs indicated. These data provide the first evidence for breakpoint variability in 1q21 rearrangements and highlight how this variability differentially impacts at least one gene.

The stories of *SRGAP2* and *BOLA2* epitomize the complex interplay between segmental duplication, evolution, and disease. Interspersed duplications have accumulated in our genomes over the recent evolutionary past, predisposing them to NAHR-based rearrangements and associated disease while also giving birth to newly-minted genes [3, 4, 43]. Although much future work will be necessary to understand these genes at the genomic, transcript, protein, and functional levels, exploring these new frontiers will enhance our knowledge of our genome and how it came to be. Beyond the horizon, many surprises and challenges undoubtedly await, but with high-quality genomes as our guides, we may one day finally capture the elusive genetic bases for the most distinctive human traits.

References

1. Ohno, S., *Evolution by Gene Duplication*. 1970, New York: Springer-Verlag.
2. Hughes, A.L., *The evolution of functionally novel proteins after gene duplication.* Proc Biol Sci, 1994. **256**(1346): p. 119-24.
3. Marques-Bonet, T., S. Girirajan, and E.E. Eichler, *The origins and impact of primate segmental duplications.* Trends Genet, 2009. **25**(10): p. 443-54.
4. Bailey, J.A. and E.E. Eichler, *Primate segmental duplications: crucibles of evolution, diversity and disease.* Nat Rev Genet, 2006. **7**(7): p. 552-64.
5. Holland, P.W., et al., *Gene duplications and the origins of vertebrate development.* Dev Suppl, 1994: p. 125-33.
6. Koszul, R., et al., *Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments.* Embo j, 2004. **23**(1): p. 234-43.
7. Lespinet, O., et al., *The role of lineage-specific gene family expansion in the evolution of eukaryotes.* Genome Res, 2002. **12**(7): p. 1048-59.
8. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.
9. Pereira-Leal, J.B., et al., *Evolution of protein complexes by duplication of homomeric interactions.* Genome Biol, 2007. **8**(4): p. R51.
10. Samonte, R.V. and E.E. Eichler, *Segmental duplications and the evolution of the primate genome.* Nat Rev Genet, 2002. **3**(1): p. 65-72.
11. Clark, A.G., *Invasion and maintenance of a gene duplication.* Proc Natl Acad Sci U S A, 1994. **91**(8): p. 2950-4.
12. Prince, V.E. and F.B. Pickett, *Splitting pairs: the diverging fates of duplicated genes.* Nat Rev Genet, 2002. **3**(11): p. 827-37.
13. Walsh, J.B., *How many processed pseudogenes are accumulated in a gene family?* Genetics, 1985. **110**(2): p. 345-64.
14. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
15. Stoltzfus, A., *On the possibility of constructive neutral evolution.* J Mol Evol, 1999. **49**(2): p. 169-81.
16. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**(1): p. 459-73.
17. Watterson, G.A., *On the time for gene silencing at duplicate Loci.* Genetics, 1983. **105**(3): p. 745-66.
18. Lynch, M. and V. Katju, *The altered evolutionary trajectories of gene duplicates.* Trends Genet, 2004. **20**(11): p. 544-9.
19. Espinosa-Cantu, A., et al., *Gene duplication and the evolution of moonlighting proteins.* Front Genet, 2015. **6**: p. 227.
20. Ji, Y., et al., *Structure of chromosomal duplicons and their role in mediating human genomic disorders.* Genome Res, 2000. **10**(5): p. 597-610.
21. Stankiewicz, P., et al., *Serial segmental duplications during primate evolution result in complex human genome architecture.* Genome Res, 2004. **14**(11): p. 2209-20.
22. Stankiewicz, P. and J.R. Lupski, *Genome architecture, rearrangements and genomic disorders.* Trends Genet, 2002. **18**(2): p. 74-82.
23. Wolfe, K.H. and D.C. Shields, *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, 1997. **387**(6634): p. 708-13.
24. Sidow, A., *Gen(om)e duplications in the evolution of early vertebrates.* Curr Opin Genet Dev, 1996. **6**(6): p. 715-22.

25. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-7.
26. Kehrer-Sawatzki, H. and D.N. Cooper, *Structural divergence between the human and chimpanzee genomes.* Hum Genet, 2007. **120**(6): p. 759-78.
27. Murphy, W.J., et al., *Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps.* Science, 2005. **309**(5734): p. 613-7.
28. Weise, A., et al., *New insights into the evolution of chromosome 1.* Cytogenet Genome Res, 2005. **108**(1-3): p. 217-22.
29. Linardopoulou, E.V., et al., *Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication.* Nature, 2005. **437**(7055): p. 94-100.
30. Szamalek, J.M., et al., *Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes.* Hum Genet, 2006. **120**(1): p. 126-38.
31. Thornton, J.W. and R. DeSalle, *Gene family evolution and homology: genomics meets phylogenetics.* Annu Rev Genomics Hum Genet, 2000. **1**: p. 41-73.
32. Cannon, S.B., et al., *The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana.* BMC Plant Biol, 2004. **4**: p. 10.
33. She, X., et al., *Shotgun sequence assembly and recent segmental duplications within the human genome.* Nature, 2004. **431**(7011): p. 927-30.
34. Cheng, Z., et al., *A genome-wide comparison of recent chimpanzee and human segmental duplications.* Nature, 2005. **437**(7055): p. 88-93.
35. Tuzun, E., J.A. Bailey, and E.E. Eichler, *Recent segmental duplications in the working draft assembly of the brown Norway rat.* Genome Res, 2004. **14**(4): p. 493-506.
36. David, L., et al., *Recent duplication of the common carp (Cyprinus carpio L.) genome as revealed by analyses of microsatellite loci.* Mol Biol Evol, 2003. **20**(9): p. 1425-34.
37. Liu, G.E., et al., *Analysis of recent segmental duplications in the bovine genome.* BMC Genomics, 2009. **10**: p. 571.
38. *The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications.* BMC Biol, 2005. **3**: p. 20.
39. Blackman, B.K., et al., *The role of recently derived FT paralogs in sunflower domestication.* Curr Biol, 2010. **20**(7): p. 629-35.
40. Marques-Bonet, T., et al., *A burst of segmental duplications in the genome of the African great ape ancestor.* Nature, 2009. **457**(7231): p. 877-81.
41. Sudmant, P.H., et al., *Evolution and diversity of copy number variation in the great ape lineage.* Genome Res, 2013. **23**(9): p. 1373-82.
42. She, X., et al., *Mouse segmental duplication and copy number variation.* Nat Genet, 2008. **40**(7): p. 909-14.
43. Marques-Bonet, T. and E.E. Eichler, *The evolution of human segmental duplications and the core duplicon hypothesis.* Cold Spring Harb Symp Quant Biol, 2009. **74**: p. 355-62.
44. Lupski, J.R., *Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits.* Trends Genet, 1998. **14**(10): p. 417-22.
45. Sharp, A.J., et al., *Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome.* Nat Genet, 2006. **38**(9): p. 1038-42.
46. Shaw, C.J. and J.R. Lupski, *Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease.* Hum Mol Genet, 2004. **13 Spec No 1**: p. R57-64.
47. Lakich, D., et al., *Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A.* Nat Genet, 1993. **5**(3): p. 236-41.
48. Wakimoto, B.T. and M.G. Hearn, *The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of Drosophila melanogaster.* Genetics, 1990. **125**(1): p. 141-54.

49. Hannes, F.D., et al., *Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant.* J Med Genet, 2009. **46**(4): p. 223-32.
50. Sharp, A.J., et al., *Characterization of a recurrent 15q24 microdeletion syndrome.* Hum Mol Genet, 2007. **16**(5): p. 567-72.
51. Higgs, D.R., et al., *A novel alpha-globin gene arrangement in man.* Nature, 1980. **284**(5757): p. 632-5.
52. Lauer, J., C.K. Shen, and T. Maniatis, *The chromosomal arrangement of human alpha-like globin genes: sequence homology and alpha-globin gene deletions.* Cell, 1980. **20**(1): p. 119-30.
53. Bunge, S., et al., *Homologous nonallelic recombinations between the iduronate-sulfatase gene and pseudogene cause various intragenic deletions and inversions in patients with mucopolysaccharidosis type II.* Eur J Hum Genet, 1998. **6**(5): p. 492-500.
54. Pentao, L., et al., *Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit.* Nat Genet, 1992. **2**(4): p. 292-300.
55. Greenberg, F., et al., *Molecular analysis of the Smith-Magenis syndrome: a possible contiguous-gene syndrome associated with del(17)(p11.2).* Am J Hum Genet, 1991. **49**(6): p. 1207-18.
56. Scambler, P.J., et al., *Velo-cardio-facial syndrome associated with chromosome 22 deletions encompassing the DiGeorge locus.* Lancet, 1992. **339**(8802): p. 1138-9.
57. Bayes, M., et al., *Mutational mechanisms of Williams-Beuren syndrome deletions.* Am J Hum Genet, 2003. **73**(1): p. 131-51.
58. Mefford, H.C., et al., *Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes.* N Engl J Med, 2008. **359**(16): p. 1685-99.
59. Weiss, L.A., et al., *Association between microdeletion and microduplication at 16p11.2 and autism.* N Engl J Med, 2008. **358**(7): p. 667-75.
60. Kumar, R.A., et al., *Recurrent 16p11.2 microdeletions in autism.* Hum Mol Genet, 2008. **17**(4): p. 628-38.
61. Miller, D.T., et al., *Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders.* J Med Genet, 2009. **46**(4): p. 242-8.
62. McCarthy, S.E., et al., *Microduplications of 16p11.2 are associated with schizophrenia.* Nat Genet, 2009. **41**(11): p. 1223-7.
63. Stefansson, H., et al., *Large recurrent microdeletions associated with schizophrenia.* Nature, 2008. **455**(7210): p. 232-6.
64. *Rare chromosomal deletions and duplications increase risk of schizophrenia.* Nature, 2008. **455**(7210): p. 237-41.
65. Brunetti-Pierri, N., et al., *Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities.* Nat Genet, 2008. **40**(12): p. 1466-71.
66. Shinawi, M., et al., *Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size.* J Med Genet, 2010. **47**(5): p. 332-41.
67. Jacquemont, S., et al., *Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus.* Nature, 2011. **478**(7367): p. 97-102.
68. Walters, R.G., et al., *A new highly penetrant form of obesity due to deletions on chromosome 16p11.2.* Nature, 2010. **463**(7281): p. 671-5.
69. She, X., et al., *A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications.* Genome Res, 2006. **16**(5): p. 576-83.
70. Zhang, L., et al., *Patterns of segmental duplication in the human genome.* Mol Biol Evol, 2005. **22**(1): p. 135-41.
71. Ciccarelli, F.D., et al., *Complex genomic rearrangements lead to novel primate gene function.* Genome Res, 2005. **15**(3): p. 343-51.

72. Johnson, M.E., et al., *Positive selection of a gene family during the emergence of humans and African apes.* Nature, 2001. **413**(6855): p. 514-9.
73. Han, M.V., et al., *Adaptive evolution of young gene duplicates in mammals.* Genome Res, 2009. **19**(5): p. 859-67.
74. *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
75. Birtle, Z., L. Goodstadt, and C. Ponting, *Duplication and positive selection among hominin-specific PRAME genes.* BMC Genomics, 2005. **6**: p. 120.
76. Semple, C.A., M. Rolfe, and J.R. Dorin, *Duplication and selection in the evolution of primate beta-defensin genes.* Genome Biol, 2003. **4**(5): p. R31.
77. Eichler, E.E., *Recent duplication, domain accretion and the dynamic mutation of the human genome.* Trends Genet, 2001. **17**(11): p. 661-9.
78. Jiang, Z., et al., *Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.* Nat Genet, 2007. **39**(11): p. 1361-8.
79. Jobling, M., M. Hurles, and C. Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples & Disease*. 2004, New York: Garland Science.
80. Liebermann, P., *The Biology and Evolution of Language*. 1984, Cambridge, Massachusetts: Harvard Univ. Press.
81. Hill, W.C.O., *Note on the male external genitalia of the chimpanzee.* Proc Zool Soc London, 1946. **116**(1): p. 129-132.
82. Dixson, A.F., *Primate Sexuality: Comparative Studies of the Prosimians, Monkeys, Apes, and Human Beings*. 1998, Oxford, UK: Oxford Univ. Press.
83. Wheeler, P.E., *The evolution of bipedality and loss of functional body hair in hominids.* J Hum Evol, 1984. **13**(1): p. 91-98.
84. Wheeler, P.E., *The influence of the loss of functional body hair on the water budgets of early hominids.* J Hum Evol, 1992. **23**(5): p. 379-388.
85. Wheeler, P.E., *The influence of bipedalism on the energy and water budgets of early hominids.* J Hum Evol, 1991. **21**(2): p. 117-136.
86. Teaford, M.F. and P.S. Ungar, *Diet and the evolution of the earliest human ancestors.* Proc Natl Acad Sci U S A, 2000. **97**(25): p. 13506-11.
87. Shellis, R.P., et al., *Variations in molar enamel thickness among primates.* J Hum Evol, 1998. **35**(4-5): p. 507-22.
88. Emes, Y., B. Aybar, and S. Yalcin, *On the evolution of human jaws and teeth: a review.* Bull Int Assoc Paleodont, 2011. **5**(1): p. 37-47.
89. Kono, R.T., *Molar enamel thickness and distribution patterns in extant great apes and humans: new insights based on a 3-dimensional whole crown perspective.* J Anthropol Sci, 2004. **112**(2): p. 121-146.
90. Dehay, C. and H. Kennedy, *Cell-cycle control and cortical development.* Nat Rev Neurosci, 2007. **8**(6): p. 438-50.
91. Rakic, P., *Evolution of the neocortex: a perspective from developmental biology.* Nat Rev Neurosci, 2009. **10**(10): p. 724-35.
92. Sidman, R.L. and P. Rakic, *Neuronal migration, with special reference to developing human brain: a review.* Brain Res, 1973. **62**(1): p. 1-35.
93. Huttenlocher, P.R. and A.S. Dabholkar, *Regional differences in synaptogenesis in human cerebral cortex.* J Comp Neurol, 1997. **387**(2): p. 167-78.
94. Petanjek, Z., et al., *Extraordinary neoteny of synaptic spines in the human prefrontal cortex.* Proc Natl Acad Sci U S A, 2011. **108**(32): p. 13281-6.
95. Liu, X., et al., *Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques.* Genome Res, 2012. **22**(4): p. 611-22.
96. Benavides-Piccione, R., et al., *Cortical area and species differences in dendritic spine morphology.* J Neurocytol, 2002. **31**(3-5): p. 337-46.

97. Elston, G.N., R. Benavides-Piccione, and J. DeFelipe, *The pyramidal cell in cognition: a comparative study in human and monkey.* J Neurosci, 2001. **21**(17): p. Rc163.
98. Dennis, M.Y., et al., *Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.* Cell, 2012. **149**(4): p. 912-22.
99. Chou, H.H., et al., *A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence.* Proc Natl Acad Sci U S A, 1998. **95**(20): p. 11751-6.
100. Irie, A., et al., *The molecular basis for the absence of N-glycolylneuraminic acid in humans.* J Biol Chem, 1998. **273**(25): p. 15866-71.
101. Stedman, H.H., et al., *Myosin gene mutation correlates with anatomical changes in the human lineage.* Nature, 2004. **428**(6981): p. 415-8.
102. McCollum, M.A., et al., *Of muscle-bound crania and human brain evolution: the story behind the MYH16 headlines.* J Hum Evol, 2006. **50**(2): p. 232-6.
103. Huby, T., et al., *Functional analysis of the chimpanzee and human apo(a) promoter sequences: identification of sequence variations responsible for elevated transcriptional activity in chimpanzee.* J Biol Chem, 2001. **276**(25): p. 22209-14.
104. Evans, P.D., et al., *Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage.* Proc Natl Acad Sci U S A, 2006. **103**(48): p. 18178-83.
105. Mekel-Bobrov, N., et al., *Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens.* Science, 2005. **309**(5741): p. 1720-2.
106. Currat, M., et al., *Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans".* Science, 2006. **313**(5784): p. 172; author reply 172.
107. Yu, F., et al., *Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens".* Science, 2007. **316**(5823): p. 370.
108. Woods, R.P., et al., *Normal variants of Microcephalin and ASPM do not account for brain size variability.* Hum Mol Genet, 2006. **15**(12): p. 2025-9.
109. Enard, W., et al., *Molecular evolution of FOXP2, a gene involved in speech and language.* Nature, 2002. **418**(6900): p. 869-72.
110. Prabhakar, S., et al., *Human-specific gain of function in a developmental enhancer.* Science, 2008. **321**(5894): p. 1346-50.
111. Rockman, M.V., et al., *Ancient and recent positive selection transformed opioid cis-regulation in humans.* PLoS Biol, 2005. **3**(12): p. e387.
112. McLean, C.Y., et al., *Human-specific loss of regulatory DNA and the evolution of human-specific traits.* Nature, 2011. **471**(7337): p. 216-9.
113. Charrier, C., et al., *Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation.* Cell, 2012. **149**(4): p. 923-35.
114. Vallender, E.J., N. Mekel-Bobrov, and B.T. Lahn, *Genetic basis of human brain evolution.* Trends Neurosci, 2008. **31**(12): p. 637-44.
115. Varki, A., D.H. Geschwind, and E.E. Eichler, *Explaining human uniqueness: genome interactions with environment, behaviour and culture.* Nat Rev Genet, 2008. **9**(10): p. 749-63.
116. Boyd, J.L., et al., *Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex.* Curr Biol, 2015. **25**(6): p. 772-9.
117. Prufer, K., et al., *The bonobo genome compared with the chimpanzee and human genomes.* Nature, 2012. **486**(7404): p. 527-31.
118. Scally, A., et al., *Insights into hominid evolution from the gorilla genome sequence.* Nature, 2012. **483**(7388): p. 169-75.
119. Locke, D.P., et al., *Comparative and demographic analysis of orang-utan genomes.* Nature, 2011. **469**(7331): p. 529-33.
120. Fortna, A., et al., *Lineage-specific gene duplication and loss in human and great ape evolution.* PLoS Biol, 2004. **2**(7): p. E207.

121. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes.* Science, 2010. **330**(6004): p. 641-6.
122. Eichler, E.E., *Segmental duplications: what's missing, misassigned, and misassembled--and should we care?* Genome Res, 2001. **11**(5): p. 653-6.
123. Doggett, N.A., et al., *A 360-kb interchromosomal duplication of the human HYDIN locus.* Genomics, 2006. **88**(6): p. 762-71.
124. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.
125. Locke, D.P., et al., *Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.* Am J Hum Genet, 2006. **79**(2): p. 275-90.
126. Zhang, Y.E., et al., *Accelerated recruitment of new brain development genes into the human genome.* PLoS Biol, 2011. **9**(10): p. e1001179.
127. Guerrier, S., et al., *The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis.* Cell, 2009. **138**(5): p. 990-1004.
128. Chen, L.T., A.G. Gilman, and T. Kozasa, *A candidate target for G protein action in brain.* J Biol Chem, 1999. **274**(38): p. 26931-8.
129. Davy, B.E. and M.L. Robinson, *Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in Hydin, a large novel gene.* Hum Mol Genet, 2003. **12**(10): p. 1163-70.
130. Guo, S. and S. Bao, *srGAP2 arginine methylation regulates cell migration and cell spreading through promoting dimerization.* J Biol Chem, 2010. **285**(45): p. 35133-41.
131. Kajii, T. and K. Ohama, *Androgenetic origin of hydatidiform mole.* Nature, 1977. **268**(5621): p. 633-4.
132. Patterson, N., et al., *Genetic evidence for complex speciation of humans and chimpanzees.* Nature, 2006. **441**(7097): p. 1103-8.
133. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
134. Fan, J.B., et al., *Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping.* Genomics, 2002. **79**(1): p. 58-62.
135. Gregory, S.G., et al., *The DNA sequence and biological annotation of human chromosome 1.* Nature, 2006. **441**(7091): p. 315-21.
136. Bailey, J.A., G. Liu, and E.E. Eichler, *An Alu transposition model for the origin and expansion of human segmental duplications.* Am J Hum Genet, 2003. **73**(4): p. 823-34.
137. Zhou, Y. and B. Mishra, *Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling.* Proc Natl Acad Sci U S A, 2005. **102**(11): p. 4051-6.
138. Brunet, M., et al., *A new hominid from the Upper Miocene of Chad, Central Africa.* Nature, 2002. **418**(6894): p. 145-51.
139. Brunet, M., et al., *New material of the earliest hominid from the Upper Miocene of Chad.* Nature, 2005. **434**(7034): p. 752-5.
140. Vignaud, P., et al., *Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad.* Nature, 2002. **418**(6894): p. 152-5.
141. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
142. Parsons, J.D., *Miropeats: graphical DNA sequence comparisons.* Comput Appl Biosci, 1995. **11**(6): p. 615-9.
143. Cooper, G.M., et al., *A copy number variation morbidity map of developmental delay.* Nat Genet, 2011. **43**(9): p. 838-46.
144. Saitsu, H., et al., *Early infantile epileptic encephalopathy associated with the disrupted gene encoding Slit-Robo Rho GTPase activating protein 2 (SRGAP2).* Am J Med Genet A, 2012. **158a**(1): p. 199-205.

145. Fischbach, G.D. and C. Lord, *The Simons Simplex Collection: a resource for identification of autism genetic risk factors.* Neuron, 2010. **68**(2): p. 192-5.
146. Geschwind, D.H., et al., *The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions.* Am J Hum Genet, 2001. **69**(2): p. 463-6.
147. Biesecker, L.G., et al., *The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine.* Genome Res, 2009. **19**(9): p. 1665-74.
148. Sharp, A.J., et al., *Segmental duplications and copy-number variation in the human genome.* Am J Hum Genet, 2005. **77**(1): p. 78-88.
149. Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. **328**(5979): p. 710-22.
150. Reich, D., et al., *Genetic history of an archaic hominin group from Denisova Cave in Siberia.* Nature, 2010. **468**(7327): p. 1053-60.
151. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nat Rev Genet, 2010. **11**(6): p. 446-50.
152. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
153. Dai, L., et al., *Is it Williams syndrome? GTF2IRD1 implicated in visual-spatial construction and GTF2I in sociability revealed by high resolution arrays.* Am J Med Genet A, 2009. **149a**(3): p. 302-14.
154. Wu, D.D., D.M. Irwin, and Y.P. Zhang, *De novo origin of human protein-coding genes.* PLoS Genet, 2011. **7**(11): p. e1002379.
155. Antonacci, F., et al., *A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk.* Nat Genet, 2010. **42**(9): p. 745-50.
156. Osoegawa, K., et al., *An improved approach for construction of bacterial artificial chromosome libraries.* Genomics, 1998. **52**(1): p. 1-8.
157. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
158. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2.3.
159. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
160. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
161. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J Mol Evol, 1980. **16**(2): p. 111-20.
162. Dopazo, J., *Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach.* J Mol Evol, 1994. **38**(3): p. 300-4.
163. Rzhetsky, A. and M. Nei, *METREE: a program package for inferring and testing minimum-evolution trees.* Comput Appl Biosci, 1994. **10**(4): p. 409-12.
164. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
165. Blekhman, R., et al., *Sex-specific and lineage-specific alternative splicing in primates.* Genome Res, 2010. **20**(2): p. 180-9.
166. Liu, S., et al., *A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species.* Nucleic Acids Res, 2011. **39**(2): p. 578-88.
167. Campbell, C.D., et al., *Population-genetic properties of differentiated human copy-number polymorphisms.* Am J Hum Genet, 2011. **88**(3): p. 317-32.
168. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
169. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.

170. Lefebvre, S., et al., *Identification and characterization of a spinal muscular atrophy-determining gene.* Cell, 1995. **80**(1): p. 155-65.
171. Olbrich, H., et al., *Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry.* Am J Hum Genet, 2012. **91**(4): p. 672-84.
172. McCarroll, S.A. and D.M. Altshuler, *Copy-number variation and association studies of human disease.* Nat Genet, 2007. **39**(7 Suppl): p. S37-42.
173. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
174. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.* Science, 2005. **307**(5714): p. 1434-40.
175. Bhattacharya, T., et al., *CCL3L1 and HIV/AIDS susceptibility.* Nat Med, 2009. **15**(10): p. 1112-5.
176. Cantsilieris, S., P.N. Baird, and S.J. White, *Molecular methods for genotyping complex copy number polymorphisms.* Genomics, 2013. **101**(2): p. 86-93.
177. Armour, J.A., et al., *Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats.* Nucleic Acids Res, 2007. **35**(3): p. e19.
178. Schouten, J.P., et al., *Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.* Nucleic Acids Res, 2002. **30**(12): p. e57.
179. Armour, J.A., et al., *Measurement of locus copy number by hybridisation with amplifiable probes.* Nucleic Acids Res, 2000. **28**(2): p. 605-9.
180. Waszak, S.M., et al., *Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity.* PLoS Comput Biol, 2010. **6**(11): p. e1000988.
181. Hardenbol, P., et al., *Multiplexed genotyping with sequence-tagged molecular inversion probes.* Nat Biotechnol, 2003. **21**(6): p. 673-8.
182. Hardenbol, P., et al., *Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay.* Genome Res, 2005. **15**(2): p. 269-75.
183. Porreca, G.J., et al., *Multiplex amplification of large sets of human exons.* Nat Methods, 2007. **4**(11): p. 931-6.
184. Turner, E.H., et al., *Massively parallel exon capture and library-free resequencing across 16 genomes.* Nat Methods, 2009. **6**(5): p. 315-6.
185. Colin, Y., et al., *Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis.* Blood, 1991. **78**(10): p. 2747-52.
186. Wagner, F.F. and W.A. Flegel, *RHD gene deletion occurred in the Rhesus box.* Blood, 2000. **95**(12): p. 3662-8.
187. Kitano, T. and N. Saitou, *Evolution of Rh blood group genes have experienced gene conversions and positive selection.* J Mol Evol, 1999. **49**(5): p. 615-26.
188. O'Roak, B.J., et al., *Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders.* Science, 2012. **338**(6114): p. 1619-22.
189. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
190. Lee, J.A., C.M. Carvalho, and J.R. Lupski, *A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.* Cell, 2007. **131**(7): p. 1235-47.
191. Zhang, F., et al., *The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans.* Nat Genet, 2009. **41**(7): p. 849-53.
192. Fledel-Alon, A., et al., *Broad-scale recombination patterns underlying proper disjunction in humans.* PLoS Genet, 2009. **5**(9): p. e1000658.
193. Carritt, B., T.J. Kemp, and M. Poulter, *Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled.* Hum Mol Genet, 1997. **6**(6): p. 843-50.
194. Edwards, M.C. and R.A. Gibbs, *Multiplex PCR: advantages, development, and applications.* PCR Methods Appl, 1994. **3**(4): p. S65-75.

195. Markoulatos, P., N. Siafakas, and M. Moncany, *Multiplex polymerase chain reaction: a practical approach.* J Clin Lab Anal, 2002. **16**(1): p. 47-51.
196. Groth, M., et al., *High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes.* Hum Mutat, 2008. **29**(10): p. 1247-54.
197. Aldhous, M.C., et al., *Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease.* Hum Mol Genet, 2010. **19**(24): p. 4930-8.
198. Fernando, M.M., et al., *Assessment of complement C4 gene copy number using the paralog ratio test.* Hum Mutat, 2010. **31**(7): p. 866-74.
199. Hiatt, J.B., et al., *Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation.* Genome Res, 2013. **23**(5): p. 843-54.
200. Itsara, A., et al., *Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing.* Am J Hum Genet, 2012. **90**(4): p. 599-613.
201. Jackson, M.S., et al., *Evidence for widespread reticulate evolution within human duplicons.* Am J Hum Genet, 2005. **77**(5): p. 824-40.
202. Schildkraut, E., C.A. Miller, and J.A. Nickoloff, *Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats.* Nucleic Acids Res, 2005. **33**(5): p. 1574-80.
203. Ezawa, K., O.O. S, and N. Saitou, *Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Genome-wide search of gene conversions in duplicated genes of mouse and rat.* Mol Biol Evol, 2006. **23**(5): p. 927-40.
204. Chen, J.M., et al., *Gene conversion: mechanisms, evolution and human disease.* Nat Rev Genet, 2007. **8**(10): p. 762-75.
205. Zufferey, F., et al., *A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders.* J Med Genet, 2012. **49**(10): p. 660-8.
206. Kumar, R.A., et al., *Association and mutation analyses of 16p11.2 autism candidate genes.* PLoS One, 2009. **4**(2): p. e4582.
207. Konyukh, M., et al., *Variations of the candidate SEZ6L2 gene on Chromosome 16p11.2 in patients with autism spectrum disorders and in human populations.* PLoS One, 2011. **6**(3): p. e17289.
208. Golzio, C., et al., *KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant.* Nature, 2012. **485**(7398): p. 363-7.
209. Migliavacca, E., et al., *A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology.* Am J Hum Genet, 2015. **96**(5): p. 784-96.
210. Blumenthal, I., et al., *Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families.* Am J Hum Genet, 2014. **94**(6): p. 870-83.
211. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.
212. Eichler, E.E. and A.W. Zimmerman, *A hot spot of genetic instability in autism.* N Engl J Med, 2008. **358**(7): p. 737-9.
213. Huddleston, J., et al., *Reconstructing complex regions of genomes using long-read sequencing technology.* Genome Res, 2014. **24**(4): p. 688-96.
214. Chin, C.S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.* Nat Methods, 2013. **10**(6): p. 563-9.
215. Gonzalez, J.R., et al., *A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity.* Am J Hum Genet, 2014. **94**(3): p. 361-72.
216. Martin, J., et al., *The sequence and analysis of duplication-rich human chromosome 16.* Nature, 2004. **432**(7020): p. 988-94.
217. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

218. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
219. Sudmant, P.H., et al., *Global diversity, population stratification, and selection of human copy-number variation.* Science, 2015. **349**(6253): p. aab3761.
220. Prado-Martinez, J., et al., *Great ape genetic diversity and population history.* Nature, 2013. **499**(7459): p. 471-5.
221. Fu, Q., et al., *Genome sequence of a 45,000-year-old modern human from western Siberia.* Nature, 2014. **514**(7523): p. 445-9.
222. Lazaridis, I., et al., *Ancient human genomes suggest three ancestral populations for present-day Europeans.* Nature, 2014. **513**(7518): p. 409-13.
223. Prufer, K., et al., *The complete genome sequence of a Neanderthal from the Altai Mountains.* Nature, 2014. **505**(7481): p. 43-9.
224. Meyer, M., et al., *A high-coverage genome sequence from an archaic Denisovan individual.* Science, 2012. **338**(6104): p. 222-6.
225. Nuttle, X., et al., *Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions.* Nat Methods, 2013. **10**(9): p. 903-9.
226. Schuster, S.C., et al., *Complete Khoisan and Bantu genomes from southern Africa.* Nature, 2010. **463**(7283): p. 943-7.
227. Jakobsson, M., et al., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. **451**(7181): p. 998-1003.
228. Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): p. 1100-4.
229. Tishkoff, S.A., et al., *The genetic structure and history of Africans and African Americans.* Science, 2009. **324**(5930): p. 1035-44.
230. Yang, M.A., K. Harris, and M. Slatkin, *The projection of a test genome onto a reference population and applications to humans and archaic hominins.* Genetics, 2014. **198**(4): p. 1655-70.
231. Hudson, R.R., *Generating samples under a Wright-Fisher neutral model of genetic variation.* Bioinformatics, 2002. **18**(2): p. 337-8.
232. Ewing, G. and J. Hermisson, *MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.* Bioinformatics, 2010. **26**(16): p. 2064-5.
233. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
234. Li, H., et al., *Human glutaredoxin 3 forms [2Fe-2S]-bridged complexes with human BolA2.* Biochemistry, 2012. **51**(8): p. 1687-96.
235. Witte, S., et al., *Inhibition of the c-Jun N-terminal kinase/AP-1 and NF-kappaB pathways by PICOT, a novel protein kinase C-interacting protein with a thioredoxin homology domain.* J Biol Chem, 2000. **275**(3): p. 1902-9.
236. Zhou, Y.B., et al., *hBolA, novel non-classical secreted proteins, belonging to different BolA family with functional divergence.* Mol Cell Biochem, 2008. **317**(1-2): p. 61-8.
237. Marchetto, M.C., et al., *Differential L1 regulation in pluripotent stem cells of humans and apes.* Nature, 2013. **503**(7477): p. 525-9.
238. Nuttle, X., et al., *Resolving genomic disorder-associated breakpoints within segmental DNA duplications using massively parallel sequencing.* Nat Protoc, 2014. **9**(6): p. 1496-513.
239. *Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders.* Neuron, 2012. **73**(6): p. 1063-7.
240. Antonacci, F., et al., *Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability.* Nat Genet, 2014. **46**(12): p. 1293-302.
241. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease.* Annu Rev Med, 2010. **61**: p. 437-55.

242. Coe, B.P., et al., *Refining analyses of copy number variation identifies specific genes associated with developmental delay.* Nat Genet, 2014. **46**(10): p. 1063-71.
243. Florio, M., et al., *Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion.* Science, 2015. **347**(6229): p. 1465-70.
244. Stefansson, H., et al., *CNVs conferring risk of autism or schizophrenia affect cognition in controls.* Nature, 2014. **505**(7483): p. 361-6.
245. Reiter, L.T., et al., *Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients.* Am J Hum Genet, 1998. **62**(5): p. 1023-33.
246. Johnson, E.E. and M. Wessling-Resnick, *Iron metabolism and the innate immune response to infection.* Microbes Infect, 2012. **14**(3): p. 207-16.
247. Ward, R.J., et al., *Iron and the immune system.* J Neural Transm (Vienna), 2011. **118**(3): p. 315-28.
248. Collard, K.J., *Iron homeostasis in the neonate.* Pediatrics, 2009. **123**(4): p. 1208-16.
249. Cunnane, S.C., *[Survival of the fattest: the key to human brain evolution]*. Med Sci (Paris), 2006. **22**(6-7): p. 659-63.
250. Cunnane, S.C. and M.A. Crawford, *Energetic and nutritional constraints on infant brain development: implications for brain expansion during human evolution.* J Hum Evol, 2014. **77**: p. 88-98.
251. Milton, K., *Micronutrient intakes of wild primates: are humans different?* Comp Biochem Physiol A Mol Integr Physiol, 2003. **136**(1): p. 47-59.
252. Bray, N., et al. *Near-optimal RNA-Seq quantification*. arXiv, 2015. 1505.02710.
253. Chaisson, M.J., et al., *Resolving the complexity of the human genome using single-molecule sequencing.* Nature, 2015. **517**(7536): p. 608-11.
254. Yong, E., *A duplicated gene shaped human brain evolution…and why the genome project missed it*, in *Not Exactly Rocket Science*. 2012.
255. King, M.C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-16.
256. Zody, M.C., et al., *Evolutionary toggling of the MAPT 17q21.31 inversion region.* Nat Genet, 2008. **40**(9): p. 1076-83.
257. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science, 2012. **335**(6070): p. 823-8.
258. Yang, L., et al., *Genome-wide inactivation of porcine endogenous retroviruses (PERVs).* Science, 2015. **350**(6264): p. 1101-4.
259. Hartenian, E. and J.G. Doench, *Genetic screens and functional genomics using CRISPR/Cas9 technology*. Febs j, 2015. **282**(8): p. 1383-93.

Appendix A. Supplemental Information for Chapter 2

Extended Experimental Procedures

FISH Experiments

FISH experiments were used to detect *SRGAP2* paralogous regions (probes 1 and 9, see Table S1), resolve the chromosomal orientation of contigs (probes 2–7), infer the evolutionary order of duplication events (probes 1, 2, 7, and 8), and validate copy number polymorphism at the *SRGAP2B* locus (probe 1). Experiments were performed using clones obtained from a G248 fosmid library [1], directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (PerkinElmer), and fluorescein-dUTP (Enzo), as previously described [2] 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 sulfate, and 3 μg sonicated salmon sperm DNA, in a volume of 10 ml. 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, Cy3, Cy5, and fluorescein fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

We used a series of FISH assays to determine the chromosomal orientation (Figure S1) and evolutionary order (Figure 2C) of *SRGAP2* duplications. Because our contigs did not map concordantly to the human reference genome and spanned multiple gaps, we could not confidently infer the chromosomal orientation of our contigs based on comparison to the reference sequence. We performed three-color interphase FISH assays to resolve this. These assays made use of two probes mapping within our sequenced contig (or extensions of our contig based on contiguous reference sequence) and one probe mapping outside of our contig, closer to the telomere. FISH analyses with probes targeting regions upstream of the *SRGAP2B* paralog indicated that this sequence is extensively locally duplicated (see yellow probe in Figure 2C—even though the region it targets is deleted in our *SRGAP2B* contig, this sequence is still present in two haploid copies at chromosome 1q21.1). Because FISH analysis could not resolve the chromosomal orientation of *SRGAP2B* or *SRGAP2D*, we instead utilized an anchored contig spanning the entire 1q21.1 region recently generated at The Genome Institute at Washington University School of Medicine (T.G. and R.W., unpublished data). Comparison of our *SRGAP2B* contig showed that this paralog is transcribed toward the centromere, whereas the *SRGAP2D* orientation remains uncertain, as the contig containing this paralog has not yet been anchored.

Generation of Paralog-Specific Sequence Contigs

Due to the highly identical nature of sequences within segmental duplications, many genes embedded within duplicated regions are not properly represented in the human reference genome [3]. To generate and distinguish sequences corresponding to *SRGAP2* paralogs, we leveraged a large-insert BAC library (CHORI-17) generated from a well-characterized complete hydatidiform mole cell line (CHM1hTERT), a resource developed to resolve paralogous regions of the genome (http://www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf). Clones were selected for sequencing based on mapping of BAC-end sequences to partial copies of *SRGAP2* within the human reference [4]. Selected clones included both concordant and discordant clones as well as clones with only one end mapped to one of the *SRGAP2* paralogous regions (*SRGAP2A*, chr1:204,582,823-204,704,406; *SRGAP2B*, chr1:142,625,200-142,805,783; *SRGAP2C*, chr1:120,637,333-120,832,584; NCBI36/hg18). *SRGAP2*-containing discordant clones were validated by PCR amplification using primers targeting either intron 1 or intron 2 of *SRGAP2*. Additional clones from the CH17-BAC library were also obtained by mapping efforts at The Genome Institute at Washington University School of Medicine directed to improving the quality of the human reference genome (T.G. and R.W., unpublished data).

Clone inserts were completely sequenced using a hierarchical clone-based strategy with high-quality capillary fluorescent-based sequencing. This entailed the construction of genomic libraries, sequencing of paired-end shotgun libraries, and assembly of inserts into a finished sequencing contig for 22 distinct CH17 clones [see Table S5 with clone IDs (ordered as they assemble into contigs) and GenBank accessions]. We used sequence quality standards for sequencing and assembly commensurate to that applied to human genome reference (estimated 1 error in 100,000 bases) [5].

The *SRGAP2* coding exons were distinguished in all clones, and variants specific to each paralog were identified. Each BAC was assigned to a specific duplicated region by performing BLAST sequence similarity searches of *SRGAP2* exons against each BAC clone sequence. Coding sequence substitutions served as features that distinguished clones originating from different *SRGAP2* paralogs. We used these paralog-specific clones as a query in a BLAST search of the human HTGS (high-throughput genomic sequence) database to identify clones from the hydatidiform mole BAC library (CH17). Clones with >99.9% sequence identity could be confidently inferred to have originated from the paralog containing that particular substitution (as opposed to a sequence identity of <99.5% observed for clones mapping to other *SRGAP2* loci). Clones thus inferred to have originated from the same paralog containing overlapping sequence were combined into contigs, and the entire process was repeated iteratively using the clones at the ends to extend overlaps. The iterations were carried out until no more hydatidiform mole clone sequences could be confidently inferred to have originated from the paralog represented by the contig. The contigs were assembled using Sequencher 4.9. Alignment quality was ensured by manual inspection and editing when necessary. This method allowed us to ultimately generate three single-haplotype contigs containing the sequences of three *SRGAP2* paralogs and their flanking sequences.

Comparison of Sequences from *SRGAP2* Contigs with the Human Reference Genome

We utilized pairwise BLAST [6] to identify missing or mismapped regions within the human reference genome. Comparing shared sequence between contigs revealed high sequence identity between *SRGAP2* paralogs (99%–99.75%) with the *SRGAP2B* and *SRGAP2D* paralogs as the most highly identical. Based on these identities, we performed pairwise BLAST alignments of our complete contigs against the human reference genome (GRCh37/hg19) and identified extended contiguous regions of the reference (at least 5,000 bp in length) having >99.6% sequence identity to a particular contig. These regions were mapped back to their respective contigs allowing us to identify any sequences missing or mismapped within the reference.

Breakpoint Analysis of *SRGAP2* Duplicated Regions

Pairwise BLAST was used to identify extended contiguous regions of high sequence identity between every pair of paralogous sequence contigs. We observe that the original duplication event (258,245 bp) encompassed the promoter and first nine exons of *SRGAP2*. A larger, second duplication event (>515 kbp) originated from the daughter *SRGAP2* paralog and included the entire original duplicated sequence. Subsequently, two large deletions (102,605 bp and 48,969 bp) occurred upstream of *SRGAP2B*. The nucleotide sequences at the ends of these regions were recorded. The contigs were then aligned at these sequences using Sequencher 4.9 and manually checked, revealing breakpoints between *SRGAP2* paralogs and breakpoints due to other structural rearrangements at a high resolution (in most cases, single-nucleotide resolution). The local sequences surrounding these breakpoints were then assessed for the presence of repetitive elements using RepeatMasker [7] and via a BLAST search against a database of human Alu repeats. To gain a better understanding of the content within the deleted regions upstream of *SRGAP2B*, we assessed the genes and potential regulatory elements residing within these deleted regions. Notably, only portions of the deleted regions are represented in the most current human reference (smaller deletion, chr1:144,275,483-144,312,889; larger deletion, chr1:120,872,120-120,936,069; 1p12-contig region represented in GRCh37/hg19). The smaller deletion (49 kbp) resides 195 kbp upstream of *SRGAP2* and contains two uncharacterized genes. The larger deletion (103 kbp) resides 34 kbp upstream of *SRGAP2*, just downstream of *FAM72*. The deleted region would have contained paralogs to *HIST72H2BA* and *FCGR1B* in addition to putative regulatory elements predicted by

conserved transcription-factor binding site predictions and ChIP-Seq data (using ‘‘Regulation’’ tracks within the UCSC Genome Browser). This deletion does not affect any obvious promoter elements of *SRGAP2* that would reside directly upstream of the transcription start site.

***SRGAP2* Duplication Timing Using a Chromosome 1q32.1 Molecular Clock**

We created a multiple-species alignment (MSA) (ClustalW [8]) of the 244 kbp *SRGAP2* genomic region shared across the three human loci, chimpanzee, and orangutan orthologous regions. The chimpanzee (October 2010) and orangutan (July 2007) orthologous sequences were identified using BLAT in the UCSC Genome Browser with the 1q32.1 (*SRGAP2A*) human sequence as the query. We manually inspected the alignment using the Jalview [9] editor and manually corrected alignment errors. We repeated the multiple sequence alignment using nonhuman primate orthologous segments, inspecting the alignment for errors each time. The final alignment was contiguous for human and chimpanzee sequences, with 12 gaps within the orangutan sequence spanning 12,639 bp of sequence [most gaps were small; two large gaps account for 9,146 bp and 1,523 bp of sequence]. From this MSA, we constructed an unrooted phylogenetic tree using the neighbor-joining method [10] (MEGA [11]; complete-deletion option). Genetic distances were computed using the Kimura two-parameter method [12] with standard error estimates (an interior branch test of phylogeny [13, 14]; n = 500 bootstrap replicates). We noticed that the branch lengths of *SRGAP2B* and *SRGAP2C* were considerably longer (>30%) than *SRGAP2A* suggesting that the rates of substitution at the chromosomal regions 1q21.1 and 1p12 were higher than 1q32.1. Using Tajima’s relative rate test (MEGA), we determined that *SRGAP2A* evolved at the same rate as orthologous counterparts in chimpanzee and orangutan ($p = 0.5345$) while both *SRGAP2B* and *SRGAP2C* evolved at an accelerated rate ($p = 0.0001$-0.0249). Using the genetic distance established for human *SRGAP2A*, we applied a correction factor to the average branch length leading to *SRGAP2B* and *SRGAP2C* in effect forcing the substitution rate of these branches to equal that of chromosome 1q32.1. In turn, we used a chimpanzee divergence time of 6 mya, noting that estimates range from ~5–7 mya since the human and chimpanzee split, based on fossil records [15-17] as well as recent genetic estimates [18], to estimate the timing of the duplication events.

The phylogenetic tree with genetic distances represented as percent of substitutions per total number of aligned sites (244,200 bp) and the standard errors:

(((((((human_*SRGAP2B*: 0.197 ± 0.0102, human_*SRGAP2C*: 0.254 ± 0.0062),:0.092 ± 0.0038), human_*SRGAP2A*:0.237 ± 0.0057),: 0.183 ± 0.0053), chimpanzee: 0.431 ± 0.0077),: 0.833 ± 0.0175), orangutan: 1.36 ± 0.0175).

To account for the increased substitution rate along the *SRGAP2B* and *SRGAP2C* branches (while conservatively leaving the standard error uncorrected), we calculated and applied a correction factor of 0.75:

$D_{SRGAP2B/C} = ½(0.197+0.254)+0.092 = 0.318$
$D_{SRGAP2A} = 0.237$
Correction factor = 0.237/0.318 = 0.75

Corrected phylogenetic tree:

(((((((human_*SRGAP2B*: 0.148 ± 0.0102, human_*SRGAP2C*: 0.191 ± 0.0062),: 0.069 ± 0.0038), human_*SRGAP2A*:0.237 ± 0.0057),:0.183 ± 0.0053), chimpanzee: 0.431 ± 0.0077),:0.833 ± 0.0175), orangutan: 1.36 ± 0.0175).
To estimate the evolutionary timing of the duplication events, we used R = D/2T:

Rate = $[D_{chimpanzee/human,SRGAP2A}]/2T = (0.237+0.183+0.431)/2T = 0.426/T$
$Rate_{T = 6mya}$ = 0.0709% substitutions/site/mya

$Rate_{T = 5mya}$ = 0.0852% substitutions/site/mya
$Rate_{T = 7mya}$ = 0.0609% substitutions/site/mya

From this, we estimated the timing of the initial *SRGAP2* duplication event:

$T = D_{human,SRGAP2A} / R_{T = 6mya} = (0.237)/(0.0709) = 3.4$ mya
$T_{lower} = D_{human,SRGAP2A} / R_{T = 5mya} = (0.237)/(0.0852) = 2.8$ mya
$T_{upper} = D_{human,SRGAP2A} / R_{T = 7mya} = (0.237)/(0.0609) = 3.9$ mya

Likewise, we estimated the timing of the secondary duplication event:

$T_{lower} = D_{human,SRGAP2B/human,SRGAP2C} / 2R_{T = 6mya} = (0.148+0.191)/(2*0.0709) = 2.4$ mya
$T_{lower} = D_{human,SRGAP2B/human,SRGAP2C} / 2R_{T = 5mya} = (0.148+0.191)/(2*0.0852) = 2.0$ mya
$T_{upper} = D_{human,SRGAP2B/human,SRGAP2C} / 2R_{T = 7mya} = (0.148+0.191)/(2*0.0609) = 2.8$ mya

There is error in these divergence rate estimates, but this is minor compared to the inherent error in the chimpanzee-human divergence time estimates. To convey this, we have reported the standard errors of genetic distances in Table 1 of the main text.

Molecular Evolution of 1p12 Genomic Region Distal to *SRGAP2C*

As a control, we estimated the genomic substitution rate for the chromosomal 1p12 region. Specifically, we obtained a 50 kbp MSA using the ENSEMBL genome browser including orthologous sequences from human, chimpanzee, gorilla, orangutan, and rhesus macaque (GRCh37/hg19; chr1:120,193,477-120,253,477). This region maps approximately 2 Mbp distal to the *SRGAP2* duplication region based on the human reference sequence. Tajima's relative rate tests indicated that the sequences were evolving at a constant rate (Bonferroni corrected p = 0.159-1.0). Creating a neighbor-joining phylogenetic tree (as described above) and assuming a human-chimpanzee divergence time of 6 mya, we estimated a chromosome 1p12 substitution rate of $9.38 \pm 1.07 \times 10^{-4}$ substitutions per site per million years (assuming a human-chimpanzee split of 6 mya), which is ~30% higher than that of 1q32.1 ($7.09 \pm 0.183 \times 10^{-4}$ substitutions per site per million years) and consistent with our locus-specific correction factor. Using the standard error of the 1q32.1 and 1p12 estimates, we calculated a range of percent differences between rates.

Molecular Evolution of *SRGAP2D* Genomic Region

Based on partial sequence of clone CH17-248H7, we constructed a smaller 9.5 kbp *SRGAP2* MSA (ClustalW) including sequence from *SRGAP2D*. The tree topology (99% bootstrap support) and sequence identity comparisons both strongly suggest *SRGAP2D* arose via a duplication of the *SRGAP2B* paralog. The increase in substitution rates across the 1p12 and 1q21.1 regions is not evident in our analysis of this much smaller genomic region. Based on an assumption that the rate is indeed accelerated on these duplicate branches, we calculated the timing of this third duplication utilizing the same correction factor as before. The upper and lower estimates of timing for this duplication were estimated using standard error of branch lengths.

Molecular Evolution of the *SRGAP2* Coding Region

We assessed the level of conservation of *SRGAP2* across mammals by estimating its rate of protein evolution in a mammalian phylogeny. Specifically, we assessed the ratio (dN/dS) of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS). Because purifying selection acts to eliminate protein-coding changes, dN/dS decreases with stronger purifying selection. Alternatively, dN/dS increases with relaxed constraint and/or positive selection.

Specifically, we created an MSA of *SRGAP2* coding exons 1 through 9 (shared between all human paralogs except *SRGAP2D* and encoding 452 amino acids) by extracting the exonic regions from the three contigs and using BLAT with exons 1–9 of human *SRGAP2A* as a reference to infer the exons in orthologous sequence from other species [chimpanzee (CGSC 2.1.3/panTro3), orangutan (WUGSC 2.0.2/ponAbe2), macaque (MGSC Merged 1.0/rheMac2), marmoset (WUGSC 3.2/calJac3), and dog (Broad/can-Fam2)]. We also obtained the mRNA sequence of the *SRGAP2* mouse ortholog (GenBank accession BC158055) and used this sequence to infer the rat coding sequence using BLAT against the rat genomic build (Baylor 3.4/rn4). A multiple-sequence alignment was created using ClustalW and Jalview (as before).

This alignment was used in conjunction with the codeml program (part of PAML 4; [19]) to test various models of selection. We estimated the overall dN/dS for the complete tree and compared likelihoods for models that allowed:

(1) free dN/dS for each branch (i.e., lineage heterogeneity);
(2) a primate-specific dN/dS;
(3) a human-specific dN/dS; and
(4) a duplicate-specific dN/dS.

Additionally, we performed tests aimed to detect site-specific signatures of positive selection across phylogeny (branch models):

(1) model 1a (neutral) versus model 2 (positive selection);
(2) model 7 (neutral) versus model 8 (with dN/dS >1); and
(3) model 8a (with dN/dS = 1) versus model 8 (with dN/dS >1).

Segmental Duplication Analysis of *SRGAP2*

To gain insight into the segmental duplication landscape around the *SRGAP2* paralogs, we calculated the percentage of bases classified as duplicated in the cytological bands 1q32.1, 1q21.1, 1p12, and 1p11.2 using the whole-genome assembly comparison (WGAC) on GRCh37/hg19 [20]. Note that *SRGAP2C* maps at the border of 1p12 and 1p11.2 in the reference genome (GRCh37/hg19). The results of this analysis show that chromosomal regions 1q21.1 and 1p11.2 are highly duplicated (63.9% and 69.4% of bases in these regions are classified as duplicated, respectively), whereas 1q32.1 and 1p12 (2.7% and 4.7% duplicated, respectively) are not.

We also sought to assess the duplication status of *SRGAP2* in mammals including chimpanzee, gorilla, orangutan, macaque, marmoset, gibbon, cow, dog, elephant, mouse, and rat. For chimpanzee, gorilla, and orangutan, we generated copy number heatmaps for several individuals using our recently described approach [21] and found no evidence of duplication or copy number variation at *SRGAP2*. Furthermore, characterizing segmental duplications by performing WSSD [22] for human, chimpanzee, gorilla, and orangutan validated this result. For macaque, marmoset, gibbon, cow, dog, elephant, and mouse, we performed WSSD against the corresponding 1q32.1 *SRGAP2* region in each species and again showed no evidence of segmental duplication at this locus. Finally, for rat we examined the segmental duplication WGAC track at the *SRGAP2* locus—again, this region was not duplicated. These results show that *SRGAP2* is duplicated specifically in the human lineage.

Characterization of *SRGAP2* Copy Number Variation Using Sequencing Data

We analyzed high-throughput Illumina shotgun sequence data from 661 individuals from 14 diverse human populations sequenced in Phase 2 of the 1000 Genomes Project in addition to nonhuman Homo species, Neanderthal, and Denisova. We applied our copy number genotyping method [21] to determine aggregate copy number for 1000 bp windows of unmasked sequence spanning the *SRGAP2A*, *SRGAP2B*, and *SRGAP2C* loci (Figures S4A–4C). In marked contrast to other nonhuman primates, duplications encompassing the promoter and first nine exons of *SRGAP2A* (the ancestral *SRGAP2* locus)

were present in all *Homo* species analyzed. Additionally, we noticed that two regions (~95 kbp and ~25 kbp) of *SRGAP2* duplicated sequence consistently showed a predicted copy number of eight total diploid copies in most, but not all individuals (see Figures S1A and S4A–4C). Using the aggregate copy number heatmap data, we identified individuals having predicted 0–4 copies of these two regions corresponding to *SRGAP2D*. From this analysis, we observe that *SRGAP2D* includes a large internal deletion including exons 2 and 3 (~115 kbp), which affects all copies examined to date, and is copy number polymorphic in the general population.

The homologous region shared among our three *SRGAP2* contigs was analyzed for variants that could distinguish different *SRGAP2* paralogs. Such singly unique nucleotide (SUN) identifiers are defined as single base-pair variants that are fixed in the population for a particular paralog [21]. Identifying and characterizing these variants is a critical first step in developing our genotyping assays based on SUN k-mers (SUNKs) and paralog-specific quantitative PCR (qPCR; details of the primer design and protocol are below). SUNKs are defined as 30-mers that specifically tag a region of the genome and thus can be used in conjunction with short-read sequencing data to genotype highly identical paralogs [21]. Briefly, the *SRGAP2* SUNK map was generated by dividing each of the newly constructed *SRGAP2* contigs into its constituent overlapping 30-mers and mapping these kmers back to the human reference genome (NCBI36/hg18) and to each of the new contigs using the mrsFAST mapper [23]. Sequence reads represented in both the contigs and the human reference (allelic regions) were masked in the reference so as to eliminate mapping to duplicate regions. Contig-specific SUNKs were then defined as 30-mers that only mapped to one contig, specifically tagging a particular paralogous locus.

Read-depth-based copy number estimates were then generated considering only these SUNKs, ensuring copy number estimates would be paralog-specific. Across all 661 individuals examined, *SRGAP2A* and *SRGAP2C* were fixed at two copies with the exception of 11 individuals who exhibited possible *SRGAP2A* duplications. Further analysis and qPCR-based copy number genotyping of the unique portion of the *SRGAP2A*, however, indicated all of these individuals have two copies of this paralog. To be certain, we also tested HapMap individuals at the lower and upper tails of the *SRGAP2C* copy number distribution using qPCR and validated that all of these individuals have two copies of this paralog. Alternatively, *SRGAP2B* appeared to be copy number polymorphic among individuals. We observed 3 homozygous and 33 heterozygous *SRGAP2B* deletions, as well as 49 individuals with three copies and 1 with four copies. A subset of individuals with *SRGAP2B* deletions and duplications were validated using FISH (described above, using probe 1, see Table S1) and a custom Agilent array targeting sequence from the *SRGAP2* contigs. From this analysis, we show the deletions and duplications affect exons 2 and 3 of *SRGAP2B*. No copies of *SRGAP2B* and *SRGAP2C* were observed among any of the nonhuman primates analyzed (including 34 nonhuman primates consisting of 4 bonobos, 7 chimpanzees, 11 gorillas, and 12 Bornean and Sumatran orangutans [T. Marques-Bonet, personal communication]).

SRGAP2B Hardy-Weinberg Equilibrium Analysis

SRGAP2B showed copy number polymorphism in all human populations, leading us to believe it is likely a nonfunctional pseudogene. To more formally explore this possibility, we calculated whether copy number allele frequencies for this paralog are in Hardy-Weinberg proportions. If selection were operating on copy number at this locus, we would expect to find the *SRGAP2B* copy number allele frequencies inconsistent with Hardy-Weinberg equilibrium. Thus, if we find the *SRGAP2B* copy number allele frequencies at Hardy-Weinberg proportions, we can rule out selection on *SRGAP2B* copy number. The absence of selection on copy number would be consistent with the lack of an important functional role for this paralog. To determine whether the *SRGAP2B* copy number allele frequencies are at Hardy-Weinberg proportions in humans, we performed the following analysis (shown here for the Yoruban in Ibadan, Nigeria, but applied separately to all populations):

p: deletion allele frequency
q: normal allele frequency
r: duplication allele frequency

The observed counts of individuals having each *SRGAP2B* diploid copy number (CN) state are:

CN 0 (genotype PP): 0 individuals
CN 1 (genotype PQ): 6 individuals
CN 2 (genotype QQ + genotype PR): 49 individuals
CN 3 (genotype QR): 11 individuals
CN 4 (genotype RR): 0 individuals

We also assume all individuals with predicted copy number 2 have two normal alleles—this assumption will lead us to underestimate (though likely not by much) allele frequencies p and r:

estimate for p = (2*0 homozygous dels+6 heterozygous dels)/(66*2 chromosomes) = 0.045
estimate for q = (6 het dels+2*49 homozygous normal+11 heterozygous dups)/(66*2 chromosomes) = 0.871
estimate for r = (11 heterozygous dups+2*0 homozygous dup)/(66*2 chromosomes) = 0.083
$p^2+q^2+r^2+2pq+2pr+2qr = 1$, as required

Expected counts:

CN 0: $p^2*66 = 0.136$ individuals
CN 1: $2pq*66 = 5.227$ individuals
CN 2: $(q^2+2pr)*66 = 50.595$ individuals
CN 3: $2qr*66 = 9.583$ individuals
CN 4: $r^2*66 = 0.458$ individuals

Use these counts to calculate a chi-square statistic: $\text{sum}((\text{observed}-\text{expected})^2/\text{expected}) = 0.969$.

The resulting p-value for the chi-square value above with two degrees of freedom is 0.6161, meaning the Hardy-Weinberg equilibrium is not rejected when considering the Yoruban population. Performing the same calculations on other populations, we obtain strong evidence that *SRGAP2B* is at Hardy-Weinberg proportions in 13 of 14 populations considered. The low p-value in the remaining population (Columbian in Medellin, Colombia) likely reflects genotyping error for a single individual rather than a meaningful biological departure from Hardy-Weinberg equilibrium. If we assume this individual (copy number estimate = 3.65) has a *SRGAP2B* genotype of copy number 3 rather than copy number 4, the p-value for this last population becomes 0.98. Thus, these data are consistent with *SRGAP2B* paralog segregating at Hardy-Weinberg equilibrium in humans as expected for a nonfunctional pseudogene.

Paralog-Specific Genotyping Using qPCR

Using the 244.2 kbp alignment between our 1q32.1, 1q21.1, and 1p12 contigs, we designed paralog-specific primers for genotyping copy number. These primers had to pass several criteria: (1) they were deemed acceptable by the primer design software Primer3 (http://frodo.wi.mit.edu/primer3/); (2) they were found to theoretically amplify only one *SRGAP2* paralog; (3) all sequences in the NCBI HTGS database (http://www.ncbi.nlm.nih.gov/HTGS) containing the targeted region of their corresponding targeted paralog had the specificity-conferring variants (i.e., no evidence suggested these variants are not fixed in the population); (4) all sequences in the HTGS database containing the targeted region of non-targeted paralogs lacked the specificity-conferring variants; and (5) they could not yield more than one product as determined by the In-Silico PCR tool of the UCSC Genome Browser using the human reference (GRCh37/hg19). See Table S6 for primer sequences.

The qPCR experiments were performed using the Roche LightCycler 480 with a primer set targeting the albumin gene (*ALB*, known to be at diploid copy number 2) as a control. Each reaction

contained 5.0 µl Roche SYBR Green Master I, 0.2 µl of each primer (10 µM), 4 µl genomic DNA (2.5 ng/µL), and 0.6 µl PCR quality water. Cycling conditions included a hot start at 95°C for 5 min, followed by 40 cycles of melting at 95°C for 15 s, annealing primers at 58°C for 20 s, elongating products at 72°C for 20 s, and concluding with a melting curve from 50C to 90C. All qPCR reactions were run in three technical replicates. Cycles-at-threshold (C_t) values were calculated using the second derivative maximum method [24]. Raw cycles-to-threshold data were converted into copy number estimates using the delta-delta C_t method using an individual with known copy number 2 at *SRGAP2A*, *SRGAP2B*, *SRGAP2C*, and *SRGAP2D* as the control (NA12878).

SRGAP2A and _SRGAP2C_ Copy Number Variant Detection in Cases and Controls

We utilized previously existing data sets as well as targeted qPCR and array CGH to assess CNVs of *SRGAP2A* (1q32.1) and *SRGAP2C* (1p12), respectively. For *SRGAP2A*, we made use of previously identified CNVs reported in the Cooper et al. study [25] to identify deletions and duplications in a cohort of 15,767 children with developmental delay (including intellectual disability and autism spectrum disorder) screened using array CGH. Four of the six CNVs were validated using a custom Agilent microarray. The remaining two CNVs were very large (>20 Mbp) and likely real. From the same study, we assessed data from 8,329 controls screened using SNP arrays. Notably, due to inefficient probe coverage across the *SRGAP2C* segmental duplication and flanking region, we were not confident in our ascertainment of CNVs of the 1p12 region.

For both *SRGAP2A* and *SRGAP2C* loci, we utilized paralog-specific qPCR assays (see Table S6 for primer sequences) using experimental procedures described above. Specifically, 1,602 children with intellectual disability and 1,794 controls (comprised of NIMH [https://www.nimhgenetics.org/available_data/controls/] and ClinSeq [26] individuals) were screened with qPCR assays targeting intron 12 of *SRGAP2A* and intron 7 of *SRGAP2C*, respectively. As a control, we used a qPCR assay for *ALB* (fixed at copy number 2). No deletions or duplications were identified for *SRGAP2A* using this assay. To validate a small subset of individuals showing a deletion or duplication of *SRGAP2C*, we performed a second qPCR screen specific to *SRGAP2C* intron 6. From this, we validated only one case and one control with a duplication.

Specifically for *SRGAP2C*, we also used array CGH data from a custom Agilent microarray to assess whether the 1p11.2 region proximal to the segmental duplication (chr1-120,843,952-121,057,437, NCBI36/hg18; 36 probes) was deleted or duplicated in a cohort of children with sporadic (n = 2,294, Simons Simplex Collection [27]) and familial (n = 579, Autism Genetic Resource Exchange [28]) autism spectrum disorder. Additionally, we screened 580 controls on the same microarray. We detected no deletions and a small number of duplications in both cases and controls (~0.1%). Individuals with a detected duplication were assayed using the *SRGAP2C*-specific qPCR assay described above, with only two sporadic autism probands showing the duplication extending across *SRGAP2C*. All *SRGAP2C* duplications were validated using an Agilent custom array targeting our *SRGAP2* contig sequences (much of which is missing from the human reference genome) in an attempt to identify any breakpoints.

Database Search of _SRGAP2_ Transcripts

The sequences of all transcripts mapping to *SRGAP2* paralogs (*SRGAP2A* at 1q32.1, *SRGAP2B* at 1q21.1, and *SRGAP2C* at 1p12) in the GRCh37/hg19 reference sequence available through the UCSC Genome Browser were downloaded and analyzed for potential alternative splice variants or novel *SRGAP2* transcripts that could be expressed from the duplicate paralogs (i.e., transcripts containing only exons 1 through 9). We identified full-length *SRGAP2A* transcripts containing 22 coding exons (GenBank accessions AB007925, BC132872, BC144343, BC132874, BC150646) as well as several additional transcripts mapping to the *SRGAP2A* locus (GenBank accessions AK057565, AK294060, AK311111, BC063527, DQ786311, AK000885, AK091814, AK293335, AK295845, BC041635, DQ786257). Some *SRGAP2A* transcripts showed alternative splice forms excluding the first three nucleotides in exon 7, resulting in an in-frame removal of an amino acid. We also discovered a single transcript, cloned from a breast cancer cell line, truncated at exon 9 and including 1,373 additional nucleotides from intron 9 and a

polyA tail (GenBank accession BC112927). We determined that the sequence of the transcript matches our *SRGAP2C* 1p12 contig (and is not a *SRGAP2A* truncated transcript) and predicts an open-reading frame (ORF) that, if translated, encodes a truncated *SRGAP2* protein (458 amino acids), including a partial F-BAR domain [29] with seven unique residues at the carboxy terminus.

Sequencing of *SRGAP2* Transcripts

We performed long-range PCR (Expand Long Template, Roche) on cDNA (generated using the Roche High Fidelity cDNA Synthesis Kit, oligo(dT) primers) from SH-SY5Y neuronal cell line total RNA, human fetal brain total RNA (collected from spontaneously aborted fetuses, 20-33 weeks, ClonTech S2437; approximately 50-60 fetuses pooled), single human adult brain mRNA (BioChain, M1234035), and single human fetal brain total RNA (BioChain, R1244035-50). We amplified a single PCR band at the expected size for SRGAP2A-specific primers (including exon 10). Alternatively, multiple PCR bands (2-3) are amplified using the duplicate-specific SRGAP2 primers (including the intron 9 extension). We PCR purified and cloned these fragments into the pCR4.0 vector (Invitrogen) and digested with EcoRI (NEB) to validate PCR insert sizes in vectors. We sequenced clones using primers spanning across the *SRGAP2* region (ABI3630 Genetic Analyzer). Primer sequences are shown in Table S6. Additional primers were used to sequence clones containing the *SRGAP2* transcripts (available upon request).

Transcripts were assigned to their respective paralogs based on diagnostic sequence variants from genome sequencing. We did not detect any splice or sequence variants within exons 1 through 10 of the *SRGAP2A* ancestral paralog (n = 11 transcripts). Alternatively, we detected three splice and numerous sequence variants of the *SRGAP2* duplicate-derived transcripts (n = 85 transcripts) based on diagnostic sequence differences. We identified ‘‘full-length’’ duplicate transcripts containing exons 1 through 9 mapping to both *SRGAP2B* (n = 4) and *SRGAP2C* (n = 47). One rare splice isoform (n = 2), which removed exon 3 and mapped to *SRGAP2C*, encodes a truncated 98-residue protein, including 11 unique amino acids at the carboxy terminus. The other splice isoform (n = 31), which removed exons 2 and 3, mapped to *SRGAP2D*. Although we do not have finished sequence for *SRGAP2D*, the sequence and structure of this transcript is consistent with *SRGAP2D* transcription, as this paralog harbors a large genomic deletion containing exons 2 and 3. From this analysis, we were able to assign 16 polymorphic and fixed PSVs to the *SRGAP2* paralogs.

In order to determine the relative abundance of ‘‘full-length’’ transcripts (i.e., including both exons 2 and 3) expressed from *SRGAP2B* and *SRGAP2C*, we performed long-range PCR amplification using primers specific to exon 3 and the intron 9 (duplicate-specific) extension, respectively, from cDNA (prepared using oligo(dT) primers) generated from human adult brain, fetal brain, and lymphoblastoid cell lines. These same cell lines had been genotyped for the *SRGAP2B* polymorphism. We performed capillary sequencing using the same primers used to amplify the cDNA and compared the chromatograms of two coding PSVs (variants 12 and 13). We show that relative expression of the *SRGAP2B* transcript is lower than *SRGAP2C* expression by quantifying peak heights of chromatogram plots for each nucleotide base corresponding to a PSV. As a control, we detected no *SRGAP2B* transcript from a HapMap lymphoblastoid cell line genotyped as copy number 0 for *SRGAP2B*.

Nonsense-Mediated Decay of the *SRGAP2D* Isoform

The *SRGAP2D* paralog produces a transcript missing exons 2 and 3 resulting in a premature stop codon at the exon 1 and 4 junction. We predict that this transcript will undergo nonsense-mediated decay (NMD); to test this, we assessed the abundance of this transcript in HapMap lymphoblastoid cell lines in normal and NMD-blocked conditions (i.e., in the presence of emetine, which blocks translation and NMD [30]). Specifically, we incubated HapMap EBV-transformed lymphoblastoid cells (1x107) in 10 ml of media (see ATCC guidelines) with or without 100 mg/ml emetine dihydrochloride hydrate (Sigma, E2375) for 7 hr at 37C. We immediately isolated total RNA from each treated and untreated cell line using Trizol (Invitrogen) and the RNeasy Mini Kit (QIAGEN). cDNA was prepared from 3 mg of total RNA using the Transcription High Fidelity cDNA Synthesis Kit (Roche) and random hexamer primers.

qPCR was performed following the same protocol described earlier with 10 ng of cDNA and primers specific to primers mapping to: (1) the *SRGAP2D* exon 1 and 4 junction and within exon 5; (2) *SRGAP2A*-specific exons 21 and 22; and (3) exons of the housekeeping gene *GAPDH*. See Table S6 for primer sequences.

First, to determine whether the cDNA isoform excluding exons 2 and 3 (‘‘deletion isoform’’) is derived from the *SRGAP2D* or *SRGAP2B* locus (e.g., from a copy having the internal deletion polymorphism), we assessed overall levels of the deletion isoform in cell lines with *SRGAP2B* copy numbers of 0, 1, and 2, respectively. We found no difference in levels of transcript across cell lines indicating that the deletion isoform is likely derived from *SRGAP2D*. Second, by blocking NMD in these cells we discovered a significant 1.1- to 1.6-fold increase of the *SRGAP2D* aberrant transcript compared to the full-length *SRGAP2A* transcript abundance (0.9-to 1.2-fold change). Overall, this indicates that NMD may be acting on the *SRGAP2D* transcript, though to a modest degree. Notably, performing this assay using lymphoblastoid cells rather than neuronal cells may limit the biological relevance of this result.

SRGAP2 Tissue-Specific Expression Analysis Using Paralog-Specific qPCR

From the transcript analysis of the *SRGAP2* paralogs, we designed paralog-specific RT-PCR primers to assess the expression of specific *SRGAP2* paralogs. In designing *SRGAP2C*-specific primers, we leveraged PSV-12 within exon 7. Similar attempts to design *SRGAP2B*-specific primers for PSVs-5, –13 and –16 resulted in non-specific amplification of both alleles. We found that the ancestral *SRGAP2A*, in addition to the *SRGAP2B* and *SRGAP2C* duplicates, were expressed in a variety of human brain tissues [using pooled total RNA (ClonTech)] including total fetal brain (ID: 636526), adult brain (ID: 636530), cerebellum (ID: 636535), frontal lobe (ID: 636563), and temporal lobe (ID: 636564). As expected, we did not detect expression of the duplicate copies in any of the nonhuman primate-derived tissues.

Spatiotemporal _SRGAP2_ Expression Using RNA-Seq Data

A subset of the SUNKs we identified (described above) were embedded in coding and UTR sequence of the *SRGAP2* paralogous transcripts, providing a unique opportunity to assess paralog-specific expression patterns using RNA-Seq data. Sequence from four different studies was analyzed encompassing 17 different human tissues (Illumina’s Human BodyMap 2.0), 7 human cell lines [31], and both chimpanzee and macaque cerebellum and liver tissues [32]. Briefly, RNA-Seq data sets were mapped to the human reference genome (NCBI36/hg18) in addition to the newly sequenced *SRGAP2* contigs and expression levels over genes were calculated in units of RPKM [33]. Among human tissues, we found that the *SRGAP2* paralogs were most highly expressed in whole-brain, cerebellum, and breast tissues, with the whole-brain and cerebellum samples showing the tightest expression levels with least variability between biological replicates. *SRGAP2A* (the ancestral paralog) shows similar expression levels in the cerebellum of chimpanzees and macaques compared to humans with little to no expression in the liver of humans, chimpanzees, or macaques. As expected, no signature of *SRGAP2* duplicate transcripts was detected in chimpanzees or macaques. Within the cerebellum, we observed *SRGAP2A* and *SRGAP2B/D* transcripts to be the most abundant and *SRGAP2C* transcripts to be the least abundant, though this analysis does not account for potential alternative splice isoforms.

1000 Genomes Populations

YRI Yoruba from Ibadan, Nigeria
ASW African American from the Southwest United States
CEU Centre d'Etude du Polymorphisme Humain collection (European)
MXL Mexican ancestry in Los Angeles, California
IBS Iberian Population in Spain
CLM Colombians from Medellin, Colombia
JPT Japanese from Tokyo, Japan

PUR Puerto Ricans from Puerto Rico
GBR British from England and Scotland
CHB Han Chinese from Beijing
CHS Han Chinese South
FIN Finnish from Finland
LWK Luhya in Webuye, Kenya
TSI Toscani in Italia

Supplemental References

1. Kidd, J.M., et al., *A human genome structural variation sequencing resource reveals insights into mutational mechanisms.* Cell, 2010. **143**(5): p. 837-47.
2. Antonacci, F., et al., *A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk.* Nat Genet, 2010. **42**(9): p. 745-50.
3. Sharp, A.J., et al., *Segmental duplications and copy-number variation in the human genome.* Am J Hum Genet, 2005. **77**(1): p. 78-88.
4. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.
5. Schmutz, J., et al., *Quality assessment of the human genome sequence.* Nature, 2004. **429**(6990): p. 365-8.
6. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
7. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4.10.
8. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2.3.
9. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
10. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
11. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
12. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J Mol Evol, 1980. **16**(2): p. 111-20.
13. Dopazo, J., *Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach.* J Mol Evol, 1994. **38**(3): p. 300-4.
14. Rzhetsky, A. and M. Nei, *METREE: a program package for inferring and testing minimum-evolution trees.* Comput Appl Biosci, 1994. **10**(4): p. 409-12.
15. Brunet, M., et al., *A new hominid from the Upper Miocene of Chad, Central Africa.* Nature, 2002. **418**(6894): p. 145-51.
16. Brunet, M., et al., *New material of the earliest hominid from the Upper Miocene of Chad.* Nature, 2005. **434**(7034): p. 752-5.
17. Vignaud, P., et al., *Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad.* Nature, 2002. **418**(6894): p. 152-5.
18. Patterson, N., et al., *Genetic evidence for complex speciation of humans and chimpanzees.* Nature, 2006. **441**(7097): p. 1103-8.
19. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
20. Bailey, J.A., et al., *Segmental duplications: organization and impact within the current human genome project assembly.* Genome Res, 2001. **11**(6): p. 1005-17.

21. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes.* Science, 2010. **330**(6004): p. 641-6.
22. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-7.
23. Hach, F., et al., *mrsFAST: a cache-oblivious algorithm for short-read mapping.* Nat Methods, 2010. **7**(8): p. 576-7.
24. Zhao, S. and R.D. Fernald, *Comprehensive algorithm for quantitative real-time polymerase chain reaction.* J Comput Biol, 2005. **12**(8): p. 1047-64.
25. Cooper, G.M., et al., *A copy number variation morbidity map of developmental delay.* Nat Genet, 2011. **43**(9): p. 838-46.
26. Biesecker, L.G., et al., *The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine.* Genome Res, 2009. **19**(9): p. 1665-74.
27. Fischbach, G.D. and C. Lord, *The Simons Simplex Collection: a resource for identification of autism genetic risk factors.* Neuron, 2010. **68**(2): p. 192-5.
28. Geschwind, D.H., et al., *The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions.* Am J Hum Genet, 2001. **69**(2): p. 463-6.
29. Guerrier, S., et al., *The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis.* Cell, 2009. **138**(5): p. 990-1004.
30. Noensie, E.N. and H.C. Dietz, *A strategy for disease gene identification through nonsense-mediated mRNA decay inhibition.* Nat Biotechnol, 2001. **19**(5): p. 434-9.
31. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
32. Blekhman, R., et al., *Sex-specific and lineage-specific alternative splicing in primates.* Genome Res, 2010. **20**(2): p. 180-9.
33. Liu, S., et al., *A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species.* Nucleic Acids Res, 2011. **39**(2): p. 578-88.
34. Saitsu, H., et al., *Early infantile epileptic encephalopathy associated with the disrupted gene encoding Slit-Robo Rho GTPase activating protein 2 (SRGAP2).* Am J Med Genet A, 2012. **158a**(1): p. 199-205.
35. Charrier, C., et al., *Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation.* Cell, 2012. **149**(4): p. 923-35.

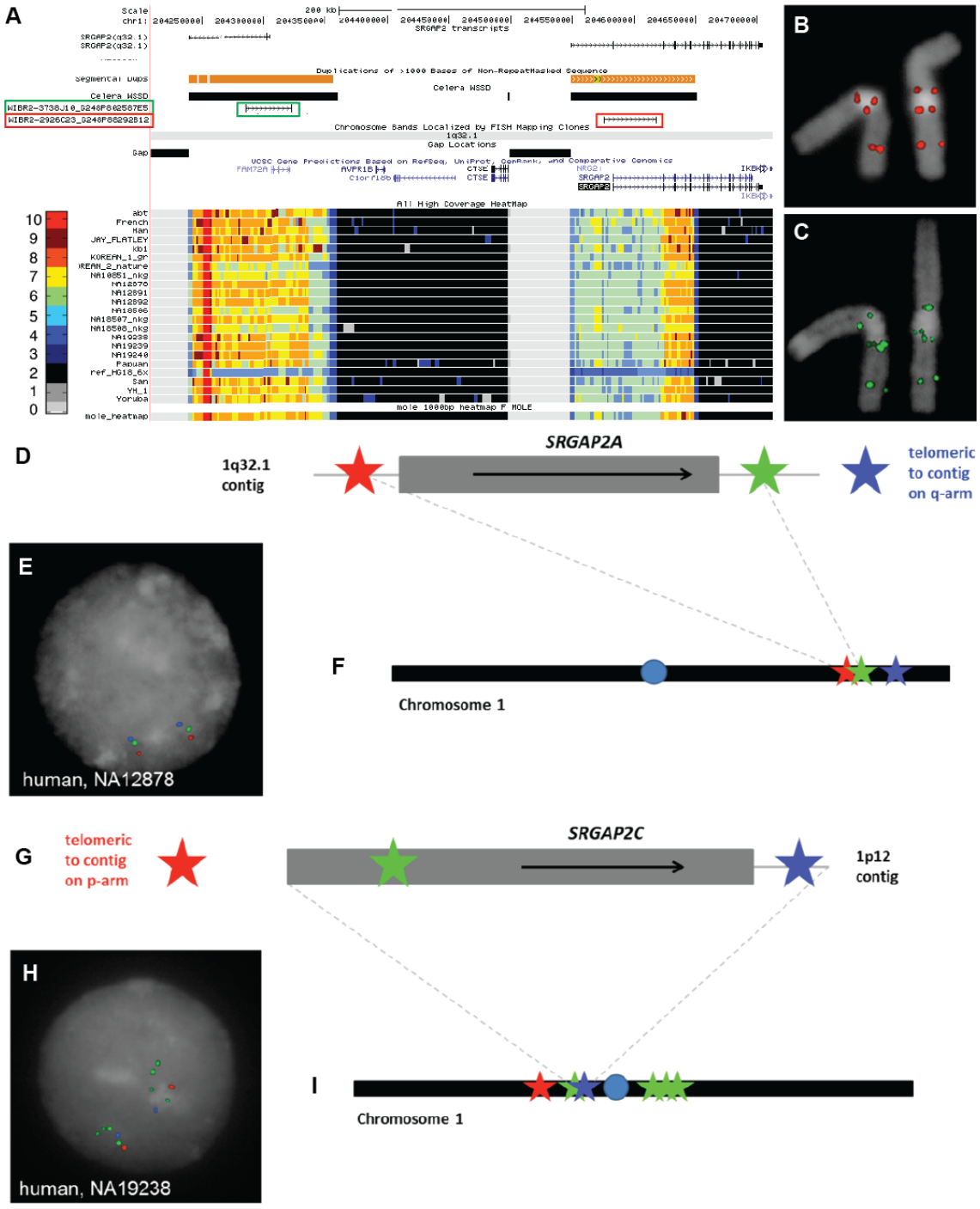


Figure S1. FISH Experiments Reveal a Fourth *SRGAP2* Paralog and Specify *SRGAP2* Paralog Orientations, Related to Figure 1. (A–C) The UCSC genome browser snapshot of the *SRGAP2A* locus in NCBI36/hg18 shows the *SRGAP2A* transcript, segmental duplications, fosmid clones used for FISH experiments, gaps in the reference genome, gene predictions, and copy number heatmaps for 22 human genomes sequenced at high coverage (including the complete hydatidiform mole genome used in this study) as well as for an ‘‘Illuminized’’ NCBI36/hg18 reference sequence. Colors indicate copy number predictions based on short-read sequence read-depth [21]. *SRGAP2A* is incomplete and partially inverted in the reference genome NCBI36/hg18. FISH analysis using a probe spanning

from intron 2 to intron 3 (B) detects three *SRGAP2* paralogs on metaphase human chromosome 1. However, FISH using a probe spanning from upstream of *SRGAP2* to intron 1 (C) detects four *SRGAP2* paralogs on metaphase human chromosome 1. These results are consistent with the heatmap data in (A) and suggest the existence of a fourth human *SRGAP2* paralog having an internal deletion of at least exon 3. (D–I) Depictions of probe locations, FISH images of interphase human chromosome 1, and schematics showing the results of the experiments to resolve the orientation of the *SRGAP2A* (D–F) and *SRGAP2C* (G–I) paralogs. Colored stars indicate relative locations of the corresponding FISH probes with regard to our *SRGAP2* contigs (gray boxes). Thin gray lines indicate extensions of our contigs based on contiguous reference sequence. Extensive local duplication upstream of the *SRGAP2B* and *SRGAP2D* paralogs (see yellow probe in Figure 2C) prevented accurate determination of their orientation using FISH. Using an anchored contig spanning the entire human 1q21.1 region recently generated at The Genome Institute at Washington University School of Medicine (unpublished data), we determined that *SRGAP2B* is oriented such that gene transcription would proceed toward the centromere.

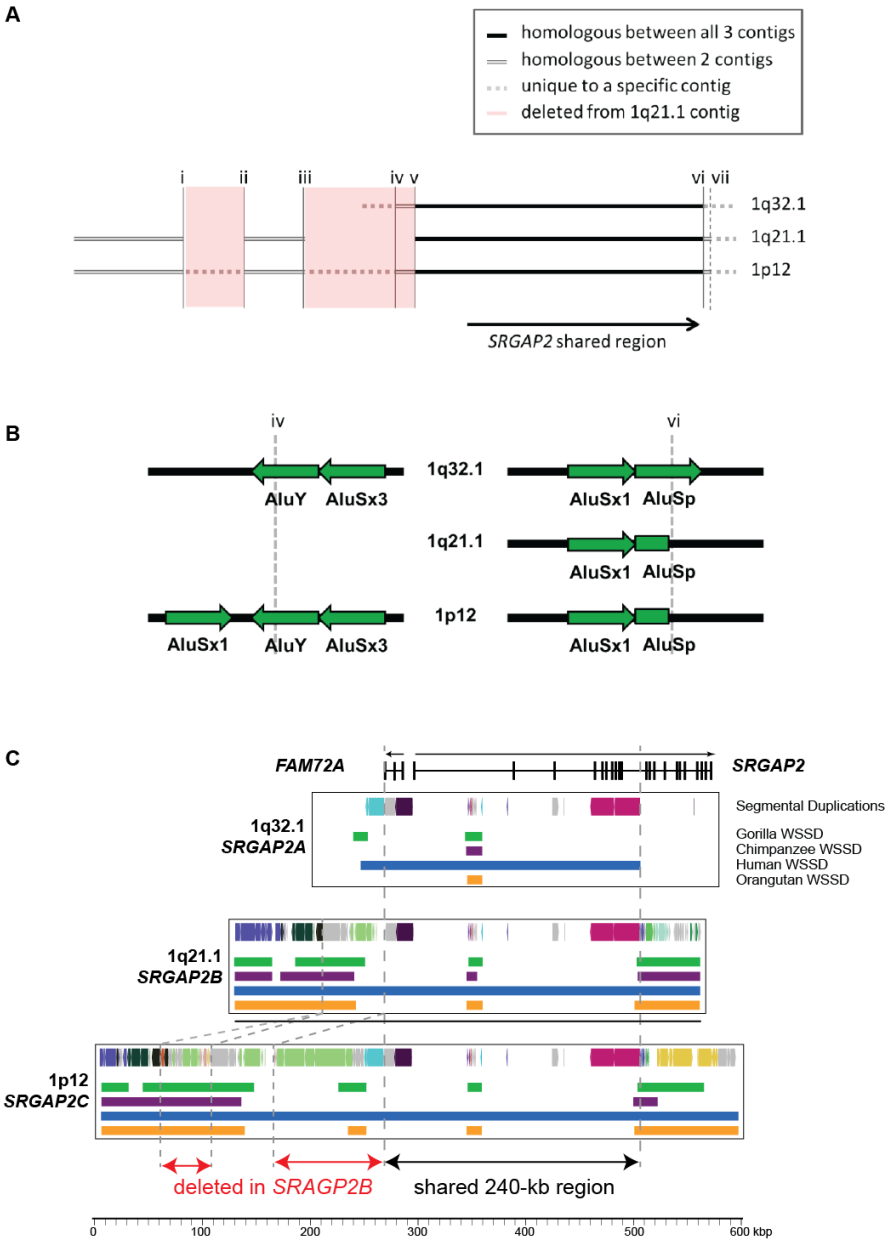


Figure S2. Breakpoint and Duplication Analyses of *SRGAP2* Contigs, Related to Figure 2. (A) Homologous regions shared between the 50 ends of the *SRGAP2* contigs are indicated, as well as their relative position with respect to the first nine exons of *SRGAP2* shared between these three paralogs. The region shared by these three paralogs spans 244.2 kbp, with the total shared sequence length depicted here over 515 kbp. Vertical lines indicate breakpoints corresponding to duplication breakpoints (iv, vi, vii) or breakpoints from other structural rearrangements (i, ii, iii, v). All breakpoints were extremely well defined except for vii; this lower resolution (within a few hundred base pairs) is indicated by a dashed vertical line. The 50 duplication breakpoint between the 1q21.1 and 1p12 paralogous regions lies beyond the edge of the contigs. Missing sequence in the 1q21.1 contig between iii and v resulted from a deletion in 1q21.1 rather than an insertion in 1p12 because part of this missing sequence (between iv and v) extends the region of homology with paralogous sequence at 1q32.1. (B) Zoomed-in views of the initial duplication breakpoints are presented with repetitive elements highlighted. These elements were identified by using RepeatMasker [7] on sequences surrounding the breakpoints—the best repeat subfamily matches are indicated. Breakpoints i, ii, and v also contained Alu elements at their boundaries. The remaining breakpoints lack repetitive features in their immediate surrounding sequences. (C) A duplication analysis of the *SRGAP2* contigs using SegDupMasker and whole-genome shotgun sequence detection (WSSD) [22] highlights the highly duplicated chromosomal environments flanking *SRGAP2* paralogs at 1q21.1 and 1p12 and affirms that the *SRGAP2* duplication is specific to the human lineage.

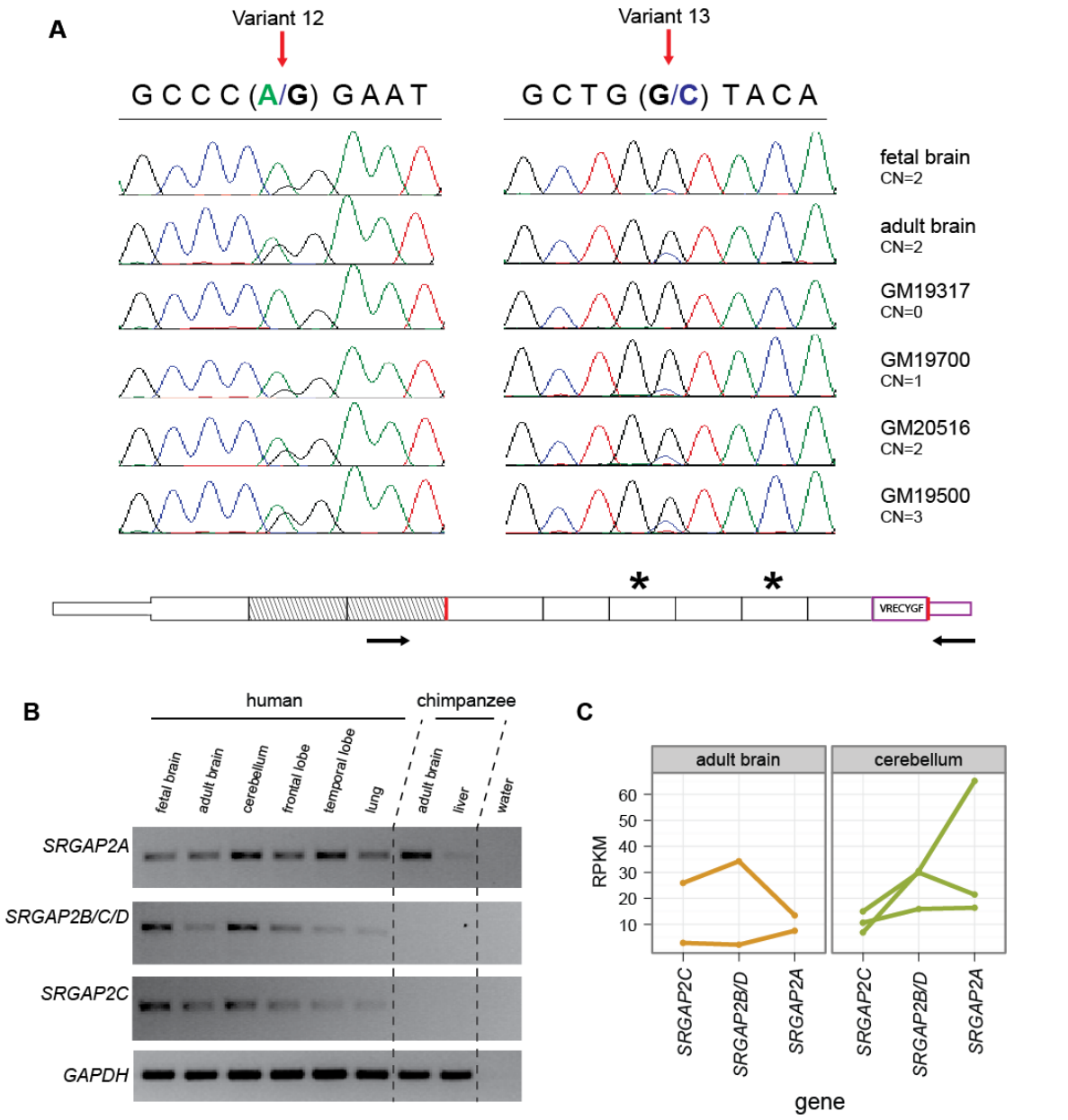


Figure S3. SRGAP2 Gene Expression Analysis, Related to Figure 3. (A) Sequencing of *SRGAP2* ''full-length'' transcript from *SRGAP2B* and *SRGAP2C* revealed reduced expression of *SRGAP2B*. To perform this experiment, we used primers targeting transcripts containing exons 3 and the 30 UTR (intron 9 extension), respectively, to avoid quantifying expression of the *SRGAP2D* transcript (with deleted exons 2 and 3). Pictured are chromatograms for coding paralog-specific variants (PSVs) 12 (*SRGAP2C*-allele = A and *SRGAP2B*-allele = G) and 13 (*SRGAP2C*-allele = G and *SRGAP2B*-allele = C) (see Figure 3B in the main text) from transcripts derived from human adult brain, fetal brain, and lymphoblastoid cells. Relative transcript abundances were determined by comparing the heights of the PSV peaks of the chromatograms. (B) RT- PCR was performed using primers specific to the ancestral and duplicate *SRGAP2* paralogs and a housekeeping gene, *GAPDH*, using cDNA derived from human and chimpanzee tissues. The following paralogs were amplified based on the existence of specific exons or by utilizing PSVs including: *SRGAP2A* (exon 8 and exon 10); SRGAP2B/C/D (exon 8 and 30 UTR extending into intron 9); and *SRGAP2C* (exon 6 containing PSV-12 and exon 7). For primer sequences used, refer the Extended Experimental Procedures. (C) Individual RPKM estimates (Liu et al., 2011) allow quantification of expression *SRGAP2* paralogs from human adult brain and cerebellum. Tissue from two and three individuals was used to test expression in adult brain and cerebellum, respectively.

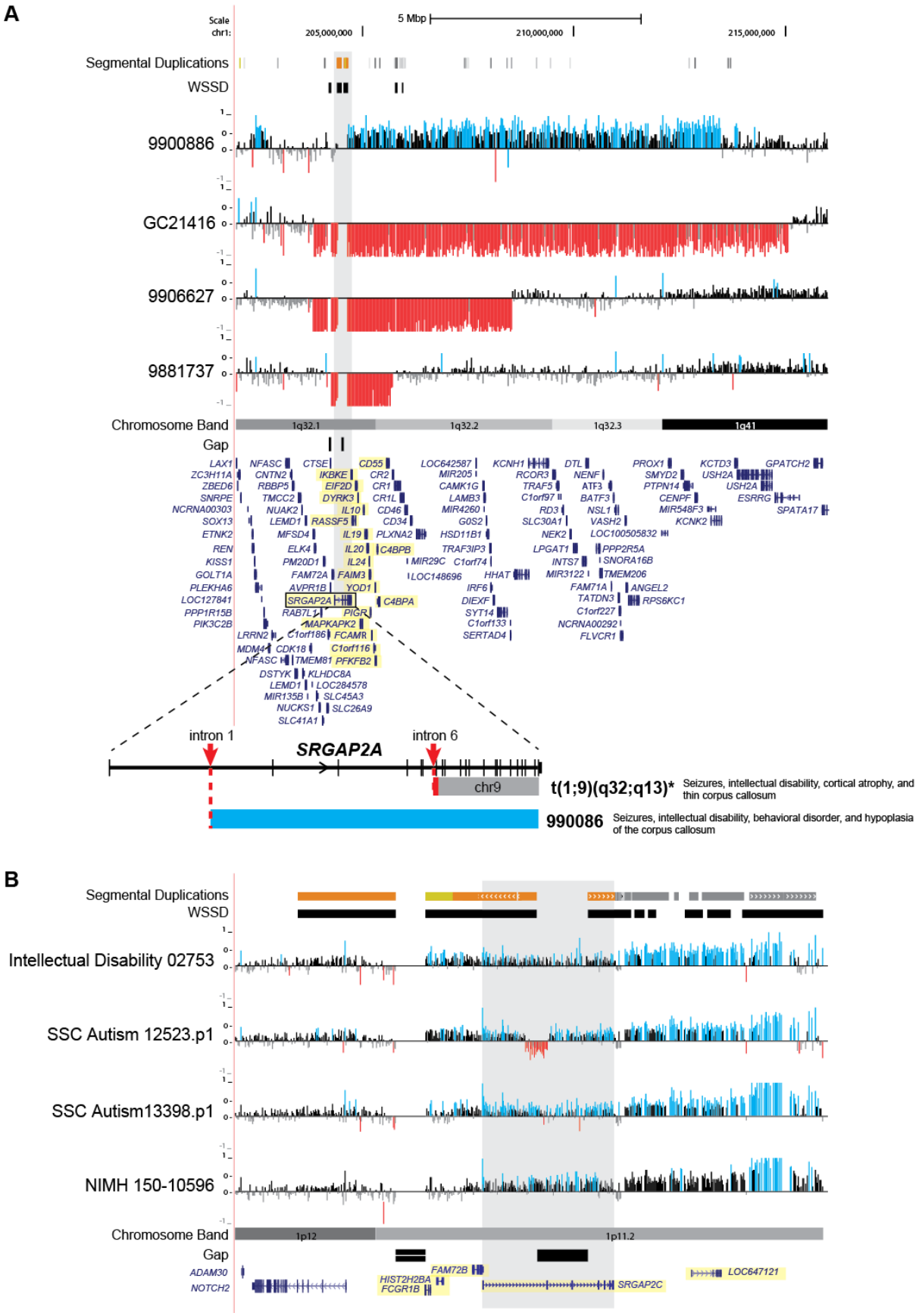

A
Scale
chr1:
5 Mbp
205,000,000
210,000,000
215,000,000
Segmental Duplications
WSSD
9900886
GC21416
9906627
9881737
Chromosome Band
1q32.1
1q32.2
1q32.3
1q41
Gap
intron 1
SRGAP2A
intron 6
chr9
t(1;9)(q32;q13)*
Seizures, intellectual disability, cortical atrophy, and thin corpus callosum
990086
Seizures, intellectual disability, behavioral disorder, and hypoplasia of the corpus callosum
B
Segmental Duplications
WSSD
Intellectual Disability 02753
SSC Autism 12523.p1
SSC Autism13398.p1
NIMH 150-10596
Chromosome Band
1p12
1p11.2
Gap
ADAM30
NOTCH2
FAM72B
HIST2H2BA
FCGR1B
SRGAP2C
LOC647121

Figure S4. Large CNVs of *SRGAP2A* and *SRGAP2C* Detected in Children with Developmental Delay and Autism, Related to Table 2. Large (>1 Mbp) deletions (red) and duplications (blue) of *SRGAP2A* (A) and *SRGAP2C* (B) were confirmed by array CGH for seven children with developmental delay and autism spectrum disorder, as well as an adult control individual. Two duplications of *SRGAP2A* (>20 Mbp) are not shown. Blue (duplication) and red (deletion) histograms depict $\log_2$ relative hybridization signals. The genes within the smallest region of overlap are highlighted in yellow. (A) Below the array CGH data for *SRGAP2A* is an expanded view of proximal breakpoints (red arrows) mapping within *SRGAP2A* for two patients. The first is a de novo t(1;9)(q32;q13) translocation breakpoint (*described by Saitsu et al. [34]) that maps within intron 6 and resulted in the deletion of exon 7 (red) and the remaining chromosome 1q-arm translocated to chromosome 9 (gray). The second is an 8.7 Mbp duplication (blue) breakpoint identified in this study that maps to *SRGAP2A* intron 1, assayed using a custom microarray targeting our *SRGAP2* contig sequences. Both patients show remarkable similarity in phenotype, including abnormalities of the corpus callosum, seizure, and intellectual disability. (B) The depicted genomic region is a hybrid of the human reference (GRCh37/hg19) and missing sequence data generated from our *SRGAP2C* contigs. Note the genome assembly gap within the human reference extends across exons 2 and 3 of *SRGAP2C*. The deletion spanning exon 2 in the 12523.p1 autistic proband likely represents polymorphism of *SRGAP2B*.

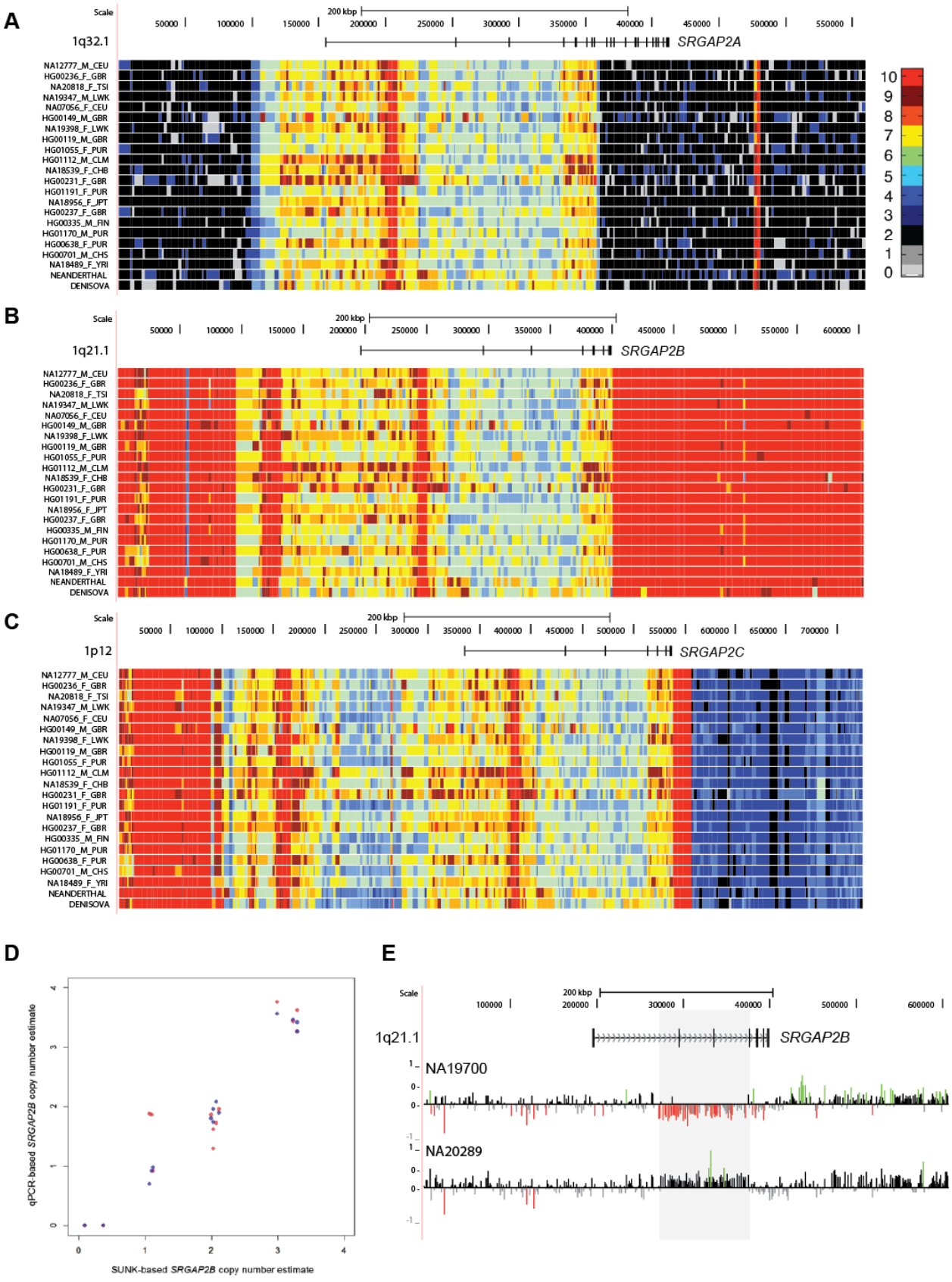

A
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000
1q32.1
SRGAP2A
NA12777_M_CEU
HG00236_F_GBR
NA20818_F_TSI
NA19347_M_LWK
NA07056_F_CEU
HG00149_M_GBR
NA19398_F_LWK
HG00119_M_GBR
HG01055_F_PUR
HG01112_M_CLM
NA18539_F_CHB
HG00231_F_GBR
HG01191_F_PUR
NA18956_F_JPT
HG00237_F_GBR
HG00335_M_FIN
HG01170_M_PUR
HG00638_F_PUR
HG00701_M_CHS
NA18489_F_YRI
NEANDERTHAL
DENISOVA
10 9 8 7 6 5 4 3 2 1 0
B
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000 600000
1q21.1
SRGAP2B
C
Scale
200 kbp
50000 100000 150000 200000 250000 300000 350000 400000 450000 500000 550000 600000 650000 700000
1p12
SRGAP2C
D
qPCR-based SRGAP2B copy number estimate
SUNK-based SRGAP2B copy number estimate
0 1 2 3 4
E
Scale
200 kbp
100000 200000 300000 400000 500000 600000
1q21.1
SRGAP2B
NA19700
NA20289

Figure S5. Copy Number Analysis of Next-Generation Sequencing Data, Related to Figure 4. Shown are heatmaps of aggregate copy number across the *SRGAP2* contigs using short-read sequences from human, Neanderthal, and Denisova genomes. Depicted are a representative sample of diverse human individuals from a total panel of 661 individuals from 14 populations (1000 Genomes Project), Neanderthal, and Denisova for the (A) 1q32.1 (*SRGAP2A*), (B) 1q21.1 (*SRGAP2B*), and (C) 1p12 (*SRGAP2C*) contigs. Gene models are shown at the top. Colors represent varying aggregate copy number predictions based on read-depth analysis of short-read sequencing data. In contrast to nonhuman primates, the first nine exons of *SRGAP2* in all the human samples analyzed are duplicated. Specifically, the genomic regions containing exons 1 as well as exons 4 through 9 are predicted as diploid copy number 7-8 while the region containing exon 2 and 3 are predicted as diploid copy number 5-6. From this analysis, we validated the existence of a fourth paralog lacking exons 2 and 3 (*SRGAP2D*). Additionally, the genomic regions flanking both sides of *SRGAP2B* at 1q21.1 show high copy number (>10 diploid copies) adding to the evidence that this region likely represented a non-ideal gene environment at the time of the initial duplication. (D) Comparison of SUNK-based and qPCR-based copy number estimates for *SRGAP2B* in multiple human individuals shows a clear correlation. Each point corresponds to an ordered pair of *SRGAP2B* copy number estimates, with the abscissa being the SUNK-based estimate and the ordinate being the qPCR estimate. Red points are from one qPCR experiment, and blue points are from a replicate qPCR experiment. The qPCR results recapitulate the four clusters seen in our SUNK analysis, clusters corresponding to different copy number states for *SRGAP2B* paralog. The overall fit of a linear model to these points has an $R^2 = 0.9087$, indicating strong concordance between the two orthogonal copy number estimation methods. These data confirm that *SRGAP2B* paralog is indeed polymorphic in humans. Note, qPCR of *SRGAP2D* also confirmed that this paralog is polymorphic in humans (not shown). (E) Array CGH of two HapMap individuals (NA19700 and NA20289) with a predicted deletion and duplication, respectively, of *SRGAP2B* was performed to validate polymorphism of this paralog. Blue (duplication) and red (deletion) histograms depict $\log_2$ relative hybridization signals mapped to our 1q21.1-sequenced contig (*SRGAP2B*). Both the deletion and duplication span the genomic region containing exons 2 and 3 of the paralog. The corresponding FISH experiment for NA19700 is depicted in Figure 4C.

Table S1. Fosmid Clones Used for FISH Experiments, Related to Figure 1

Probe	Fosmid clone	Genomic coordinates (NCBI36/hg18)	Target Sequence Description
1	WIBR2-2926C23_G248P88292B12	chr1: 204,575,058-204,618,304	duplicated *SRGAP2* sequence (predicted copy number 6 in human) from intron 2 to intron 3
2	WIBR2-2044O01_G248P86756H1	chr1: 205,518,260-205,556,982	region telomeric to 1q32.1 *SRGAP2A* contig
3	WIBR2-3685H16_G248P801507D8	chr1: 204,455,995-204,491,660	region slightly beyond 5' end of 1q32.1 *SRGAP2A* contig (5' of 1q32.1 paralog)
4	WIBR2-2212C22_G248P86986B11	chr1: 204,976,699-205,015,867	region slightly beyond 3' end of 1q32.1 *SRGAP2A* contig (3' of 1q32.1 paralog)
5	WIBR2-3549F23_G248P802137C12	chr1: 119,969,004-120,006,662	region telomeric to 1p12 *SRGAP2C* contig
6	WIBR2-1864B19_G248P86489A10	chr1: 147,506,053-147,549,258	region within 1p12 *SRGAP2C* contig, near the 5' end (5' of 1p12 paralog)
7	WIBR2-1489L21_G248P83865F11	chr1: 120,931,054-120,966,849	region slightly beyond 3' end of 1p12 *SRGAP2C* contig (3' of 1p12 paralog)
8	WIBR2-2397J12_G248P82711E6	chr1: 120,697,113-120,735,077	region just outside of original 258 kbp duplicated sequence (targets sequence where the original duplication landed)
9	WIBR2-3738J10_G248P802587E5	chr1: 204,285,032-204,323,561	duplicated *SRGAP2* sequence (predicted copy number 8 in human) from upstream of the gene to intron 1

Table S2. Maximum-Likelihood Estimates of Selection of *SRGAP2* Orthologs, Related to Figure 2

Model comparison	Model 1	Model 2	p-value*	dN/dS
Purifying selection	dN/dS = 1	one dN/dS	6.95E-126	All = 0.01221
Lineage heterogeneity	one dN/dS	free dN/dS	1.16E-05	Human *SRGAP2A* = 0.0001 Human *SRGAP2B* = 0.6302 Human *SRGAP2C* = ∞
Primate specific	one dN/dS	primate dN/dS	1.46E-03	Primates = 0.0415 Rest = 0.00542
Human specific	one dN/dS	human dN/dS	3.05E-10	Human = 0.7358 Rest = 0.0044
Human *SRGAP2B/C* specific	**one dN/dS**	**duplicate dN/dS**	**1.32E-11**	**Human *SRGAP2B/C* = 2.2310** **Rest = 0.0044**
Human *SRGAP2B* specific	one dN/dS	1q21.1 dN/dS	1.79E-03	Human *SRGAP2B* = 0.62909 Rest = 0.01000
Human *SRGAP2C* specific	one dN/dS	1p12 dN/dS	2.53E-09	1p12 = 999 Rest = 0.00662
Site-specific positive selection	model 1a	model 2	0.999	n/a
	model 7	model 8	1	n/a
	model 8a	model 8	1	n/a

* A likelihood ratio test (chi-squared test of the log-likelihood ratio of two models) was used to compare model 1 (null model) and model 2 (alternative model). Significant p-values reflect a higher fit of the data to the alternative model over the null model.

Table S3. Sequence Analysis of Human *SRGAP2* mRNA Transcripts from Neuronal Cells, Related to Figure 3

Variant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Exon	**1**			**2**		**3**	**5**			**6**			**8**	**9**	**Intron 9 extension**	
ANCESTRAL *SRGAP2A*																
***SRGAP2A* contig (CH17)**	**C**	**G**	**G**	**G**	**C**	**C**	**G**	**C**	**C**	**G**	**T**	**G**	**G**	**T**	**a**	**g**
SH-SY5Y Cell Line [3]	C	G	G	G	C	C	G	C	C	G	T	G	G	T	n/a	n/a
Pooled Fetal Brain [8]	C	G	G	G	C	C	G	C	C	G	T	G	G	T	n/a	n/a
DUPLICATE *SRGAP2C*																
***SRGAP2C* contig (CH17)**	**T**	**G**	**G**	**A**	**C**	**T**	**C**	**T**	**C**	**A**	**C**	**A**	**G**	**T**	**A**	**A**
I.M.A.G.E. cDNA clone (Accession:BC112927)*	T	A	G	A	C	T	C	T	C	A	C	A	G	T	A	A
Pooled Fetal Brain [1]	T	G	G	A	C	T	C	T	C	A	C	A	G	T	A	A
SH-SY5Y Cell Line [2]	T	A	G	A	C	T	C	C	T	G	C	A	G	T	A	A
Single Adult Brain [14]			G	A	C	T	C	C	T	G	C	A	G	T	A	A
Single Fetal Brain [17]	T	A	G	A	C	T	C	C	T	G	C	A	G	T	A	A
Single Fetal Brain [3]	T	A	G	A	C	T	C	A	T	G	C	A	G			
Single Fetal Brain [1]^				G	C	T	C	C	T	G	C	A	G			A
Pooled Fetal Brain [8]	T	A	G	A	C	T	C	C	T	G	C	A	G	T	A	A
Pooled Fetal Brain [1]^	T	A	G	A	C	T	C	C	T	G	C	A	G	C	A	A
Pooled Fetal Brain [2]	T	A	G	A	C	n/a	C	C	T	G	C	A	G	T	A	A
DUPLICATE *SRGAP2B/D*																
***SRGAP2B* contig (CH17)**	**T**	**G**	**G**	**G**	**T**	**C**	**C**	**C**	**C**	**G**	**C**	**G**	**C**	**C**	**G**	**G**
***SRGAP2D* contig (CH17)**													**C**	**C**	**A**	**G**
SH-SY5Y Cell Line [1]	T	G	G	G	T	C	C	C	C	G	C	G	C	C	G	G
Single Fetal Brain [1]				G	T	C	C	C	C	G	C	G	C	C	G	G
Pooled Fetal Brain [2]*	T	G	G	G	T	C	C	C	C	G	C	G	C	C	G	G
SH-SY5Y Cell Line [2]	T	G	A	n/a	n/a	n/a	C	C	C	A	C	G	C	C	A	G
Single Adult Brain [3]	T	G	A	n/a	n/a	n/a	C	C	C	A	C	G	C	C	G	G
Single Fetal Brain [6]	T	G	G	n/a	n/a	n/a	C	C	C	A	C	G	C	C	A	G
Pooled Fetal Brain [8]	T	G	A	n/a	n/a	n/a	C	C	C	G	C	G	C	C	A	G
Pooled Fetal Brain [11]	T	G	A	n/a	n/a	n/a	C	C	C	A	C	G	C	C	A	G
Pooled Fetal Brain [1]^	T	G	A	n/a	n/a	n/a	n/a	C	C	G	C	G	C	C	A	G
UNKNOWN																
Pooled Fetal Brain [1]^	T	G	G	G	T	C	C	C	C	G	C	G	G	T	A	A

Total clones sequenced: Ancestral *SRGAP2A*: SH-SY5Y = 3 clones; pooled fetal brain = 8 clones; Duplicates *SRGAP2B/C/D*: SH-SY5Y = 5 clones; pooled fetal brain = 35 clones; single fetal brain = 28; single adult brain = 17.

Colored boxes represent types of substitutions compared to the ancestral state (from chimpanzee sequence): red = nonsynonymous; blue = synonymous; yellow = noncoding. The green boxes including "n/a" represent variants not represented in transcript due to splicing and/or deleted exons. The empty gray boxes are variants that did not have high-quality coverage during sequencing of clones. The numbers in brackets are the total transcripts observed from the tissue source.

*These cDNA clones were used in functional assays in the accompanying manuscript [35].

^ Transcript only observed once in any tissue type.

Table S4. Copy Number Variants Detected in Cases with Developmental Delay and Controls, Related to Table 2

ID	Sex	Age	Cytoband	Genome coordinates (NCBI36)	Size (bp)	Type	Phenotype	Inheritance	Notes
9906627 Signature Genomics	F	23 months	1q32.1 - 1q32.2	chr1:203,821,004 -208,526,029	4,705,025	Deletion	Developmental delay, dysmorphic features, and multiple congenital anomalies	*De novo*	No other large CNVs in the genome
GC21416 Signature Genomics	F	3 years	1q32.1 - 1q41	chr1:203,845,072 -215,071,039	11,225,967	Deletion	Developmental delay, growth delay, hypotonia, and carnitine deficiency	*De novo*	No other large CNVs in the genome
9881737 Signature Genomics	M	6 months	1q32.1 - 1q32.2	chr1:204,164,223 -205,670,688	1,506,465	Deletion	Marked hypertelorism in both baby and father. Also, father has neurocognitive defects (may be from serious traumatic brain injury at age 11)	Paternal	No other large CNVs in the genome
9900886 Signature Genomics	F	10 years	1q32.1 - 1q41	chr1:204,650,122 -213,516,382	8,866,260	Duplication	MRI showed asymmetry: stable, abnormal left hippocampus, prominence of the left anterior temporal horn, probable mild left periventricular leukomalacia, and stable hypoplasia of the posterior body of the corpus callosum. The patient has a history of seizures, ADHD, behavior problems, and learning disabilities	Unknown	Patient also has a 283 kbp duplication at 1q32.1 (2 Mbp proximal to described duplication) including *KISS1*, *GOLT1A*, *PLEKHA6*, *PPP1R15B*, and *PIK3C2B*, and a 7.7 kbp *NRG3* intronic deletion
9896507 Signature Genomics	M	Newborn	1q21.1-1q44	chr1:143,793,178 -247,169,918	103,376,741	Duplication	Patient exhibits multiple congenital abnormalities	Unknown	Mosaic (18/50 uncultured cells)
9885509 Signature Genomics	F	3 weeks	1q31.1-1q32.2	chr1:184,399,825 -208,376,099	23,976,275	Duplication	Patient exhibits multiple congenital abnormalities	Unknown	No other large CNVs in the genome.
02753 Intellectual Disability	M	3 years and 10 months	1p12-1p11.2	chr1:120,498,679 -121,186,957	688,278*	Duplication	Patient shows regression of speech and behavior. Psychological evaluation led to the diagnosis of intellectual disability and generalized development disturbance	Unknown	No other large CNVs in the genome
12523.p1 SSC Autism	M	6 years and 1 month	1p12-1p11.2	chr1:120,498,679 -121,186,957	688,278*	Duplication	Patient is cognitively normal but his adaptive behavior skills fall in the low range; he reportedly shows significant hyperactivity and inappropriate speech. ADOS score = 9 (on 1-10 scale; >4 clinical; 10 most impaired)	Paternal	Patient also has 1.8 Mbp deletion at 11p14.3 including *AN05*, *SLC17A6*, *FANCF*, *GAS2*, and *SVIP*
13398.p1 SSC Autism	M	8 years and 7 months	1p12-1p11.2	chr1:120,498,679 -121,186,957	688,278*	Duplication	Patient shows attentional deficits, anxiety/depression, and aggression. ADOS score = 9 (on 1-10 scale; >4 clinical; 10 most impaired)	Maternal	Maternal side has history of migraines, eating disorder, obsessive compulsive disorder, and post-traumatic stress syndrome. No other large CNVs in the genome
150-10596 NIMH control	M	73 years	1p12-1p11.2	chr1:120,498,679-121,186,957	688,278*	Duplication	This control is self-reported as neurologically normal	Unknown	No other large CNVs in the genome

*Due to several gaps and incorrect annotations of the human reference 1p12.1-1p11.2 genomic region (e.g., inversion and a gap within *SRGAP2C*; see **Figure 1C** in main text), there is uncertainty in these size estimates. They are likely greater than reported here and, notably, include numerous genes.

Table S5. *SRGAP2* Paralog-Specific BAC Clones, Related to Experimental Procedures

***SRGAP2* Contig**	**GenBank Accession number**	**Clone**	**Sequence status at time of publication**
SRGAP2A	AC244035	CH17-84K15	complete
SRGAP2A	AC244158	CH17-67I7	complete
SRGAP2A	AC244017	CH17-255A18	complete
SRGAP2A	AC244016	CH17-251H16	complete
SRGAP2A	AC244023	CH17-465H19	complete
SRGAP2A	AC244018	CH17-286M8	complete
SRGAP2A	AC244024	CH17-67D14	complete
SRGAP2A	AC244019	CH17-397E23	complete
SRGAP2A	AC244159	CH17-94O18	complete
SRGAP2A	AC244034	CH17-76K2	complete
SRGAP2B	AC243754	CH17-61O17	fragmented working draft
SRGAP2B	AC241585	CH17-195P21	complete
SRGAP2B	AC244020	CH17-400H17	complete
SRGAP2B	AC242498	CH17-254B7	complete
SRGAP2B	FP700111	CH17-164J11	contiguous working draft
SRGAP2C	AC241377	CH17-118O6	complete
SRGAP2C	FP700108	CH17-465D18	complete
SRGAP2C	AC244453	CH17-469K7	complete
SRGAP2C	AC240103	CH17-366F13	complete
SRGAP2C	AC243994	CH17-219N22	complete
SRGAP2C	AC244021	CH17-437K3	complete
*SRGAP2D**	AC244015	CH17-248H7	complete

* An additional *SRGAP2D* BAC clone (CH17-266P3; GenBank Accession Number AC246680), not included in our analysis, was sequenced after acceptance of the manuscript. This clone verifies the *SRGAP2D* deletion spanning across exons 2 and 3.

Table S6. *SRGAP2* Primers, Related to Experimental Procedures

Experiment	Forward Primer		Reverse Primer	
	***SRGAP2* Target**	**Sequence**	***SRGAP2* Target**	**Sequence**
Validate discordant *SRGAP2*-containing BACs	Intron 2	GGATTGGCTTTGATTGCTGT	Intron 3	TGGGGGTCTGGTGTACAGAT
	Intron 1	CATGTTTGCATGTGGTAGGC	Intron 1	CTCAGAGCAACCAGGGAGTC
SRGAP2B paralog-specific qPCR	Intron 2	AGACCTCTACTTCTCAATGCCTCA	Intron 2	TGTGCACACATTTTAACACTTGG
SRGAP2C paralog-specific qPCR	Intron 6	GTAAGTGCCGTGTACATGTATGG	Intron 6	AAATGGGTGTTTCACAGTTCAGG
	Intron 7	CGGACCACTGTCAAAGCATA	Intron 7	GGCAGAAGAGTGAGCTAGCAG
SRGAP2D paralog-specific qPCR	Intron 6	GACAACACCAGATAAACCTGAAAAC	Intron 6	TTCAACGGTTAAACACACCCTAC
SRGAP2A paralog-specific qPCR	Intron 12	TCAGTTCCTTGGCTGAAACC	Intron 12	TGCCAAACTGATGTCTCTGG
Sequencing *SRGAP2A* transcripts	Exon 1	CATGTTGTGCGGAAGGACT	Exon 10	TGCCCTCCAGGTACTCTTTC
Sequencing *SRGAP2* duplicate transcripts	Exon 1	(same as above)	Intron 9 extension	TCTGAGTATGCCCACATTCG
	Exon 5	CCAGTCAACTGCTGGAATCTC		
SRGAP2A-specific RT-PCR	Exon 8	TCCACTCTAAAGATTGAAAACGAA	Exon 10	(same as above)
	Exon 21	CGGCTGGATAGTCCACAGAT	Exon 22	TGCCGTTCTAGTTCCCGTAG
SRGAP2-duplicate-specific RT-PCR	Exon 8	(same as above)	Intron 9 extension	(same as above)
SRGAP2C-specific RT-PCR	Exon 7*	AAGGCCATCAAAGCCCA	Exon 8	TGCACCAGATTCCTGAACAA
SRGAP2D NMD assay	Exon 1	(same as above)	Exon 5	CCGAGTAGAGCTCGTTCAGG
	Exon 1/4 junction	CGATACTCAGGTCAAAGAGTAA	Exon 5b	CCAATTTGCTTCTCCTCCTG
ALB qPCR control	n/a	GTGGGCTGTAATCATCGTCT	n/a	TGCTGGTTCTCTTTCACTGAC
GAPDH RT-PCR control	n/a	AGCCACATCGCTCAGACACC	n/a	GTACTCAGCGCCAGCATCG

* This oligonucleotide contains a *SRGAP2C* PSV.

Appendix B. Supplemental Information for Chapter 3

MIP Design

SRGAP2 exon-targeting MIPs were designed as previously described [1]. MIPs used for paralog-specific copy-number inference were designed in a similar fashion, with the following additional considerations. Careful selection of paralogous regions to target for MIP capture is critical for applying our copy-number genotyping method to any particular gene family of interest. Suitable target regions contain genetic variation between paralogs that has fixed in the human species. Obtaining paralog-specific read counts from a targeted region requires that the region contain genetic variation such that at least one paralog can be distinguished from all others. Ensuring that these counts reflect underlying relative paralog-specific copy numbers demands that variants used for distinguishing paralogs have very low levels of polymorphism.

To identify regions containing paralog-distinguishing variation, we aligned *SRGAP2* sequences [2], *RH* sequences (GRCh37/hg19, chr1:25594516–25655519 and chr1:25688914–25751819) and *RHD* flanking segmental duplications (GRCh37/hg19, chr1:25585374–25594516 and chr1:25655517–25664845) using Clustal 2.1 [3]. Regions of the alignments where sequences identical between all paralogs (20 bp each side) flanked a 112-bp region where at least a single paralog had a distinct sequence were selected as potential targets and input to the MIP design pipeline [1]. This pipeline attempted to design MIPs to capture each of these potential target regions, outputting MIP oligonucleotides, information about their corresponding arm hybridization sequences and their capture targets, and scores corresponding to their predicted capture performances. We eliminated from consideration any MIPs determined to have arm hybridization sequences with copy counts in the genome (GRCh37 augmented with *SRGAP2* contig sequences) >8 to avoid capturing repeat sequences and to restrict MIP hybridization to *SRGAP2* and *RH* loci. We also ensured that all MIP arm hybridization sequences were complementary to sequences identical between all paralogs of interest—any MIPs not meeting this criterion were eliminated from further consideration. Finally, we eliminated from consideration all MIPs with the lowest design score (–1) and most MIPs having a target region with <35% or >55% GC content.

For remaining MIPs under consideration for design, we analyzed polymorphism at potential SUNs within the corresponding capture target regions. Briefly, potential SUNs distinguishing each paralog were extracted from the alignment and scored with regard to likely fixation status via analysis of 12 high-coverage genomes (Supplementary Table 1). For each *SRGAP2* and *RH* paralog, we computed all 30-mer sequences found within that paralog and absent from the rest of the genome (singly unique nucleotide *k*-mers [SUNKs, [4]]). We then mapped 12 unrelated high-coverage genomes to *SRGAP2* and *RH* paralog sequences (masked using RepeatMasker [5] and Tandem Repeats Finder [6]) using mrsFAST [7] and parsed mapping output to assess the presence of each SUNK in each genome analyzed. A SUNK was considered present if observed in at least a single read mapped with no mismatches. Using only high-coverage genomes for this analysis minimizes the possibility of simply not having sequenced SUNKs that are truly present in a genome. Because each potential SUN typically contributes to (as a single base in the sequence of) many 30-mer SUNKs, the presence or absence of such SUNKs can serve as a proxy for the presence or absence of each potential SUN. Thus, a score from 0 to 12 was calculated for each potential SUN, corresponding to the average number of high-coverage genomes supporting a potential SUN's presence (Supplementary Fig. 8). For example, if a particular potential SUN contributed to four different 30-mer SUNKs, and these SUNKs were determined to be present in 11, 9, 11, and 12 high-coverage genomes, respectively, the score for that potential SUN would be 10.75 ((11 + 9 + 11 + 12)/4). True SUNs are paralog-distinguishing SNVs that have fixed in the human population. We defined potential SUNs having scores ≥11 (for *SRGAP2A*, *SRGAP2B*, and *SRGAP2C*) or ≥8 (for *SRGAP2D*, *RHD*, and *RHCE*) as true SUNs. (The threshold is lower for these latter paralogs owing to a paucity of higher-scoring potential SUNs across the spatial extent of duplicated sequence, reflecting in part heterozygous *SRGAP2D* and *RHD* deletions in some of the individuals sequenced to high coverage.) Biologically, these defined true SUNs are most likely to be fixed in a particular paralog in the human species and thus most

useful for copy-number genotyping. Given the observation that the majority (84.8%) of putative autosomal SUNs genome-wide were present in 12 of 12 high-coverage genomes previously analyzed [4], however, SUN scoring, though useful, is not necessary for successful application of our method.

MIPs used for copy-number genotyping were selected from remaining MIPs under consideration on the basis of the paralog-specificity, SUN content, and relative genic location of their corresponding target regions. *SRGAP2A* and *SRGAP2C* were prioritized in the *SRGAP2* copy-number genotyping MIP design owing to the likely pseudogenicity of *SRGAP2B* and *SRGAP2D* [2]. All MIPs were ordered from Integrated DNA Technologies as previously described [1]. Supplementary Table 2 provides specific details regarding MIPs designed for this study and their pooling.

MIP pooling, 5′ phosphorylation, and multiplex capture

MIPs were pooled (Supplementary Table 2), phosphorylated, and used to capture targeted sequences as previously described [1], with the following modifications. Initial capture reactions used in the 48-individual experiment were performed with genomic DNA input levels of 50 ng and 100 ng, with subsequent reactions involving HapMap samples using 200 ng DNA input and the reaction involving sample Troina2665 using 100 ng DNA input. MIPs were added to capture reactions at a ratio of 800 MIP copies per haploid genome copy. Incubation of capture reactions at 60 °C was performed for 23–24 h, and incubation of exonuclease reactions at 37 °C was performed for 45 min. Supplementary Table 9 summarizes MIP capture experiments performed for this study and details sample sets assayed.

Amplification, barcoding, pooling, cleanup, and sequencing

Captured sequences were amplified, barcoded, pooled, and purified as previously described [1], with the following specifications. PCR was performed in a 25-μL reaction. Libraries with excessive off-target captures were not observed; thus, the standard Agencourt purification protocol was followed. Final library DNA concentrations were quantified using the Qubit dsDNA HS assay (Life Technologies). Sequencing of pools of capture reactions was performed using either a MiSeq or a HiSeq 2000, depending on the number of individual capture reactions included in the pool for sequencing and the number of MIPs used in each individual capture reaction. Supplementary Table 9 provides specific details regarding sequencing performed for different MIP capture experiments in this study.

Paralog-specific copy-number genotyping

Initial 151-bp reads (MiSeq) or 101-bp reads (HiSeq 2000) were trimmed from their 3′ ends to 76 bp to eliminate low-quality data from the ends of reads while ensuring coverage of each targeted base in nearly all cases. All MIPs are designed such that a 152-bp region (target sequence plus hybridization arms) is sequenced. With 151-bp reads, all bases except the first and last base in this 152-bp region are sequenced during both the forward and reverse reads. Thus, retaining only the first 76 bp from each read eliminates low-quality data from the ends of reads while ensuring coverage of each targeted base in all cases except those in which there is a net insertion. Trimmed reads were mapped to *SRGAP2* and *RH* paralog sequences using mrFAST 2.5 [8] in paired-end mode with the maximum allowed edit distance set to 4 and the minimum and maximum inferred distances allowed between paired-end sequences set to 144 and 160, respectively.

Mapping output was parsed to yield counts of reads mapping to each paralog for each MIP for each individual. The following stringent filters were applied to ensure accuracy: the mapping location of a read pair was required to be within 4 bp of the expected mapping location, the strandedness of reads had to be consistent with expectation based on MIP design, the inferred insert size had to be within 2 bp of its expected value (152 bp), any bases covered by forward and reverse trimmed reads had to have the same base call, the quality scores at all base positions showing variation between paralogs (base positions that affect mapping paralog-specificity) had to be at least Q30, no mismatches could occur at likely fixed true SUN positions, and reported barcode sequences had to perfectly match a known barcode sequence. Read pairs violating any of these filters were not included in final counts. For *SRGAP2* paralog-specific copy-number analysis, final counts served as input to a genotyping program that generated *SRGAP2* paralog-

specific copy-number calls for each individual across the spatial extent of duplicated *SRGAP2* sequences. For *RH* paralog-specific copy-number analysis, genotyping calls were made in a similar automated fashion, except no copy-number state transitions were allowed. Thus, all internal *RH* gene conversion events were called on the basis of manual visual inspection of paralog-specific count frequency plots.

The *SRGAP2* genotyping program generates paralog-specific copy-number calls using a maximum-likelihood approach together with dynamic programming. For each individual, log-likelihoods of observing the paralog-specific read-count data for each MIP are calculated under 400 different possible hidden underlying *SRGAP2* paralog-specific copy-number states, where *SRGAP2A* and *SRGAP2C* can have copy numbers from 0 to 3 and *SRGAP2B* and *SRGAP2D* can have copy numbers from 0 to 4 ($4 \times 5 \times 4 \times 5 = 400$ combinations) [2]. For each paralog-specific copy-number state, log-likelihoods were calculated as logarithms of multinomial probabilities. Specifically, for each paralog-specific copy-number state, a multinomial probability of the observed data was computed for each MIP, with the number of trials equal to the total number of mapped reads for that MIP, and the vector of outcome probabilities equal to the copy numbers of specific paralogs over the aggregate copy number for the gene family given the paralog-specific copy-number state. (An outcome in this case is observing a read coming from a particular paralog.) The *SRGAP2D* internal deletion is built into the log-likelihood calculations: all copy-number states for MIPs in this region have *SRGAP2D* copy number set to 0. Log-likelihood values below −30 are set to −30 to limit the ability of count data from a single MIP to potentially single-handedly invalidate a particular *SRGAP2* paralog-specific copy-number state as possibly underlying the count data.

Next, for each individual, log-likelihoods are used to construct a weighted directed acyclic graph, with prior probabilities based on observed *SRGAP2* copy-number genotype data from previous experiments [2] incorporated into the log-likelihoods for the first (most 5′ with respect to *SRGAP2*) MIP. The graph is constructed by iteratively considering log-likelihoods for the next MIP and tracking the highest scoring paths ending at each copy-number state allowing 0, 1, and 2 transitions between copy-number states as well as the values of the corresponding log-likelihoods of these paths until the graph spans all MIPs. Allowed transitions between copy-number states are restricted to copy-number gains or losses affecting a single paralog or cases where the copy numbers of two paralogs change, but the total number of *SRGAP2* copies remains constant. All transitions meeting these criteria and transition probabilities associated with remaining in the same state have probability 1; all other transitions have probability 0. Three highest-scoring paths through the likelihood graph are calculated: one for 0 allowed total transitions between copy-number states, one for 1 allowed transition between copy-number states, and one for 2 allowed transitions between copy-number states. Restricting the nature of allowed copy-number state transitions reflects the fact that true biological events should fall into one of two categories (single paralog–affecting duplication/deletion or interlocus gene conversion). Restricting the number of transitions reflects the fact that a single individual is most likely to have, at most, a single duplication, deletion, or interlocus gene conversion restricted to within *SRGAP2*. If an individual were to have multiple events restricted to within *SRGAP2*, the program would still flag this individual as having a complex *SRGAP2* paralog-specific copy-number genotype, and the second internal event would be apparent upon subsequent visual inspection of paralog-specific count frequency plots. The program ultimately identifies the highest-scoring paths through the likelihood graph (most likely paralog-specific copy-number states across the spatial extent of duplicated *SRGAP2* sequence) allowing 0, 1, and 2 transitions and their corresponding log-likelihood scores. Heuristics (Supplementary Table 10) are used to assess increases in likelihood of the one-transition and two-transition paths compared to the zero-transition path and to determine whether they signal an event within *SRGAP2* and warrant calling the paralog-specific copy-number genotype for an individual as complex. In most cases, the scores of the one-transition and two-transition paths will not be substantially higher than that of the zero-transition path, and the genotype for an individual will be called as simple (a single copy-number state across the entirety of duplicated *SRGAP2* sequence).

The program also calculates a logarithm of odds confidence score associated with the simple (zero-transition) genotype call for each individual. Specifically, this score is equal to the log-likelihood of the chosen zero-transition path minus the highest log-likelihood for a zero-transition path having a distinct

set of associated multinomial probabilities. For example, if an individual was called as having two copies of each *SRGAP2* paralog, the confidence score would be the log-likelihood of the zero-transition path for this copy-number state minus the highest log-likelihood of a zero-transition path among all other copy-number states except those having equal copy numbers for each *SRGAP2* paralog. The logic behind this requirement is that likelihoods of zero-transition paths with the same set of associated multinomial probabilities will differ only because they have distinct prior probabilities—that is, the paralog-specific read-count frequency data, independent of any prior knowledge, support each such path equally well. Confidence scores should be interpreted relative to confidence scores for other individuals genotyped for the same gene family using the same set of MIPs (Supplementary Table 7 and Supplementary Fig. 5) rather than in an absolute sense.

The *RH* genotyping program works the same way as the *SRGAP2* genotyping program, except that there are 25 different possible *RH* paralog-specific copy-number states (each of *RHD* and *RHCE* is allowed to vary in copy number from 0 to 4), prior probabilities used were based on our estimates of *RH* paralog-specific copy number from the 1000 Genomes Project (1KG) data, and no transitions between copy-number states were allowed, such that all *RH* genotypes are called as simple (a single copy-number state across the entirety of duplicated *RH* sequence).

Fluorescence *in situ* hybridization

Metaphase spreads were obtained from lymphoblast and fibroblast cell lines from human HapMap individuals NA19700, NA19703, NA19901, NA20127, NA20334, NA19005, NA19190, NA19201, and NA12878 (Coriell Cell Repository). FISH experiments were performed using fosmid clones (WIBR2-2926C23_G248P88292B12 and WIBR2-3738J10_G248P802587E5) [2] directly labeled by nick translation with Cy3-dUTP (Perkin-Elmer), Cy5-dUTP (Perkin-Elmer), and fluorescein-dUTP (Enzo) as described previously [9] with minor modifications. Briefly: 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37 °C in 2× SSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 μg sonicated salmon sperm DNA in a volume of 10 μL. Posthybridization washing was at 60 °C in 0.1× SSC (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, Cy3, Cy5 and fluorescein fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

Other orthogonal validations

Array CGH data, qPCR data, and whole-genome shotgun sequence data from the 1KG used for validation purposes were collected and processed as previously described [2, 4].

Paralog-specific SNV genotyping

We trimmed initial reads from their 3′ ends to 100 bp to eliminate some low-quality data from the ends of reads while ensuring coverage of each targeted base. Trimmed reads were mapped separately to individual *SRGAP2* paralog sequences using the Burrows-Wheeler Aligner [10] (v.0.5.9, paired-end mapping) with the following options: -e 50 -l 17 -q 20 -d 5 -i 5 -I. Mapping output, combined with each individual's *SRGAP2* paralog-specific copy-number genotype determined as described above, was parsed to yield sequence genotypes for each copy of each paralog for each MIP for each individual. The Hungarian method [11] was used to optimally assign distinct sequences to copies of different *SRGAP2* paralogs according to observed counts of distinct sequences, treating paralog-specific mapping edit distances as the costs of sequence assignments to copies of different paralogs. In cases where equally optimal but biologically distinct sets of assignments could be made, each assignment set and its corresponding paralog-specific sequence genotypes was reported and flagged as having some ambiguity. All detected SNVs were annotated with regard to location and likely functional impact. Reported SNVs in Supplementary Table 6 have the following properties: (i) they were called on the basis of MIP sequence data from NA18507, (ii) they occur at alignment positions where all paralogs share the same nucleotide,

(iii) they occur in close proximity to a SUN or on a paralog-distinguishing haplotype such that their paralog-of origin can be accurately inferred, (iv) they were unambiguously assigned to a particular paralog, and (v) all copies of the paralog they were assigned to have no ambiguity in sequence at the variant site.

Preparation of final *SRGAP2* MIP pool

Data from the 48-individual genotyping experiment revealed that most MIPs in the initial pool captured their corresponding targets well (Supplementary Fig. 9). Considering only the capture reactions using 100-ng DNA input, the mean and median mapped read counts per MIP were 18,447 and 15,809, respectively. On average, this translates to approximately 350 mapped reads per MIP per individual. To optimize our MIP pool before extending our genotyping efforts to thousands of samples, we rebalanced exon-targeting MIPs that failed to efficiently capture their corresponding targets and removed *SRGAP2* copy-number genotyping MIPs that did not meet a high performance standard.

Specifically, we increased the amount of any exon-targeting MIPs having a total mapped read count lower than 2,500 times the number of paralogs that include the targeted exon. For example, if a MIP targeted exon 1, shared between all four *SRGAP2* paralogs, but had fewer than 10,000 reads from the initial capture reactions using 100-ng DNA input, we rebalanced it. Rebalancing was performed such that this count threshold would be achieved if mapped read count per MIP increases proportionally with the amount of MIP added to the pool: for example, if doubling the amount of MIP added to the pool results in twice the number of corresponding mapped reads. We thus added seven exon-targeting MIPs to achieve a relative amount of 2× in the final pool and another five such MIPs to achieve a relative amount of 5×. Eleven MIPs, however, would still fail to meet the count threshold even if their corresponding mapped read counts increased fivefold. These worst-performing exonic MIPs were added to achieve a relative amount of 50× in the final MIP pool to maximize their chances for successful capture.

To evaluate the performance of *SRGAP2* copy-number genotyping MIPs, we compared observed paralog-specific count frequencies for each MIP with corresponding expected frequencies for 31 genomes from the 48-individual experiment (Supplementary Table 3). These genomes were selected because we had very high confidence in their true paralog-specific copy-number genotypes: genotyping results were concordant between all methods used for 30 of these genomes, and FISH results supported MIP results in the remaining case. For each *SRGAP2* copy-number genotyping MIP, we calculated the mean and s.d. of per-genome error in paralog-specific count frequencies (Supplementary Fig. 1). We removed MIPs having mean per-genome errors ≥ 0.125 or corresponding s.d. ≥ 0.25 from our final set, with a few exceptions. For example, we retained some MIPs in the *SRGAP2C* deletion region having mean errors or s.d. slightly above these values because we wanted to maximize our power to genotype this event. Reducing the number of MIPs used for genotyping *SRGAP2* in our final pool in this manner increases our capacity to assay additional genes of interest and larger numbers of individuals in the same experiment while ensuring *SRGAP2* genotyping remains highly accurate. Selection of a high-performing final MIP set from all initial MIPs tested, though useful for increasing multiplexing potential, is not necessary for successful application of our method (Supplementary Fig. 2).

Cost estimation

The approximate cost per gene per individual associated with using MIPs can be estimated as follows. Each MIP is 70 bp, and each synthesized base costs $0.09. Thus, each MIP costs $6.30. We usually multiplex ~2,000 MIPs in a single MIP pool, so the cost of generating a typical MIP pool is $12,600. Because a very small amount of the MIP pool is used in each capture reaction, a single order of oligonucleotides can be used to assay tens of thousands of samples; thus, we assume this cost is effectively fixed (i.e., independent of the number of samples tested). The oligo cost per sample thus depends on the number of samples tested. Assuming we assay 4,000 samples, for example, the per-sample oligo cost is $3.15. The cost of reagents associated with the experimental protocol is $2.57 per sample. The cost of a lane of sequencing using the HiSeq is $1,388. We have found that up to 192 samples can be multiplexed per lane to obtain high coverage per MIP per sample under the assumption that the MIP pool

used in capture experiments contained 2,000 MIPs. Thus, the sequencing cost per sample is approximately $7.23. Adding these results, we obtain a cost of $12.95 per sample. Assuming each gene can be effectively assayed by 50 MIPs, on average, each MIP pool covers 40 genes. Thus, the final cost per gene per sample in this scenario is ~$0.32. If we eventually assay 10,000 samples using this same MIP pool, the final cost per gene per sample works out to $0.28. Even if we were to assay only 1,000 samples, the final cost per gene per sample would still be less than $1 (~$0.56).

Internal deletion and duplication genotyping by WGS

We leveraged data from the 1KG to evaluate WGS-based discovery of novel structural variation within duplicated genes and to compare these results with our MIP data. Specifically, we genotyped *SRGAP2B* copy number in an individual genotyped by MIPs as having two copies of *SRGAP2B* with an 83-kbp internal *SRGAP2B* duplication, and we genotyped *SRGAP2C* copy number in seven individuals genotyped by MIPs as having two copies of *SRGAP2C*, one harboring the 38-kbp internal deletion. All WGS-based paralog-specific copy-number estimates [4] for these individuals were 2; however, specifying the regions affected by these events before genotyping allowed for successful identification of the internal events in 7 of 8 cases (Supplementary Table 11). These data provide additional support for the internal duplications and deletions called by our MIP-based method and suggest that naïve paralog-specific copy-number genotyping using low-coverage WGS data cannot reliably discover them.

Application of our method to other gene families of interest

To use our method to study a gene family of interest other than *SRGAP2* or *RH*, one would first need to obtain accurate genomic sequences for as many paralogs as possible. Having reliable sequence data for all paralogs allows MIP design to be optimized to achieve complete paralog specificity and maximize genotyping power. Second, one would align paralogous sequences and identify SUN-containing regions to guide MIP design. We provide a program (https://github.com/xnuttle/mips_cnv_typer/) to identify such regions from aligned sequences. Third, one would attempt to design MIPs to all such regions as well as exons using the publicly available MIP design software [1], select a final set of MIPs according to criteria detailed above, and order them from a commercial oligo provider. Fourth, one would perform the MIP experiments and analyses described (https://github.com/xnuttle/mips_cnv_typer/) to obtain genotypes for each duplicated segment. If possible, we recommend testing every new MIP set on a panel of genomes having known paralog-specific copy numbers to ensure accuracy and reproducibility.

We found *RH* more difficult to genotype for copy number than *SRGAP2* using our approach. Two factors likely contribute to this observation. First, *RH* paralog-specific copy-number genotyping included data from only 35 MIPs, whereas that for *SRGAP2* incorporated data from either 90 or 142 MIPs. Second, fewer independent MIP capture events occur per genome for *RH* than *SRGAP2* because there are fewer total genomic copies of *RH* than *SRGAP2*. Thus, there are effectively fewer experimental trials for *RH* than *SRGAP2*, resulting in increased sampling error in *RH* paralog-specific read-count data. Designing more MIPs targeting *RH* and increasing DNA input would mitigate these issues and improve future *RH* genotyping performance. These issues warrant consideration in applying our method to other gene families of interest.

CCL3L1 [12-15], beta-defensins [16, 17], and *C4 [18]* present a few novel challenges for our method: (i) *CCL3L1* and beta-defensins are much smaller than *SRGAP2* and *RH*, such that only a few MIPs may be able to interrogate SUN-containing regions within them; (ii) unlike *SRGAP2* and *RH*, beta-defensins and *C4* have no obvious family member fixed or nearly fixed at diploid copy number 2 in the human population, making copy-number determination based on relative counts more ambiguous. For these gene families, it will be necessary to perform absolute in addition to relative read depth analysis, perhaps via singular value decomposition analysis as has been done to normalize exome capture variability from a large number of samples [19]. Another possible strategy would be to use some MIPs as MLPA probes (Supplementary Fig. 10) targeting genes of interest and regions of known invariant diploid copy number to calibrate aggregate or paralog-specific copy-number estimates on the basis of absolute

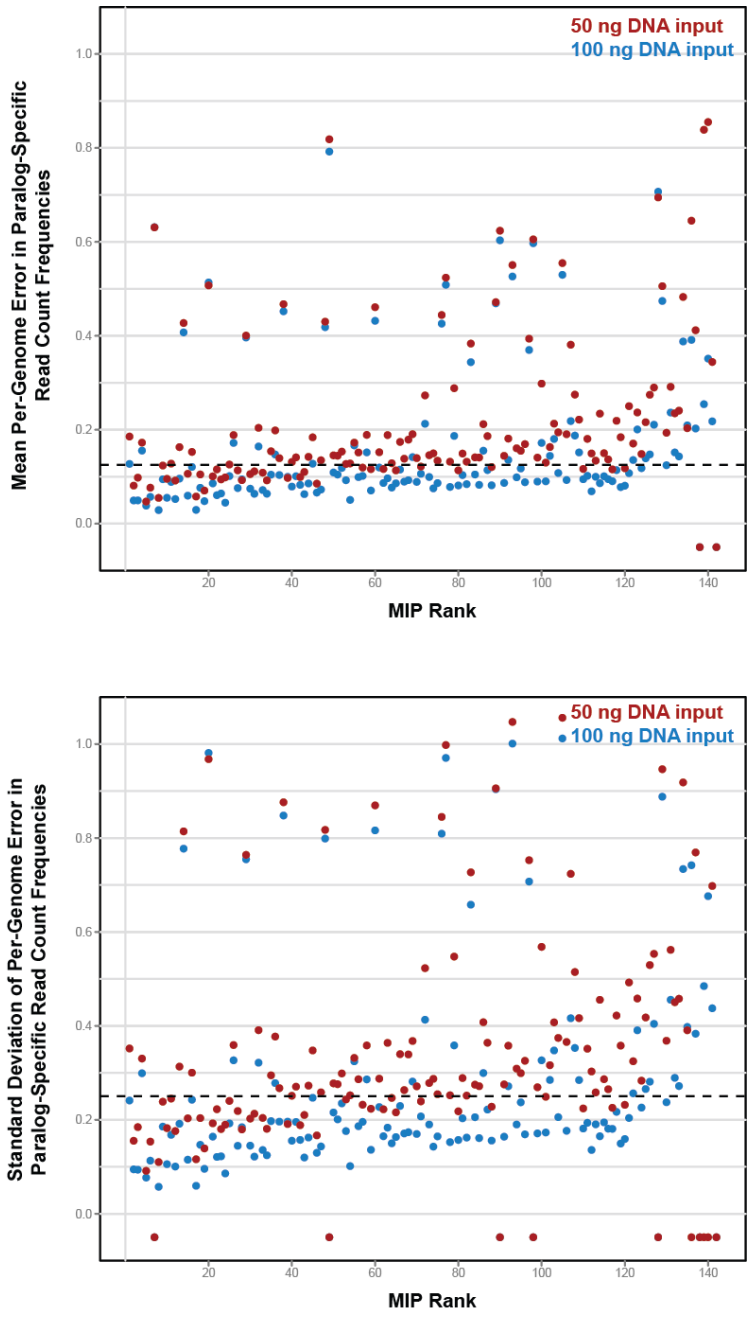
read depth data. In addition, genotyping copy number of the blocks of duplicated sequence containing *CCL3L1* and beta-defensins should provide accurate copy-number genotypes for these genes, as common copy-number variation at these loci occurs at the level of such blocks rather than affecting individual genes within them [16]. This approach leverages the much larger sample of SUNs these blocks contain compared to the genes themselves.

Identification of missing paralogous sequences in GRCh37

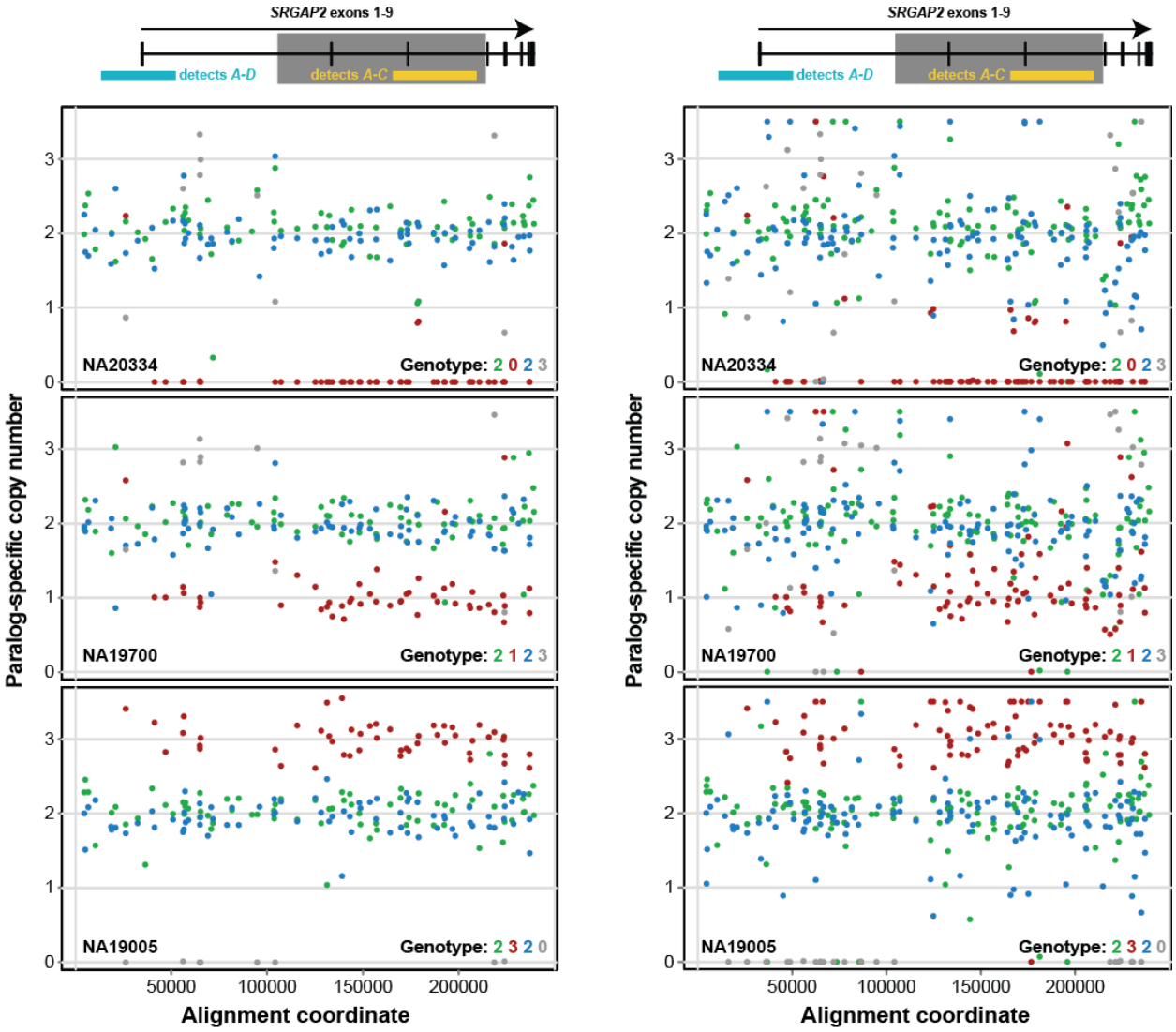
We genotyped regions in GRCh37 that had been previously described as missing paralogous sequence in NCBI36 [4] to identify regions of the reference genome still lacking complete sequence characterization. We successfully lifted over 326 of the original 333 regions from NCBI36 to GRCh37 and calculated paralog-specific copy numbers for each region with 885 individuals from the 1KG. A region was considered 'missing' from GRCh37 if the paralog-specific copy number for that region was greater than 2 for at least 90% of the individuals we genotyped. Using this definition, we found 21 regions that are still missing paralogous sequence in GRCh37. Comparing these regions with public NCBI patches to GRCh37 reveals that 7 of the 21 regions are completely covered by a fix patch and will likely be resolved in GRCh38.

Identification of SUNs from GRCh37

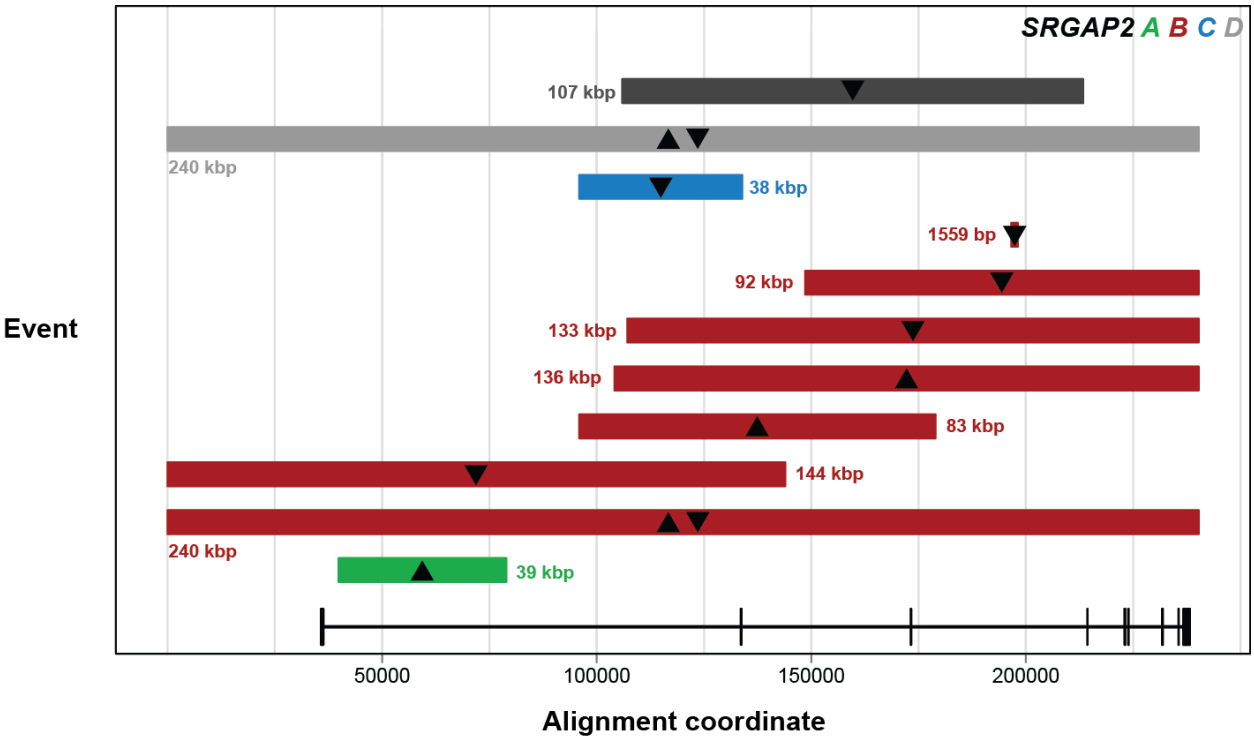
We used previously calculated SUNKs and segmental duplications for GRCh37 to calculate the set of all SUNs that uniquely identify individual segmental duplications. For each pair of segmental duplications, we globally aligned the corresponding sequences and identified all mismatches, insertions and deletions. We identified the diagnostic differences between related duplications by intersecting the coordinates of all differences with coordinates for SUNKs across GRCh37. With this approach we identified ~4 million SUNs. After filtering out any of these SUNs that were within 36 bp of repeats identified by RepeatMasker [5] or Tandem Repeats Finder [6], we identified ~3.8 million SUNs.



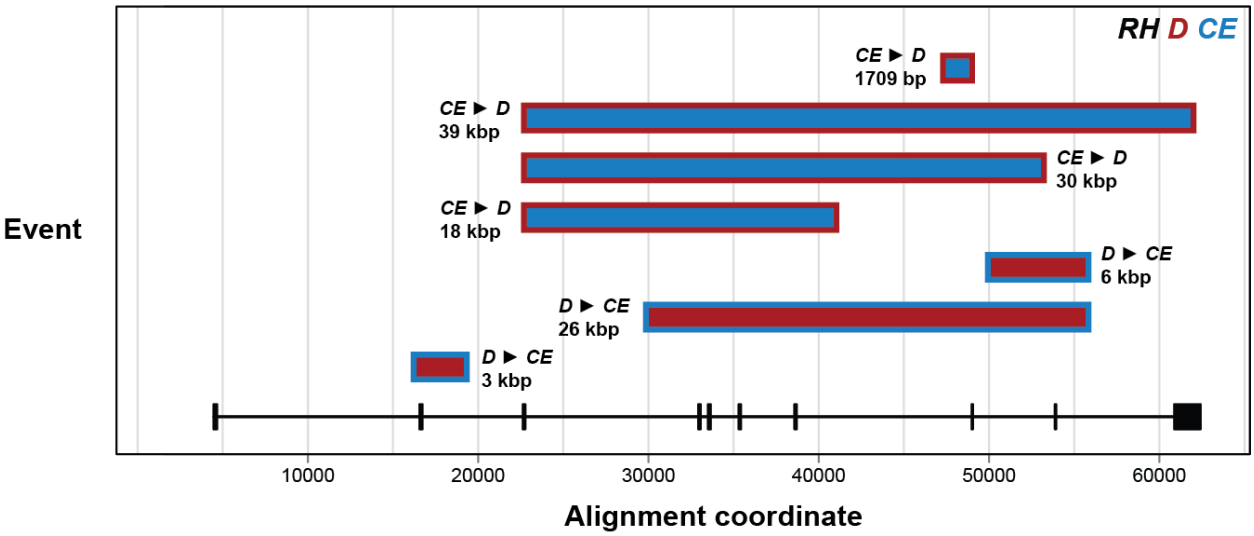
Supplementary Figure 1. Performance assessment of *SRGAP2* copy number genotyping MIPs. For a given genome assayed, error was calculated for each MIP as the sum of the absolute values of the differences between observed and expected mapped read count frequencies for each *SRGAP2* paralog and for a non-paralog-specific category (not all MIPs targeted sequence where all four *SRGAP2* paralogs can be distinguished). The per-genome means (top) and standard deviations (bottom) of these error values are plotted for each MIP using data from 31 individuals assayed in the initial 50 ng (red) and 100 ng (blue) replicate capture experiments. Negative plotted values correspond to mean errors and standard deviations of errors greater than 1.1. MIPs are ranked in the plot by total corresponding mapped read count in the 100 ng capture data for the 31 individuals, with MIPs having the highest such counts on the left. Dashed lines indicate thresholds we imposed in selecting MIPs for inclusion in our final pool. These error data highlight the increase in accuracy attained by using 100 ng of DNA rather than 50 ng of DNA for the capture reactions. Most likely, more independent capture events occur and sampling error accordingly declines with increased DNA input.



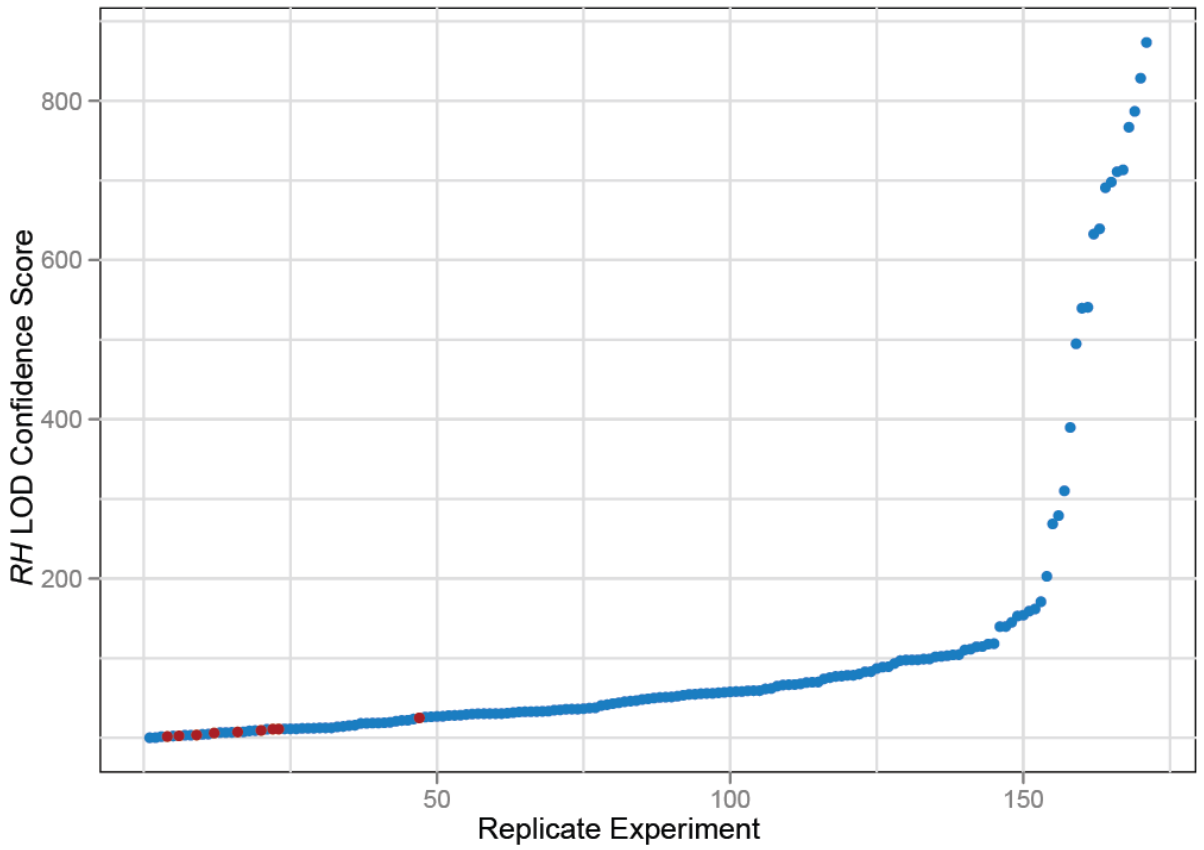
Supplementary Figure 2. Comparison of the full *SRGAP2* MIP set with the final selected set. The left panels show paralog-specific copy number estimates for 90 high-performing MIPs across ~240 kbp of aligned *SRGAP2* genomic sequence, as in Figure 2. The right panels show corresponding data for the full set of 142 MIPs. All values >3.5 were set to 3.5 for plotting purposes. Even though the right panels show more noise, the same automated genotype call (consistent with FISH) is made regardless of whether data from the final MIP set only or the full MIP set is considered. Extending this analysis, we compared genotype calls made from the same experiment using data from the full *SRGAP2* MIP set to those made using only data from the final selected set. With one exception, the genotypes were identical for 48 individuals tested when comparing the full set with the selected set. Interestingly, for the one discordancy orthogonal data supported the genotyping call from the full set as opposed to the selected set.



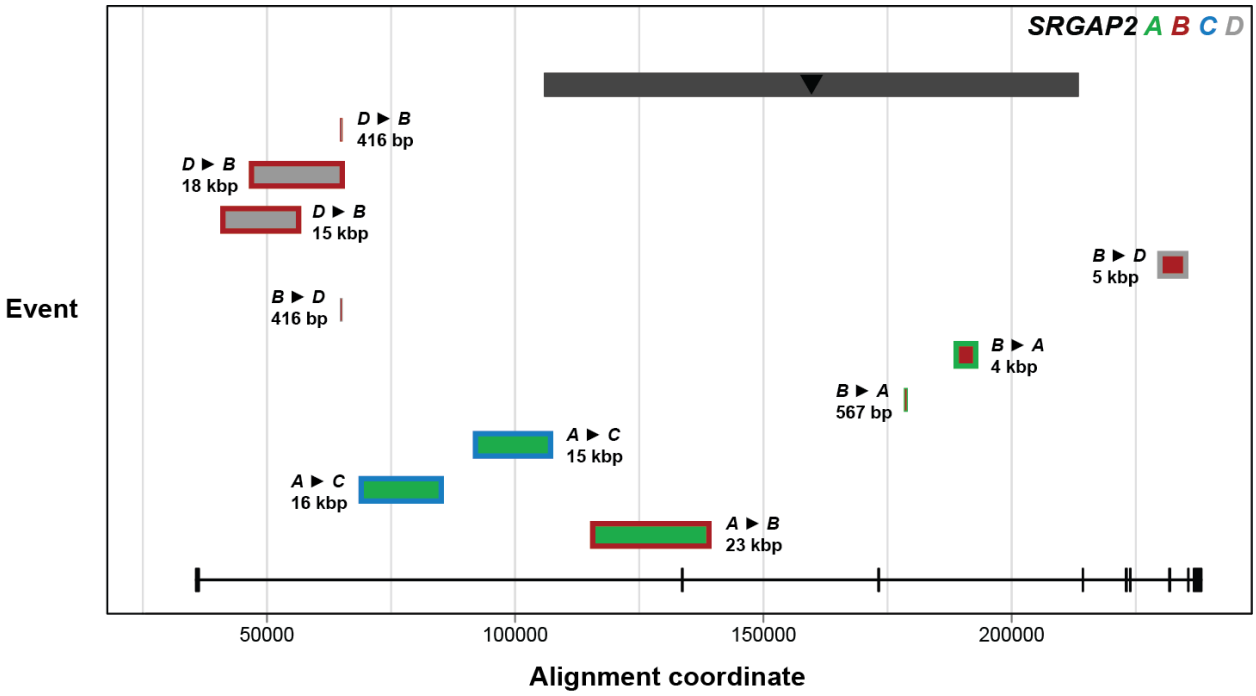
Supplementary Figure 3. Structural variation in *SRGAP2* paralogs. Locations of duplications (depicted by colored boxes with upward-pointing triangles) and deletions (depicted by colored boxes with downward-pointing triangles) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *SRGAP2* exons. Dashed lines indicate events that extend beyond the extent of duplicated sequence shared between all four *SRGAP2* paralogs. Reported approximate sizes of all events are minimum estimates, calculated as the number of base pairs between the centers of MIP target sequences for the 5'-most and 3'-most MIPs signaling each event (except for events extending beyond duplicated *SRGAP2* sequence, where *SRGAP2* duplication boundaries are used in this calculation). The precisions of these size estimates are governed by the spacing and paralog-specificity of MIPs targeting surrounding regions, but typically allow for breakpoint resolution within a few kbp to a few tens of kbp. The dark gray box depicts the *SRGAP2D* internal deletion. Its breakpoints are known with very high-precision from clone-based capillary sequencing [2].



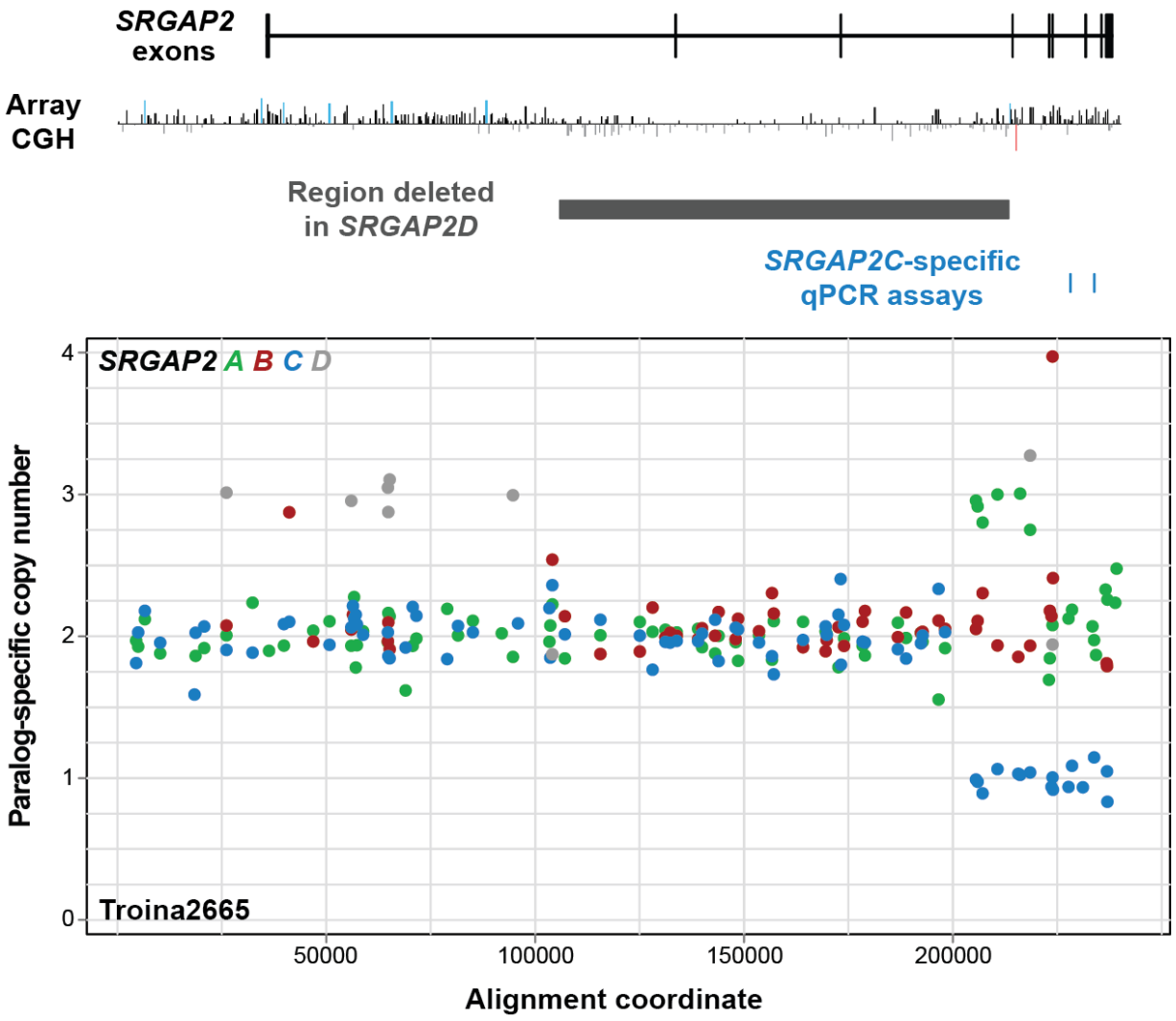
Supplementary Figure 4. Signatures of interlocus gene conversion in *RH* paralogs. Locations of putative *RH* interlocus gene conversion events (depicted by two-colored boxes) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *RH* exons (corresponding to *RHD* transcript variant 1). Inner fill colors indicate putative conversion donors, while border colors indicate corresponding putative conversion acceptors. Reported approximate sizes of all events are minimum estimates, calculated as described in the legend to Supplementary Fig. 3.



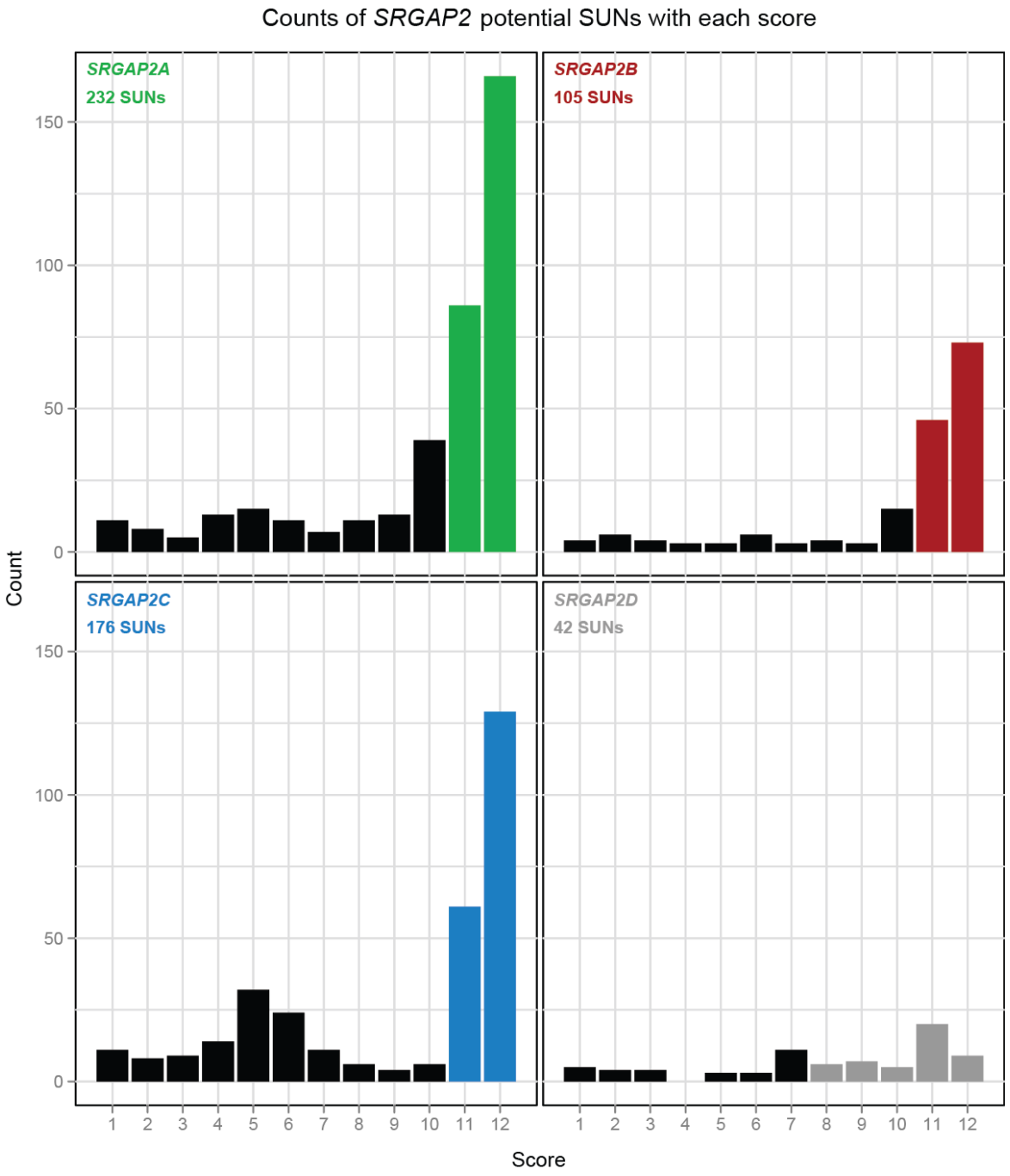
Supplementary Figure 5. Distribution of LOD confidence scores for MIP-based *RH* paralog-specific copy number genotypes from 171 replicate experiments. Discordancies are shown in red. The highest scores correspond to individuals having homozygous deletion of *RHD*. These data allow potential genotyping errors to be readily distinguished from high-confidence genotype calls.



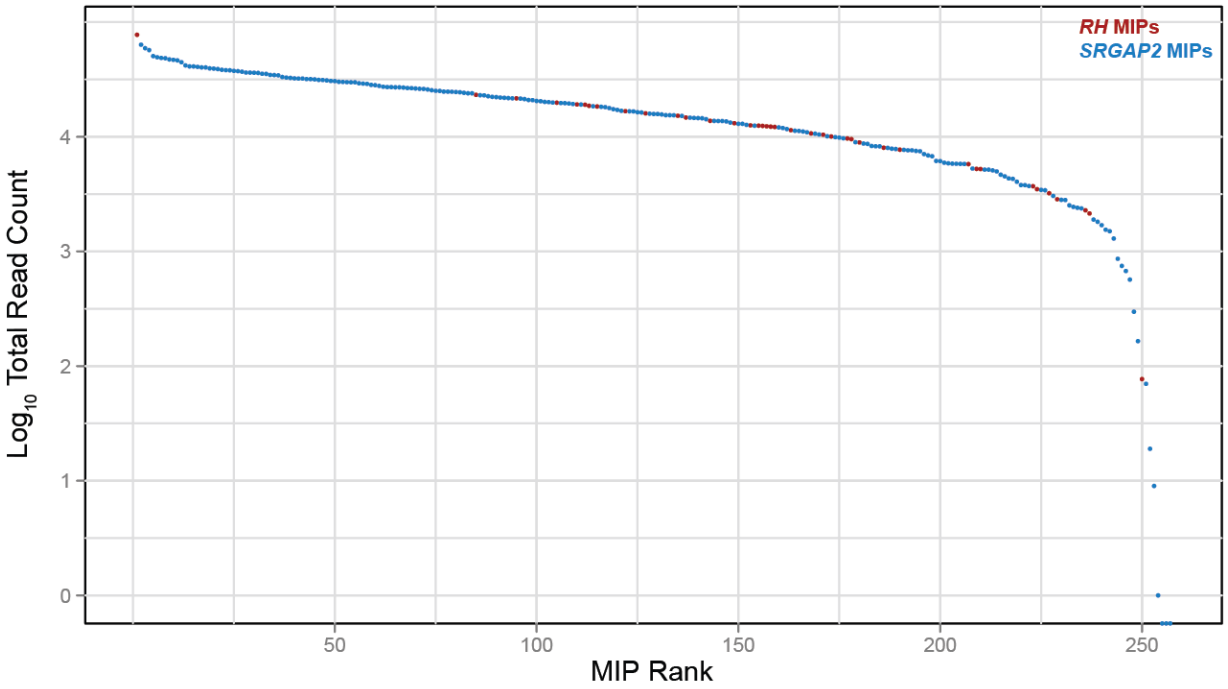
Supplementary Figure 6. Signatures of interlocus gene conversion in *SRGAP2* paralogs. Locations of putative *SRGAP2* interlocus gene conversion events (depicted by two-colored boxes) identified from MIP-based genotyping of 1,056 HapMap individuals are plotted relative to duplicated *SRGAP2* exons. Colors and reported sizes follow the convention described in Supplementary Fig. 4. The dark gray box depicts the *SRGAP2D* internal deletion. We note that our power to detect gene conversion events between *SRGAP2B* and *SRGAP2D*, paralogs having ~99.6% sequence identity both located within chromosome 1q21.1, was limited. This limited power largely reflects our prioritization of *SRGAP2A* and *SRGAP2C* in designing MIPs for copy number genotyping.



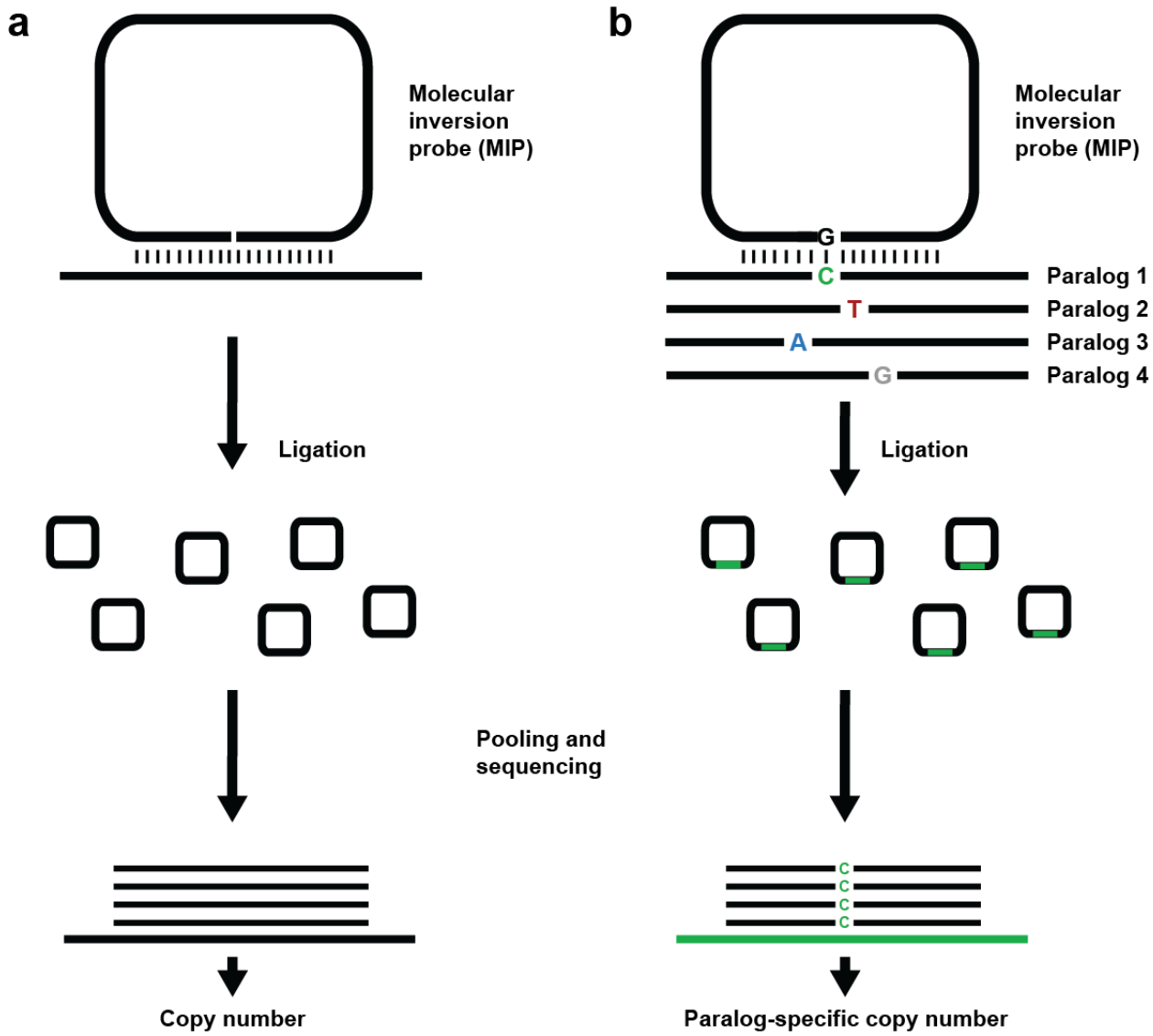
Supplementary Figure 7. Array CGH and qPCR validation of an interlocus gene conversion signature. The array CGH profile for *SRGAP2* loci predicts a gain in an individual with intellectual disability, likely involving *SRGAP2D* because the array signal disappears over the *SRGAP2D* internal deletion region. However, two independent *SRGAP2C*-specific qPCR assays targeting introns 6 and 7 predict a *SRGAP2C* deletion, a result seemingly inconsistent with the array data. MIP genotyping provides further support for the *SRGAP2D* duplication and suggests that gene conversion involving *SRGAP2C* as an acceptor explains the qPCR results. MIP data from this individual show evidence for multiple putative interlocus gene conversion events affecting the last few duplicated *SRGAP2* exons.



Supplementary Figure 8. Score histograms for *SRGAP2* potential SUNs. All *SRGAP2* potential SUNs having at least a single overlapping 30-mer SUNK were scored on a scale of 0-12. Scores were calculated as the sum over all 30-mer SUNKs overlapping the potential SUN of the number of high-coverage genomes analyzed supporting the SUNK's presence, divided by the total number of 30-mer SUNKs overlapping the potential SUN. This score can thus be interpreted as the average number of high-coverage genomes supporting a potential SUN's presence. Low scores reflect low allele frequency, sequence masking at or near a potential SUN position, or some combination of these factors, while high scores indicate a likely high potential SUN allele frequency and thus high value for copy number genotyping. The histograms show the distributions of potential SUN scores rounded to the nearest integer for all four *SRGAP2* paralogs. Colored bars correspond to potential SUNs defined to be true SUNs. Counts of SUNs scoring < 0.5 are omitted from the plot, as SUNs with these scores may indeed be present in several analyzed high-coverage genomes but could not be assessed due to sequence masking.



Supplementary Figure 9. Counts of total reads mapped to each MIP target for the 100 ng 48-individual capture experiment. All reads included in these counts passed all filters described in the copy number genotyping section of the Appendix B. All *SRGAP2*-targeting and *RH*-targeting MIPs are ranked in the plot by total corresponding mapped read count, with MIPs having the highest such counts on the left. These data provide insight into the relative capture efficiencies of different MIPs and inform MIP rebalancing. The tight distribution of total corresponding mapped read count values (within 1.5 logs for the 227 MIPs having highest such counts) suggests capture efficiency was fairly uniform between MIPs. MIPs having the fewest corresponding mapped read counts were almost all exon-targeting with the lowest design score (-1), used because no higher-scoring alternative MIPs could be designed that would still target the desired exonic sequence.



Supplementary Figure 10. MIP-based multiplex ligation-dependent probe amplification. a) MIP arms could be designed to hybridize to adjacent sequences, such that hybridization followed by ligation results in circularly closed molecules. Barcoding, pooling, and sequencing these molecules, mapping reads to corresponding reference sequence, and quantifying read depth should provide insights into copy number of targeted loci in a manner akin to MLPA. This approach would allow for up to ~2000 sites to be assayed in this manner simultaneously. Furthermore, these probes could be combined with conventional MIP probes in the same reaction. b) If the MLPA-MIP were designed such that the final base of one hybridization arm was complementary to a SUN, this assay might be able to achieve paralog-specificity.

Supplemental References

1. O'Roak, B.J., et al., *Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders.* Science, 2012. **338**(6114): p. 1619-22.
2. Dennis, M.Y., et al., *Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.* Cell, 2012. **149**(4): p. 912-22.

3. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2.3.
4. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes.* Science, 2010. **330**(6004): p. 641-6.
5. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4.10.
6. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
7. Hach, F., et al., *mrsFAST: a cache-oblivious algorithm for short-read mapping.* Nat Methods, 2010. **7**(8): p. 576-7.
8. Alkan, C., et al., *Personalized copy number and segmental duplication maps using next-generation sequencing.* Nat Genet, 2009. **41**(10): p. 1061-7.
9. Antonacci, F., et al., *A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk.* Nat Genet, 2010. **42**(9): p. 745-50.
10. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
11. Kuhn, H.W., *The Hungarian Method for the assignment problem.* Nav. Res. Logist. Q., 1955. **2**(1-2): p. 83-97.
12. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.* Science, 2005. **307**(5714): p. 1434-40.
13. Bhattacharya, T., et al., *CCL3L1 and HIV/AIDS susceptibility.* Nat Med, 2009. **15**(10): p. 1112-5.
14. Carpenter, D., et al., *Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders.* BMC Genomics, 2011. **12**: p. 418.
15. Nordang, G.B., et al., *Association analysis of the CCL3L1 copy number locus by paralogue ratio test in Norwegian rheumatoid arthritis patients and healthy controls.* Genes Immun, 2012. **13**(7): p. 579-82.
16. Groth, M., et al., *High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes.* Hum Mutat, 2008. **29**(10): p. 1247-54.
17. Aldhous, M.C., et al., *Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease.* Hum Mol Genet, 2010. **19**(24): p. 4930-8.
18. Fernando, M.M., et al., *Assessment of complement C4 gene copy number using the paralog ratio test.* Hum Mutat, 2010. **31**(7): p. 866-74.
19. Krumm, N., et al., *Copy number variation detection and genotyping from exome sequence data.* Genome Res, 2012. **22**(8): p. 1525-32.
20. Wang, J., et al., *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-5.
21. Park, H., et al., *Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing.* Nat Genet, 2010. **42**(5): p. 400-5.
22. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
23. Ahn, S.M., et al., *The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group.* Genome Res, 2009. **19**(9): p. 1622-9.
24. Schuster, S.C., et al., *Complete Khoisan and Bantu genomes from southern Africa.* Nature, 2010. **463**(7283): p. 943-7.
25. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
26. Fischbach, G.D. and C. Lord, *The Simons Simplex Collection: a resource for identification of autism genetic risk factors.* Neuron, 2010. **68**(2): p. 192-5.

Appendix C. Supplemental Information for Chapter 4

Table of Contents

1. Sequencing and assembly of the 16p11.2 region

We generated high-quality reference sequences over the 16p11.2 region for orangutan, chimpanzee, and multiple human haplotypes by sequencing and assembling large-insert clones using a previously described strategy [1]. We examined clone paired-end sequence mapping data [2] and/or performed hybridization experiments to identify bacterial artificial chromosomes (BACs) likely harboring sequence from the 16p11.2 region. We utilized three BAC libraries constructed from human genomic material: CH17 (from the complete hydatidiform mole CHM1 [3]), VMRC53 (from the HapMap female NA12878), and RP11 (from a male, the primary library sequenced as part of the Human Genome Project). We also used BAC libraries from chimpanzee (CH251, from a male named Clint) and orangutan (CH276, from a female named Susie). All candidate BACs were sequenced using a Nextera protocol [4] and massively parallel Illumina sequencing technology, and reads were mapped and analyzed as previously described [1, 5]. This procedure allowed us to select tiling paths of clones spanning the region. This process was complicated by the presence of segmental duplications having high sequence identity within the 16p11.2 region and between 16p11.2 and other chromosome 16 loci (**Fig. S1**). However, because the Nextera data provide sequence information across the entirety of the clone insert (~170 kbp in length) rather than merely at the ends, it was possible to distinguish truly overlapping clones from their allelic and paralogous counterparts and, thus, establish single-haplotype tiling paths.

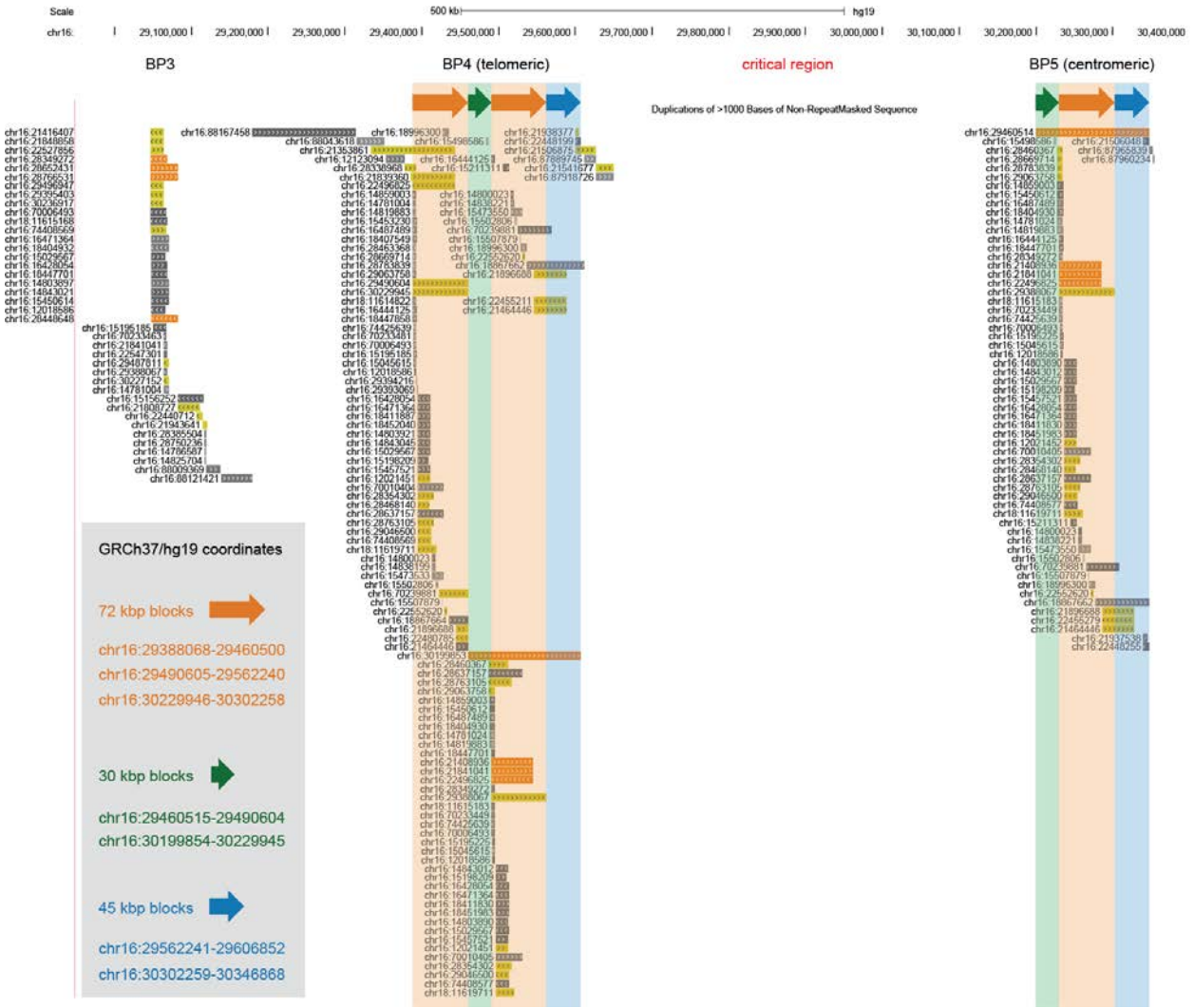


Figure S1. Segmental duplication architecture flanking the ~550 kbp 16p11.2 autism critical region. Schematic depicts the human reference sequence (GRCh37) at this locus and the complex blocks of segmental duplications at BP4 and BP5 [6] (Segmental Dups UCSC Genome Browser track). Each colored box indicates sequence duplicated between the region it spans and another genomic locus. Box colors correspond to sequence identity between duplication pairs (orange = 99% or above, yellow = 98%–99%, gray = 90%–98%), and direction of markings indicates orientation of duplicated sequences (right-pointing, directly oriented; left-pointing, inversely oriented). Thick arrows correspond to duplication blocks of particular interest to this study defined and discussed in the text and in **Fig. S2**.

BAC clones were sequenced using capillary sequencing or Pacific Biosciences (PacBio) single-molecule, real-time (SMRT) technology [7]. SMRT-sequenced clones were assembled using HGAP and error-corrected using Quiver to generate one complete single contig per clone as previously described [7]. Sequences from overlapping clones were assembled into larger haplotype contigs using Sequencher (Gene Codes Corporation). To ensure proper assembly, we assembled clones into the same contig only if they exhibited >99.9% sequence identity over their shared region of overlap. Due to the complexity of the 16p11.2 locus in chimpanzee and our discovery of a large inversion polymorphism, we performed additional rounds of genomic library colony hybridization, Nextera clone sequencing [1, 4, 5], and SMRT BAC sequencing and assembly [1, 7] to fill gaps remaining after the initial set of chimpanzee clones was sequenced and assembled into contigs. Ten chimpanzee clones could not be fully resolved owing to large (>30 kbp), high-identity tandem duplications within them that could not be spanned by SMRT sequence read lengths (**Fig. 4.1a** and **Table S1**). We encountered this same problem sequencing human haplotypes but were able to overcome it by sequencing shorter clones (~40 kbp) from a fosmid library (ABC12) corresponding to the same human individual (NA12878).

In total, we incorporated 106 clones into our final 16p11.2 contig assemblies, including 70 BACs sequenced using SMRT technology, 21 BACs previously sequenced using capillary technology, and 15 fosmids sequenced using SMRT technology (**Fig. 4.1** and **Table S1**). A small gap in unique sequence in one contig (corresponding to the effectively haploid hydatidiform mole, CHM1 [3]) was filled using consensus sequence from SMRT whole-genome sequencing (WGS) of the same genome [8]. To enable detailed analysis of the evolutionary history of 16p11.2, we also compiled or generated sequence data over loci paralogous to duplicated sequences within 16p11.2. Specifically, we assembled publicly available BAC sequences into contigs covering the 16p12.1 locus in human and chimpanzee (**Table S1**) and sequenced or collected sequence data from several BACs corresponding to ancestral paralogs of 16p11.2 duplicated sequences (**Table S1**). All contig sequences and assembled clone inserts are publicly available in GenBank under the accession codes provided at the end of the manuscript.

2. Structural variation

2.1 Segmental duplication analyses

Two approaches were used to annotate and characterize segmental duplications within each assembled haplotype. First, we identified all regions within our sequences homologous to known human segmental duplications by analyzing each sequence using DupMasker [9] (default settings). Second, we identified segmental duplications by applying a whole-genome assembly comparison (WGAC) pipeline [10] to a minimal genome assembly consisting of only our contig sequences, with each treated as a separate "chromosome". These analyses revealed that the 16p11.2 locus in human and chimpanzee, but not orangutan, has acquired more than 1 Mbp of duplicated sequences originating from over a dozen ancestral genomic loci (**Fig. 4.1**). The most complex duplication architecture was observed in chimpanzee.

Three segmental duplications are of particular interest: i) a 30 kbp segment containing the genes *BOLA2*, *SLX1*, and *SULT1A3* (as well as a potential fusion gene *SLX1-SULT1A3*); ii) a 72 kbp segment including

NPIP and the 3' end of *SMG1P*; and iii) a 45 kbp segment harboring the 5' end of *SMG1P* (**Fig. S2**). These segmental duplications constitute the largest block of duplicated sequences found at multiple locations within the 16p11.2 BP1-BP5 region [6] in humans. They flank the autism critical region in direct orientation and, thus, are strong candidates for mediating nonallelic homologous recombination (NAHR) underlying recurrent microdeletions and microduplications [11]. In contrast, the largest segments of duplication within the chimpanzee haplotypes occur in inverted orientation and would promote recurrent inversions, consistent with the observed inversion polymorphism between the two chimpanzee haplotypes. Interestingly, both of these duplications harbor species-specific duplicated sequences—the 30 kbp block including *BOLA2* in humans and an ~160 kbp block originating from chromosome 16p12.1 in chimpanzee.

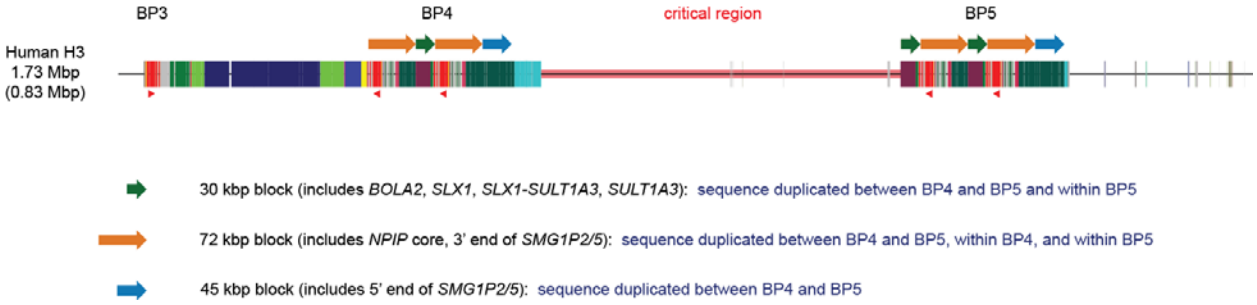


Figure S2. Simplified schematic of segmental duplication architecture in human chromosome 16p11.2 at BP4 and BP5. Duplication blocks of particular interest to this study are shown as thick arrows. Criteria (blue text) defining each block were determined based on duplication patterns within the most derived human haplotype (H3).

Comparative analysis of human haplotypes indicates structural variation affects both sides of the critical region, results in large (95-102 kbp) blocks of highly identical, directly oriented sequences adjacent to and flanking the critical region, and always involves the same 102 kbp unit, including *BOLA2*, *SLX1*, and *SULT1A3* (**Fig. 4.1b**). These data, together with our *BOLA2* copy number estimates from WGS data (**Fig. 4.3a**), suggest that humans with extreme *BOLA2* copy numbers differ by >500 kbp as a result of copy number variation. The duplication architecture of the region predisposes it to NAHR both within BP4 and BP5, resulting in tandem expansions and contractions of the variable unit on each side of the critical region, and between BP4 and BP5, resulting in disease-associated microduplications and microdeletions (**Fig. S3**).

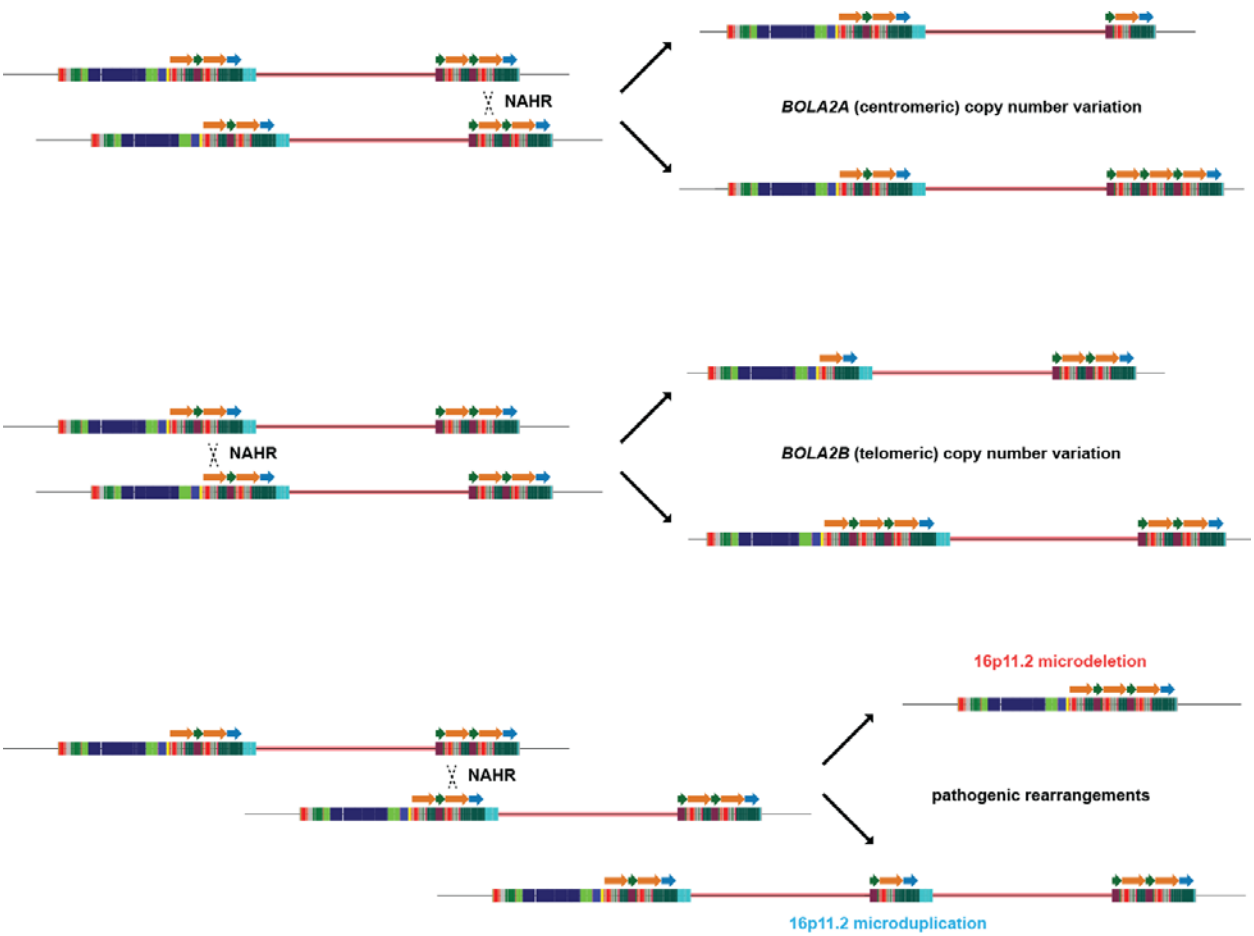


Figure S3. NAHR model underlying normal and disease-associated copy number variation at 16p11.2.

2.2 Inversion analysis

We visualized structural differences between sequenced haplotypes using Miropeats [12] and refined breakpoints by sequence alignment using Clustal 2.1 [13]. A series of three-color fluorescence *in situ* hybridization (FISH) experiments (**Table S2** and section 9.1) were performed to validate the order and orientation of inversions (**Figs. S4-S5**). The results confirm accurate assembly of our contigs and imply extensive reorganization of the 16p11.2 region over the past ~15 million years of great ape evolution.

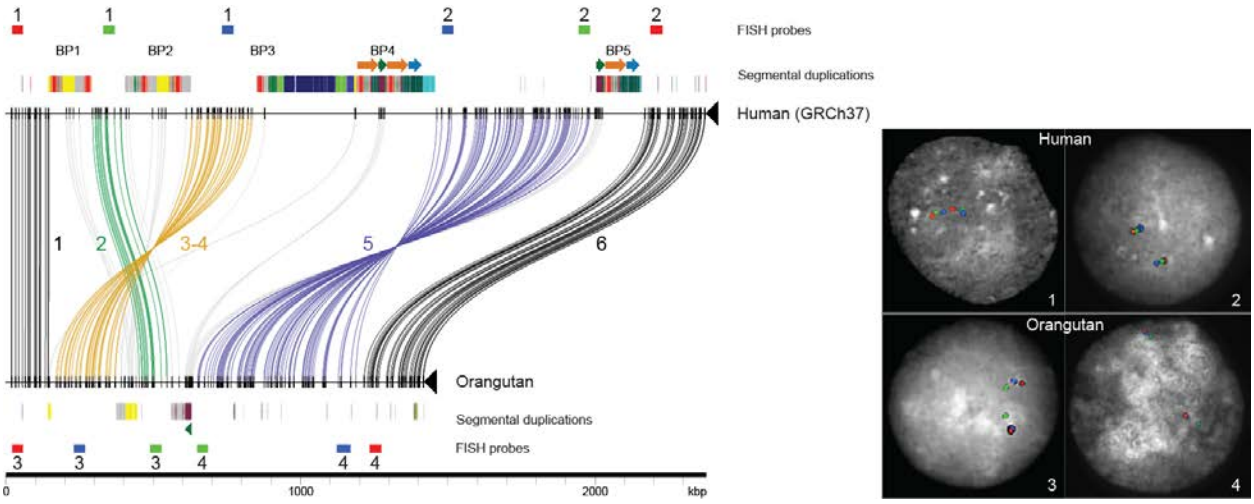


Figure S4. Comparison of 16p11.2 structure between human and orangutan. Sequences (thin horizontal lines) from human (GRCh37 chr16:28195661-30573128) and orangutan (contig sequence) at 16p11.2 are compared using Miropeats [12] (s = 1,000) and annotated with locations of human segmental duplications and FISH probes used to validate the organization of the region. Lines connecting the sequences show regions of homology, and line colors highlight differences in the order and orientation of distinct gene-rich regions of unique sequence across the locus (numbered 1-6). Numbers below FISH probes correspond to numbers within the images on the right, specifying which probes were used in each experiment. Experiment 1 used the same probes as experiment 3, and experiment 2 used the same probes as experiment 4. Three-color interphase FISH on human and orangutan chromosomes confirms the accuracy of our assembled orangutan contig.

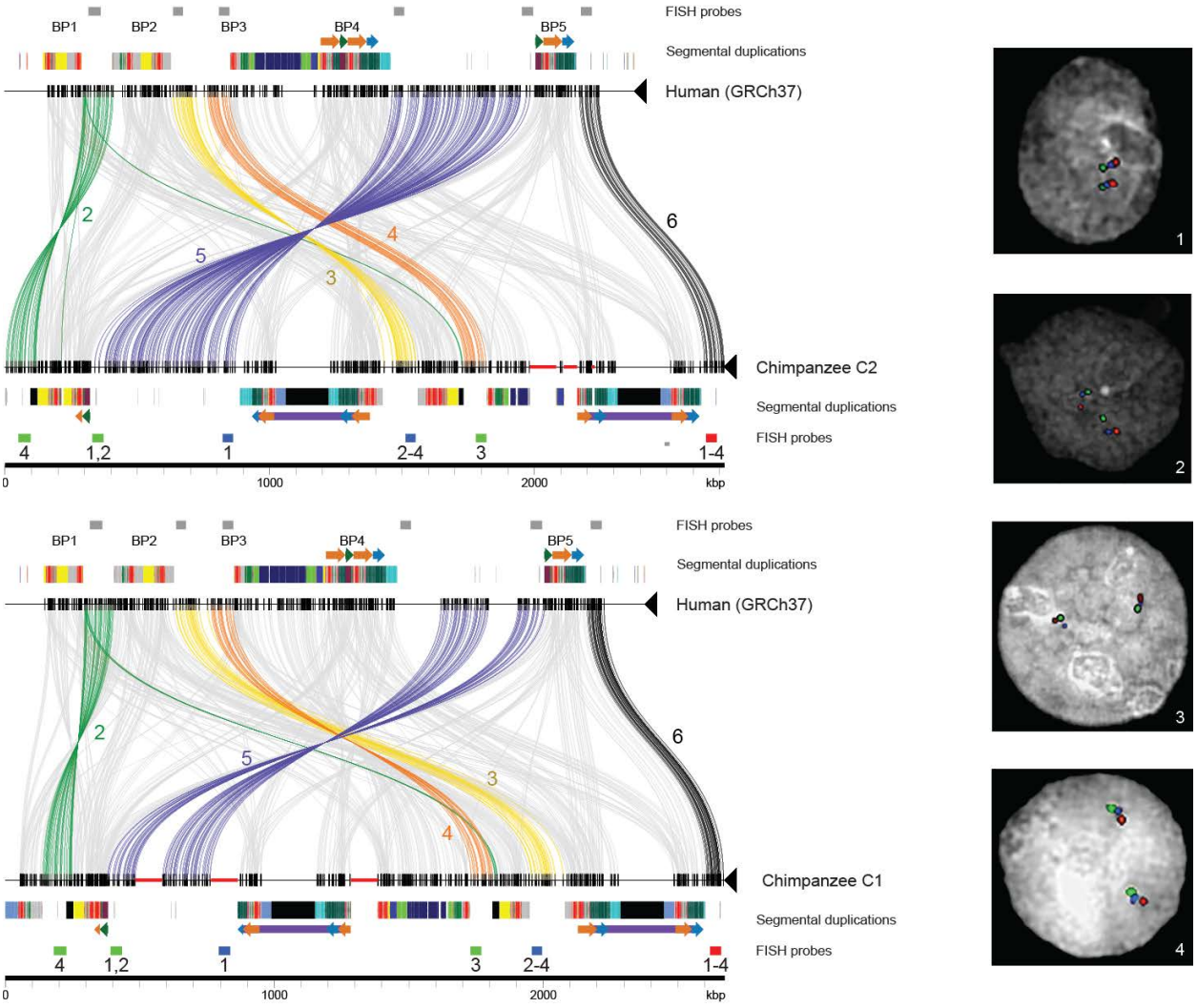


Figure S5. Comparison of 16p11.2 structure between human and chimpanzee. Sequences (thin horizontal lines) from human (GRCh37 chr16:28195661-30573128) and two chimpanzee structural haplotypes at 16p11.2 are compared using Miropeats [12] (s = 1,500) and annotated with locations of human segmental duplications and FISH probes used to validate the organization of the region. Thick red horizontal lines indicate gaps in the chimpanzee contigs, and black boxes correspond to chimpanzee-specific segmental duplications (i.e., sequences not duplicated in humans). Lines connecting the sequences show regions of homology, and line colors highlight differences in the order and orientation of distinct gene-rich regions of unique sequence across the locus (numbered 2-6). Numbers below FISH probes correspond to numbers within the images on the right, specifying which probes were used in each experiment. Gray rectangles show mapping locations of FISH probes in human. Three-color interphase FISH on chimpanzee chromosomes confirms the accuracy of our assembled contigs.

3. Evolutionary reconstruction

We put forward a model for the evolution of the chromosome 16p11.2 region in great apes (**Fig. 4.2a**, **Fig. S6**, and **Table S3**) and detail the evidence supporting each hypothesized step below.

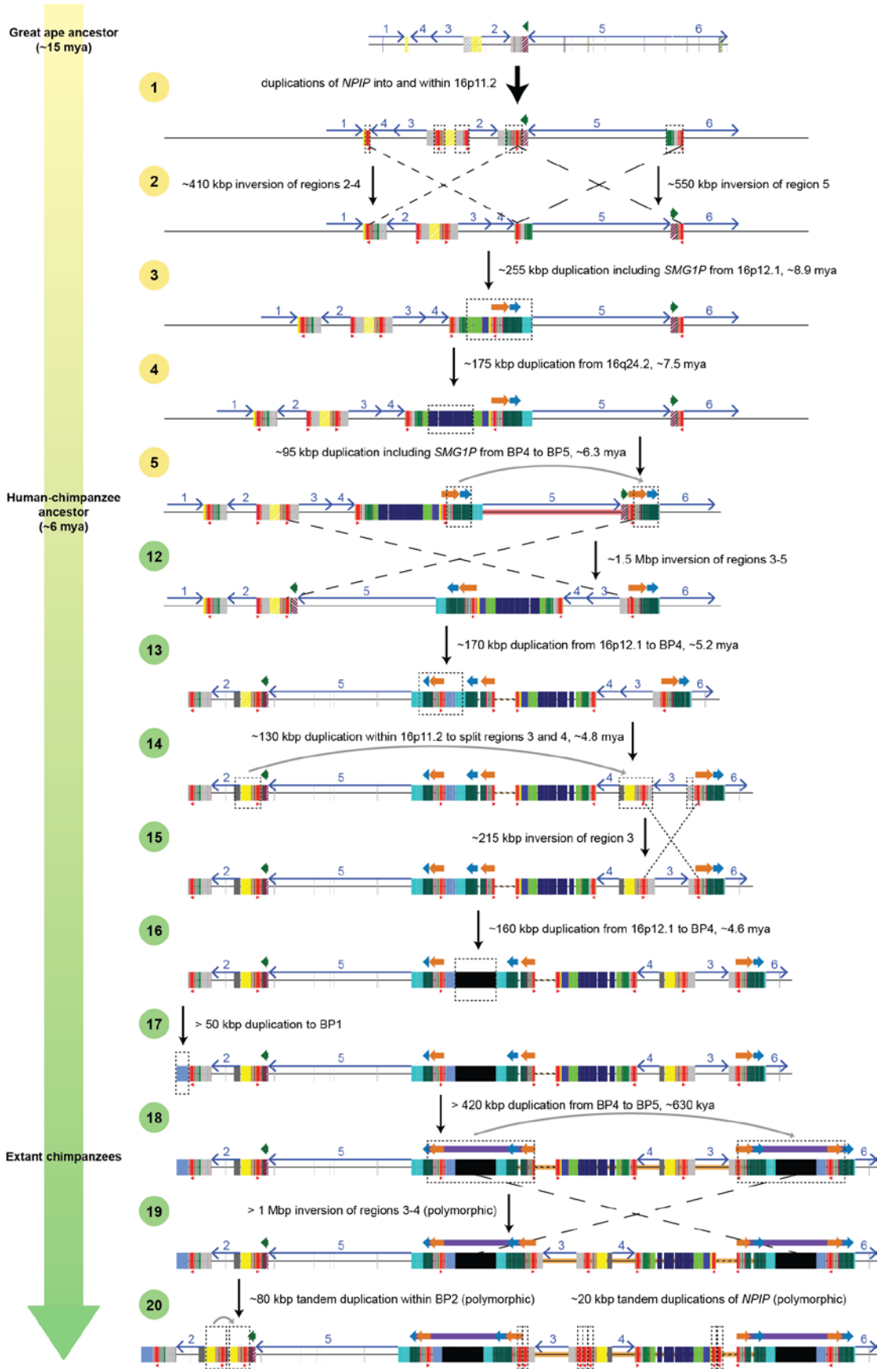

Great ape ancestor
(~15 mya)
1
duplications of NPIP into and within 16p11.2
2
~410 kbp inversion of regions 2-4
~550 kbp inversion of region 5
3
~255 kbp duplication including SMG1P from 16p12.1, ~8.9 mya
4
~175 kbp duplication from 16q24.2, ~7.5 mya
5
~95 kbp duplication including SMG1P from BP4 to BP5, ~6.3 mya
Human-chimpanzee
ancestor
(~6 mya)
12
~1.5 Mbp inversion of regions 3-5
13
~170 kbp duplication from 16p12.1 to BP4, ~5.2 mya
14
~130 kbp duplication within 16p11.2 to split regions 3 and 4, ~4.8 mya
15
~215 kbp inversion of region 3
16
~160 kbp duplication from 16p12.1 to BP4, ~4.6 mya
17
> 50 kbp duplication to BP1
18
> 420 kbp duplication from BP4 to BP5, ~630 kya
Extant chimpanzees
19
> 1 Mbp inversion of regions 3-4 (polymorphic)
20
~80 kbp tandem duplication within BP2 (polymorphic)
~20 kbp tandem duplications of NPIP (polymorphic)

Figure S6. Model for the evolution of chromosome 16p11.2 from the great ape ancestor to modern chimpanzee. This schematic depicts structural changes within the 16p11.2 BP1-BP5 region[5] over time leading to the present-day chimpanzee architecture (see **Fig. 4.2a** for a corresponding schematic for human chromosome 16p11.2 evolution and **Table S3** for a listing of all structural rearrangement events). Evidence supporting each evolutionary event is detailed below. Steps 1-5 represent changes that occurred prior to human-chimpanzee divergence. Steps 6-11 occurred specifically along the human lineage, while Steps 12-20 correspond to chimpanzee-specific changes.

Evolution of chromosome 16p11.2 from the great ape ancestor to the human-chimpanzee ancestor (Steps 1-5)

Ancestral ape genome organization

Comparison of our assembled orangutan sequence contig with orthologous sequence from the mouse genome assembly (GRCm38/mm10 chr7:126110001-127410000) reveals identical gene order and orientation between these two species (**Fig. S7**). Similar to orangutan, mouse sequence over this region is largely devoid of segmental duplications. Both observations indicate the orangutan organization likely represents the great ape ancestral state and the dramatic changes that restructured this region in human and chimpanzee occurred after divergence with orangutan. To determine the ancestral *BOLA2* locus, we examined the order and orientation of genes closest to *BOLA2* copies at BP4 and BP5 but not duplicated between the two loci in human. Considering these genes in our nonhuman primate contigs and in mouse, we observed conserved order gene synteny only between *BOLA2* and *CORO1A* (**Fig. S7**). Because *CORO1A* is present at BP5 but not at BP4 in humans, we conclude that BP5 represents the ancestral locus.

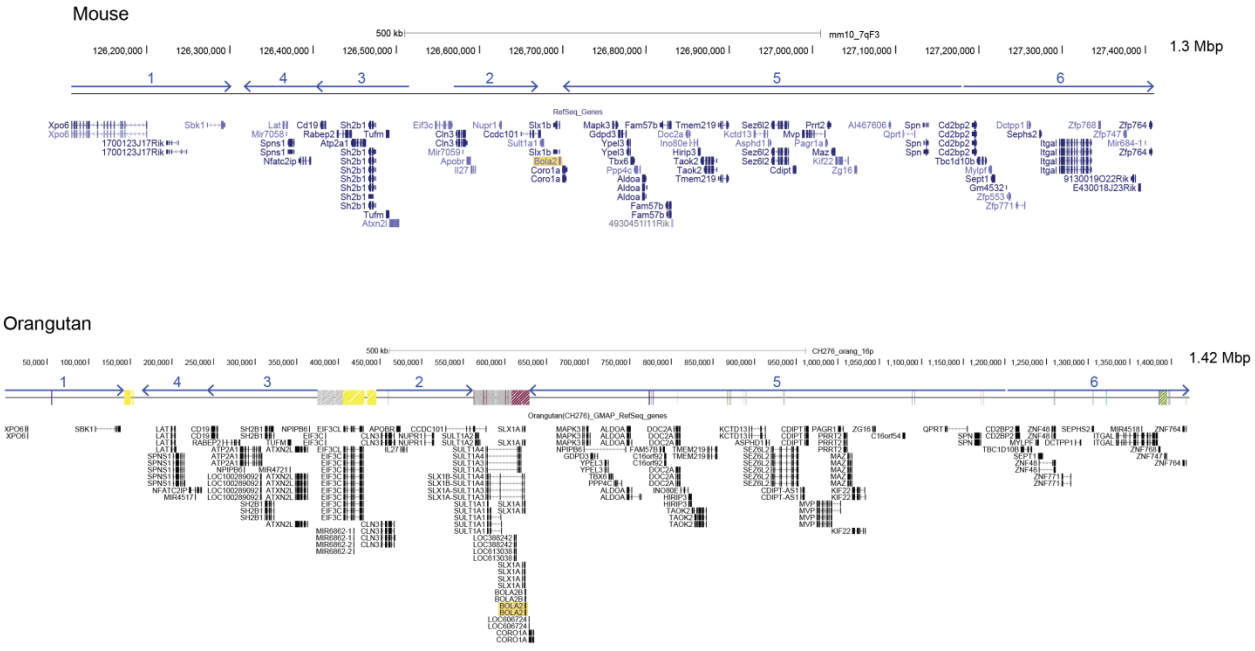


Figure S7. Conserved order synteny between orangutan and mouse. Comparative genomic analysis suggests the orangutan 16p11.2 configuration represents the ancestral great ape state. *BOLA2* (highlighted in yellow in the gene tracks below each sequence) maps adjacent to *CORO1A* in nonhuman primates and in mouse, consistent with the ancestral human *BOLA2* paralog mapping adjacent to *CORO1A* at BP5 and the duplicate *BOLA2* paralog mapping to BP4. RefSeq mRNA sequences from GRCh37 were mapped to the orangutan contig using GMAP [14], and results were used to annotate gene locations in orangutan.

3.1 Step 1: Expansion of the LCR16a segmental duplication

A striking feature of human and chimpanzee assembled sequence contigs is the abundance of ~20 kbp chromosome 16 low-copy repeat (LCR16a) sequences containing the *NPIP* family [15] (red triangles in **Fig. 4.1**, **Fig. 4.2a**, and **Fig. S6**). These *NPIP* core duplicons [16] are absent from the 16p11.2 locus in orangutan and mouse. Previous phylogenetic analyses [15, 17] revealed that LCR16a expanded in the human-chimpanzee-gorilla common ancestor and showed that about two-thirds of all copies are orthologous among African great apes. Interestingly, all of the inversion breakpoints between our contig sequences map within regions of segmental duplications including inversely oriented *NPIP* core duplicons—consistent with their involvement in mediating most of the evolutionary inversions. Strikingly, the LCR16a segmental duplication is associated with 14 evolutionary events affecting 16p11.2 and/or genes therein (**Table S3**), not including its likely role in driving inversions. These events include duplications from another locus on chromosome 16 into 16p11.2 (Steps 1, 3, and 13), duplications within 16p11.2 (Steps 5, 7, 9, 10, 11, 14, 18, and 20), interlocus gene conversion (Step 6), and duplication from 16p11.2 to chromosome 17 (not included in our model schematics).

3.2 Step 2: Evolutionary inversions before human-chimpanzee divergence

For each unique gene-rich region (numbered 2-5 in **Fig. 4.1a**) bracketed by segmental duplications, we determined the most likely number of inversions using the following logic. If the orientation of the region in a particular species is the same as the orientation of the orthologous region in the orangutan proxy for the great ape ancestor, the region either did not invert or must have inverted an even number of times along that species' lineage after divergence with orangutan. In contrast, if the orientation in the species of interest is opposite that in orangutan, the region must have inverted an odd number of times over the same evolutionary period. For each region of unique sequence, for both human and chimpanzee, we applied maximum parsimony, selecting the path requiring the fewest rearrangements to reconcile modern human and chimpanzee organizations with the ancestral great ape state. Our full evolutionary model (**Fig. 4.2a**, **Fig. S6**, and **Table S3**) suggests a total of six evolutionary inversions, including two along the branch leading from the great ape ancestor to the human-chimpanzee ancestor, one specific to the human lineage, and three specific to the chimpanzee lineage.

The two distinct inversions occurring along the lineage between the great ape ancestor and the human-chimpanzee ancestor affect unique regions 2-4 (17 genes) and unique region 5 (30 genes including *BOLA2*, *SLX1*, and *SULT1A3*), respectively. Neither precise timing estimates for these events nor their order relative to each other or to most other events along the human-chimpanzee ancestral lineage can be inferred. Because the breakpoints of these inversion events map within regions of segmental duplications including inversely oriented *NPIP* core duplicons, it appears likely that NAHR between inverted *NPIP* copies mediated these inversions. Thus, these inversions are displayed together in this step, predating the human-chimpanzee common ancestor, but occurring after the dispersal of *NPIP* across the 16p11.2 locus.

3.3 Step 3: Duplicative transposition between chromosome 16p12.1 and 16p11.2

A complex higher-order duplication block (~255 kbp) located between BP3 and BP4 in humans and at the orthologous chimpanzee position maps to two locations in these genomes: chromosome 16p11.2 and chromosome 16p12.1 (**Fig. S8a**). Although partial fragments of this block are found elsewhere on chromosome 16, the identical structure (with respect to order and orientation of smaller segmental duplications) at these two positions implies that the components (individual duplicons) first evolved at one of these loci, followed by the larger cassette duplicating to the other locus. We estimated the timing of this duplication using a molecular clock approach, generating a phylogenetic tree (**Fig. S8b**) incorporating sequence over a region of the cassette that is unique sequence in orangutan and, therefore, orthologous to both chimpanzee and human.

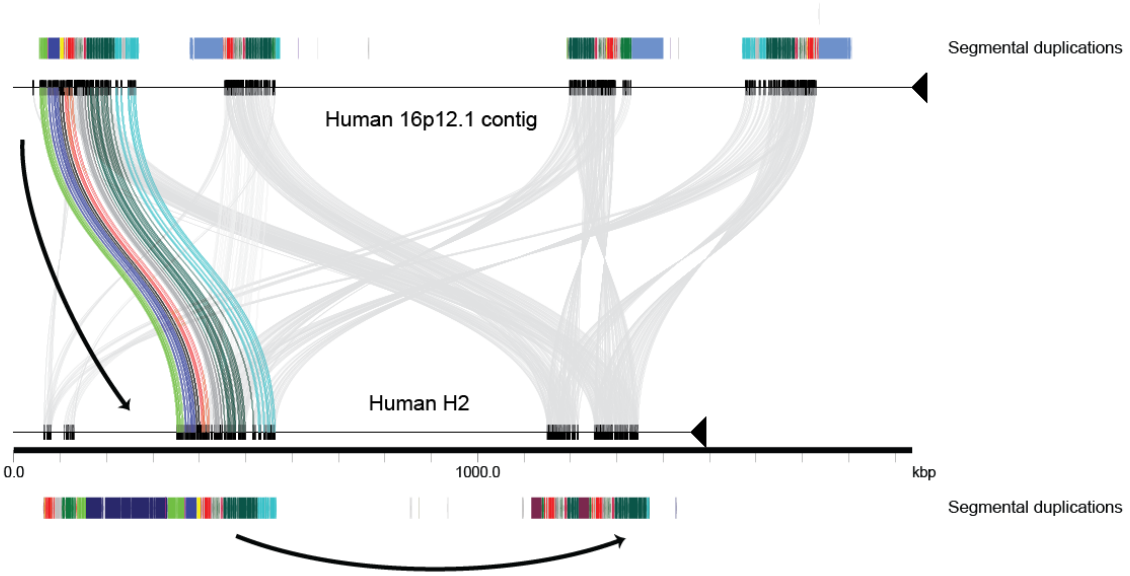
To estimate the evolutionary age of duplications into and within the 16p11.2 locus, we calibrated local molecular clocks based on sequence divergence of paralogs and orthologs and assumed divergence times

of 6 mya between human and chimpanzee [18] and 15 mya between human or chimpanzee and orangutan [5]. Specifically, we generated multiple sequence alignments including relevant sequences from our contigs (**Table S1**), reference genome assemblies, and sequenced BACs (**Table S1**) using Clustal 2.1 [13] and fixed alignment errors manually (including removing regions of poor alignment quality) using Jalview [19]. We built a series of neighbor-joining phylogenetic trees (**Figs. S8-S9**, **Fig. S11**, **Fig. S13**, and **Figs. S15-S17**) using MEGA5 [20] with the complete deletion option, calculating genetic distances using the Kimura 2-parameter model with standard error estimates based on an interior branch test of phylogeny with 500 bootstrap replicates. For each tree, we performed several Tajima's relative rate tests to assess the validity of the molecular clock assumption.

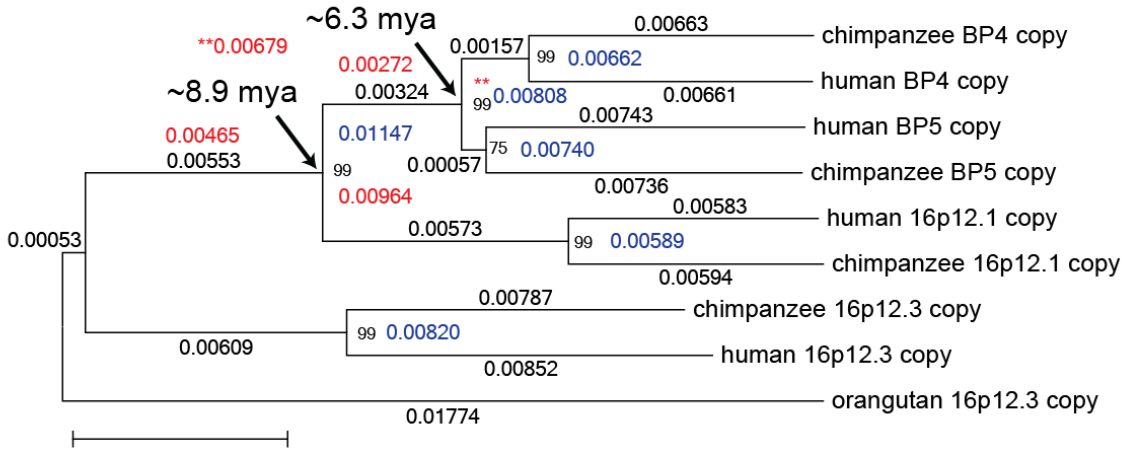
For phylogenetic trees where the molecular clock was supported, we estimated divergence times corresponding to duplication events of interest using the equation T = K/2R, where T is time (in millions of years), K is divergence (substitutions per site), and R is the substitution rate (substitutions per site per million years). It follows that the divergence times corresponding to duplication events of interest can be estimated without explicitly calculating the substitution rate if assumptions are made regarding divergence times between species. Specifically, the fraction of sequence divergence after the duplication event relative to the total divergence between duplicated sequences of interest and a single-copy orthologous outgroup sequence is equal to the fraction of time that elapsed after the duplication event relative to the total divergence time between the species having the duplicated sequences and the outgroup species. We present details of duplication timing calculations for each phylogenetic tree using this approach in corresponding figures. For trees where the molecular clock was not supported, we scaled sequence divergences to match the substitution rate of branches corresponding to the ancestral locus. These divergence corrections are noted in corresponding figures and associated text.

For this initial duplication from 16p12.1 to 16p11.2 including the 5' end of the gene *SMG1* (i.e., including *SMG1P*), we scaled duplicate branch lengths to the branch lengths for the ancestral *SMG1* locus (16p12.3), as the tree did not pass Tajima's relative rate test. We estimate that duplication of the cassette occurred ~8.9 mya. We cannot confidently infer the directionality of this duplication, so the 16p11.2 duplicons may in fact be older if 16p11.2 is the ancestral locus for the cassette. However, the fact that later duplication events have clearly transferred large pieces of sequence to 16p11.2 within the BP3-BP4 region or its chimpanzee counterpart (Steps 4, 13, and 16—twice from 16p12.1) suggests the direction of this duplication was most likely from 16p12.1 to 16p11.2.

a



b



0.005 substitutions per site

15253 alignment positions

includes end of 45 kbp block
(part not involved in interlocus gene conversion between BP4 and BP5)

chimpanzee BP4 copy = chimpanzee ancestral:1193427-1221394
human BP4 copy = human H2:493179-521219
human BP5 copy = human H1:1357796-1386137
chimpanzee BP5 copy = chimpanzee ancestral:2569455-2595492
human 16p12.1 copy = human 16p12.1 contig:196366-212974
chimpanzee 16p12.1 copy = chimpanzee 16p12.1 contig 1:248485-265045
chimpanzee 16p12.3 copy = CH251-511B8:57762-86119
human 16p12.3 copy = GRCh37/hg19 chr16:18909541-18937920
orangutan 16p12.3 copy = CH253-5C15:100835-129718

branch lengths
corrected branch lengths
average branch length for all sequences in clade (corrected average branch lengths)

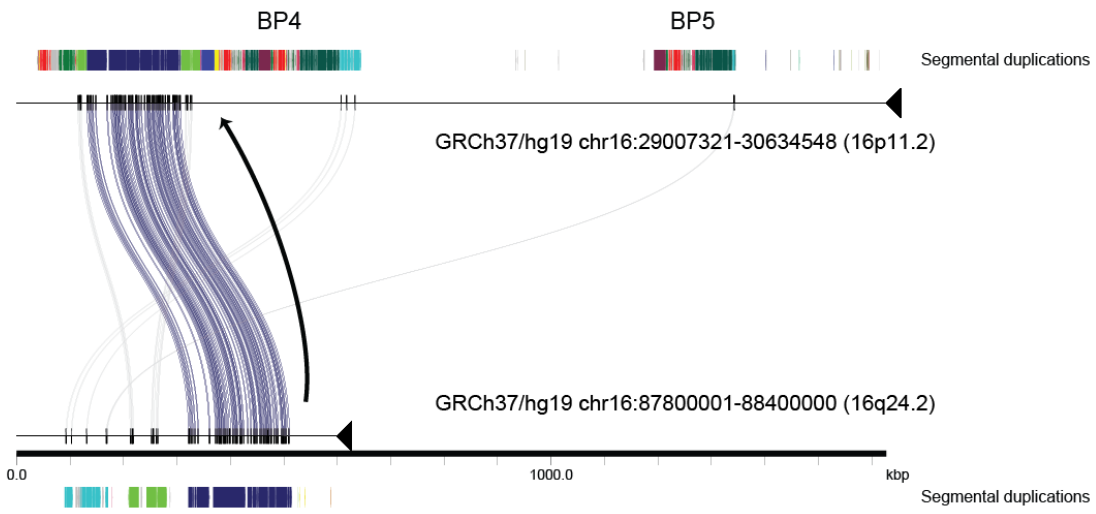
T1 = ((0.00964 subs/site)/(0.00964+0.00465+0.00053+0.01774 subs/site))*(30 million years)
T1 = ~8.9 million years
T2 = ((0.00679 subs/site)/(0.00679+0.00272+0.00465+0.00053+0.01774 subs/site))*(30 million years)
T2 = ~6.3 million years

Figure S8. Characterization of the duplication including *SMG1P* from 16p12.1 to 16p11.2 and the duplication including *SMG1P* from BP4 to BP5. a) Sequences (thin horizontal lines) from human 16p12.1 (top, human 16p12.1 contig) and human 16p11.2 (bottom, human H2 contig) are compared using Miropeats [12] (s = 1,000). Lines connecting the sequences show regions of homology, and colored lines highlight most sequence included in the duplication event from 16p12.1 to 16p11.2. Examination of chimpanzee 16p12.1 sequence (chimpanzee 16p12.1 contig 1, not shown) suggests this duplication included the full extent of the lime-green-colored duplication block in addition to the region connected with colored lines here. A later duplication copied *SMG1P* from BP4 to BP5. b) An unrooted neighbor-joining phylogenetic tree based on a 15,253 bp multiple sequence alignment including sequences from 16p11.2, 16p12.1, and 16p12.3 (the ancestral *SMG1* locus). Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades, while red numbers show either raw or clade-averaged branch lengths corrected to account for the difference in substitution rate between ancestral and duplicate *SMG1* loci. Branch lengths were used to estimate times corresponding to *SMG1P* duplications into and within 16p11.2 as shown (see equations for T1 and T2, respectively). Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified.

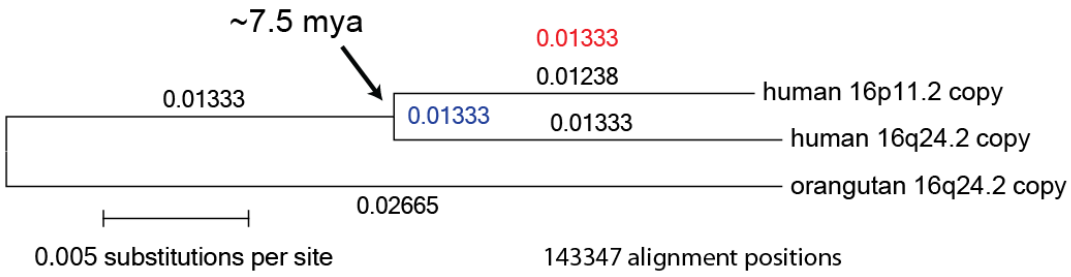
3.4 Step 4: Duplicative transposition from chromosome 16q24.2 to 16p11.2

The second largest component of the complex block of duplications between BP3 and BP4 in humans is an ~175 kbp segment originating from chromosome 16q24.2 (**Fig. S9a**). Tajima's relative rate test indicated that the 16p11.2 copy did not evolve at the same relative rate as the 16q24.2 copy. Adjusting the 16p11.2 branch length accordingly, we estimate this duplication event occurred ~7.5 mya, after the duplication of *SMG1P* from 16p12.1 to 16p11.2 but before the duplication of *SMG1P* within 16p11.2 (**Fig. S9b**). Indeed, sequence analysis indicates that this segmental duplication disrupts the contiguity of the ~255 kbp *SMG1P* duplication block (Step 3; see lime-green colored duplicon in **Fig. 4.2a**), unequivocally making it a secondary event after the initial duplication of *SMG1P* into 16p11.2. Consistent with our timing estimate suggesting this duplication from 16q24.2 occurred near the time of human-African great ape speciation, FISH experiments (**Table S2**) show human and chimpanzee have this ~175 kbp segment duplicated between 16q24.2 and 16p11.2, while orangutan and gorilla lack the duplicate copy at 16p11.2 (**Fig. S9c**).

a



b



branch lengths
corrected branch lengths
average branch length for all sequences in clade

human 16p11.2 copy = human H4:149636-325017
human 16q24.2 copy = GRCh37/hg19 chr16:88121422-88310654
orangutan 16q24.2 copy = CH276-223L14:26032-186639

T = ((0.01333 subs/site)/(0.01333+0.1333+0.02665 subs/site))*(30 million years) = ~7.5 million years

c

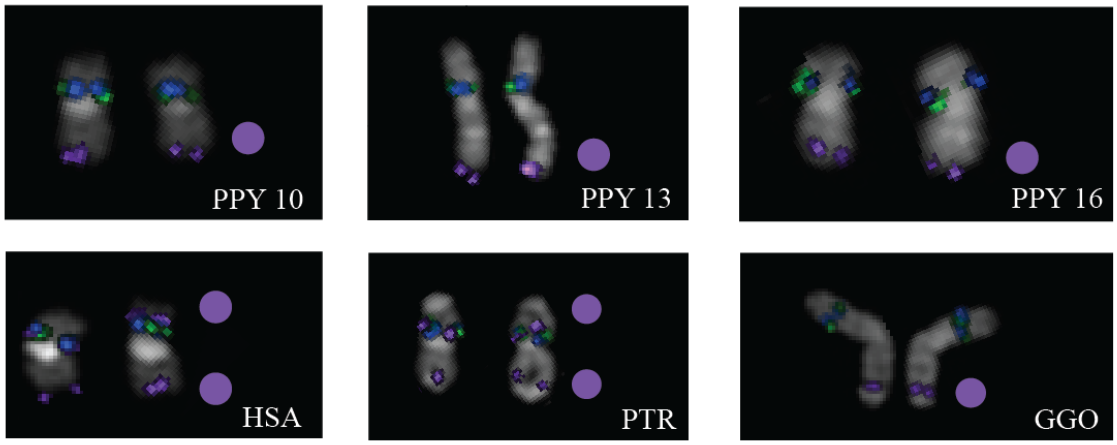


Figure S9. Characterization of the ~175 kbp duplication from 16q24.2 to 16p11.2. a) Sequences (thin horizontal lines) from human 16p11.2 (top, GRCh37) and human 16q24.2 (bottom, GRCh37/hg19) are compared using Miropeats [12] (s = 1,000). Lines connecting the sequences show regions of homology, and colored lines highlight sequence included in the duplication event from 16q24.2 to 16p11.2. b) An unrooted neighbor-joining phylogenetic tree based on a 143,347 bp multiple sequence alignment including sequences from chromosome16p11.2 and 16q24.2 (the ancestral locus). Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), while the red number shows the human 16p11.2 branch length corrected to account for the difference in substitution rate between ancestral and duplicate loci. The blue number corresponds to the node at which human sequences branch and indicates the average branch length for all human sequences after correction for different substitution rates. Branch lengths were used to estimate the time corresponding to the duplication from 16q24.2 to 16p11.2 as shown. Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified. c) FISH experiments using a probe (purple) mapping within this ~175 kbp segment together with probes (blue and green) mapping within regions of unique sequence at 16p11.2 confirm this segment is duplicated between 16q24.2 and 16p11.2 in human (HSA) and chimpanzee (PTR) metaphases. This is consistent with its presence in our 16p11.2 haplotype contigs for these two species. Conversely, this segment is present only at 16q24.2 in gorilla (GGO) and orangutan (PPY) metaphases. We conclude that the duplication most likely occurred in the common ancestor of human and chimpanzee after divergence from gorilla.

3.5 Step 5: Duplicative transposition from BP4 to BP5 within chromosome 16p11.2

The duplication of an ~95 kbp segment that included *SMG1P* from BP4 to BP5 is the least confident step in our evolutionary reconstruction. Approximately 67 kbp of this segment at BP5 was subsequently subjected to interlocus gene conversion along the human lineage (see Step 6 for evidence), obscuring the phylogenetic signal of its duplication history. Phylogenetic analysis of sequence within the ~95 kbp segment not affected by conversion suggests this duplication occurred ~6.3 mya, around the time of human-chimpanzee speciation (**Fig. S8b**). Although there are at least seven *SMG1P* copies on human chromosome 16, BP4 and BP5 *SMG1P* copies share sequence with the ancestral *SMG1* locus that is not shared with other *SMG1P* duplicates. Furthermore, 16p11.2 *SMG1P* copies are more highly identical to one another than to the ancestral locus. These observations are consistent with the phylogenetic inference that the BP5 copy originated as a result of duplication from BP4.

Human-specific evolution of chromosome 16p11.2 (Steps 6-11)

3.6 Step 6: Complex interlocus gene conversion event between chromosome 16p12.1 and 16p11.2

The duplication architecture at BP5 in humans (**Fig. S1**) exhibits an unusual property. Except for the 30 kbp segment including *BOLA2*, most sequence duplicated between BP5 and BP4 is also duplicated between BP5 and chromosome 16p12.1. Strikingly, this duplication between BP5 and 16p12.1 does not exhibit uniform sequence identity across the alignment, instead showing >99% identity over most of the 72 kbp block and lower identity (<99%) over the remainder of the 72 kbp block and over the 45 kbp block. This sequence identity pattern suggests the duplications at BP5 are likely a product of more than one evolutionary event. A simple scenario involving a single duplication of the 72 kbp and 45 kbp blocks in concert from BP4 to BP5 fails to explain the non-uniform sequence identity with 16p12.1. Visualizing the patterns of sequence identity between the paralogous loci suggests a complex human-specific interlocus gene conversion involving BP5, BP4, and sequence from 16p12.1 (**Fig. S10** and **Fig. S11a**). Combining our sequence identity analysis with phylogenetic timing (**Fig. S11b**), we propose that BP5 acted as a conversion acceptor for 54 kbp of sequence from 16p12.1 and 36 kbp of sequence from BP4 ~1.8 mya.

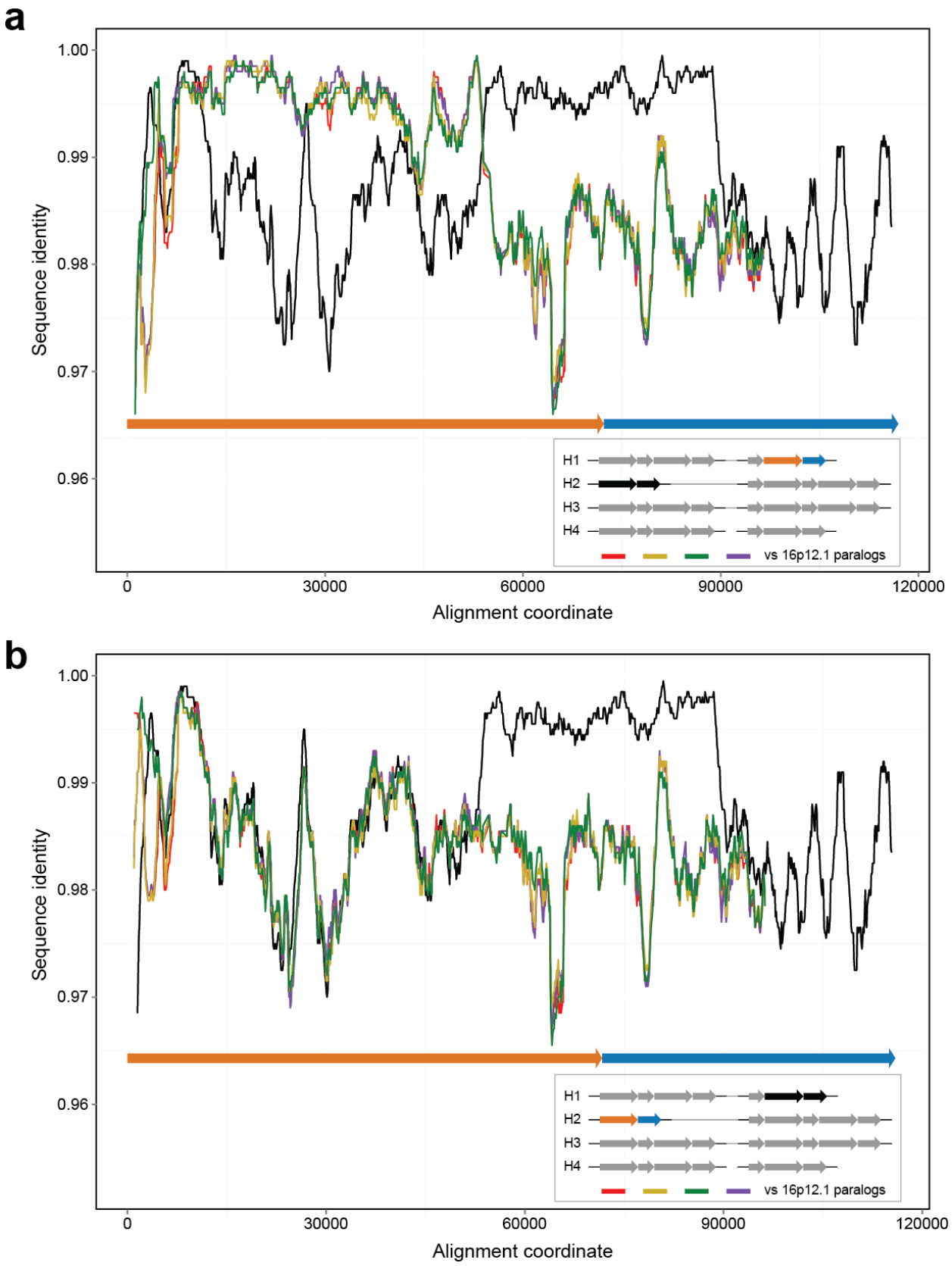

a
1.00
0.99
0.98
0.97
0.96
Sequence identity
H1
H2
H3
H4
vs 16p12.1 paralogs
0
30000
60000
90000
120000
Alignment coordinate
b
1.00
0.99
0.98
0.97
0.96
Sequence identity
H1
H2
H3
H4
vs 16p12.1 paralogs
0
30000
60000
90000
120000
Alignment coordinate

Figure S10. Sequence identity analysis reveals a complex interlocus gene conversion event involving sequence from BP4, BP5, and 16p12.1. a) BP5 sequence over the 72 kbp and 45 kbp blocks was aligned to BP4 sequence as well as four paralogous sequences from 16p12.1. For each pair of aligned sequences including BP5, sequence identity was calculated and plotted over 2 kbp windows, sliding by 100 bp. Lines show identity between BP5 and BP4 (black) or between BP5 and 16p12.1 sequences (other colors) across the extent of sequence shared between these regions. BP5 initially exhibits high identity with 16p12.1 sequences, sharply transitions to having the same level of high identity with BP4, and ends without having that level of high identity with BP4 or 16p12.1 sequences. This spatial pattern is most consistent with interlocus gene conversion, with BP5 acting as a conversion acceptor for 16p12.1 sequence and BP4 sequence likely in a single complex conversion event. b) These same analyses were performed for BP4 sequence. BP4 exhibits a high level of identity only with BP5 (black line) over a 36 kbp region where the conversion event affecting BP5 involved BP4 serving as the donor. BP4 does not exhibit the same high level of identity with any sequence from 16p12.1 (colored lines).

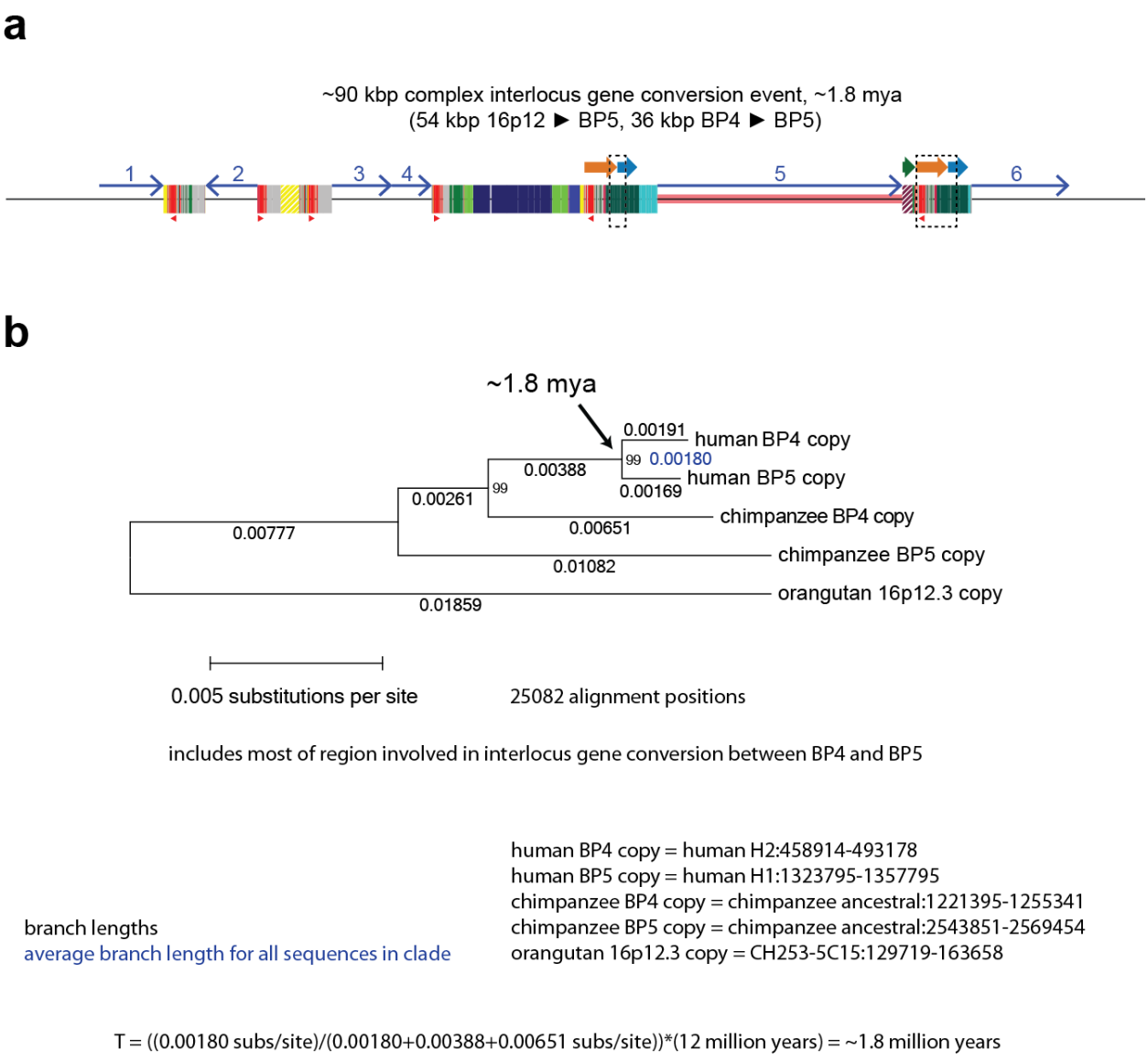


Figure S11. Phylogenetic characterization of the complex interlocus gene conversion event. a) Schematic shows 16p11.2 regions involved in the conversion event. Sequences from 16p12.1 (54 kbp, not shown) and BP4 (36 kbp, first dashed box) served as conversion donors, affecting ~90 kbp of sequence at BP5 (second dashed box). b) An unrooted neighbor-joining phylogenetic tree based on a 25,082 bp multiple sequence alignment including sequences

from BP4, BP5, and 16p12.3 (the ancestral *SMG1* locus) over the BP4-BP5 conversion region. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). The blue number corresponds to the node at which human sequences branch and indicates the average branch length for all human sequences. Branch lengths were used to estimate the time corresponding to the complex interlocus gene conversion event as shown. Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified.

3.7 Step 7: Duplicative transposition from BP2 to BP1 within chromosome 16p11.2

The high sequence identity (>99.6%) of an ~115 kbp duplication from BP2 to BP1 suggests this event likely occurred specifically along the human lineage (< ~1.8 mya).The resulting architecture rendered the interstitial region susceptible to NAHR-mediated inversion associated with asthma and obesity [21].

3.8 Step 8: Human ~450 kbp inversion polymorphism

The ~450 kbp inversion between BP1 and BP2 has been previously reported as an inversion polymorphism in humans. It was estimated to have occurred ~1.35 mya [21, 22].

3.9 Step 9: Tandem 102 kbp segmental duplication at BP5

The duplication architecture of BP4 haplotypes including *BOLA2B* matches the structure of BP5 haplotypes having a tandem duplication of the variable 102 kbp unit (**Fig. 4.1b**). In fact, the junction sequences between the end of the 72 kbp block and the start of the 30 kbp block (i.e., junctions between *SMG1P* and *BOLA2*) at BP4 are identical to the same junction sequence at BP5 in the H2 haplotype contig. Since the BP5 *BOLA2* paralog is ancestral (**Fig. S7**), we hypothesize that the *Homo sapiens*-specific duplication of *BOLA2* across the critical region originated from a BP5 haplotype having a tandem duplication of the variable 102 kbp unit. To test this hypothesis, we performed a series of sequence alignments and sliding window sequence identity analyses. The results corroborate our proposed scenario for the formation of the *BOLA2B* paralog, suggesting tandem duplication of the 102 kbp unit at BP5 preceded the duplication including *BOLA2* and the junction sequence from BP5 to BP4 (**Fig. S12a**). Because we do not observe duplicate *BOLA2* copies in archaic hominins, this tandem duplication most likely occurred specifically in *Homo sapiens*. It is possible this tandem duplication occurred before the human-archaic hominin species split but is absent from archaic hominins due to incomplete lineage sorting. We have no evidence that this tandem duplication is polymorphic in archaic hominins, although sampling of such genomes is limited.

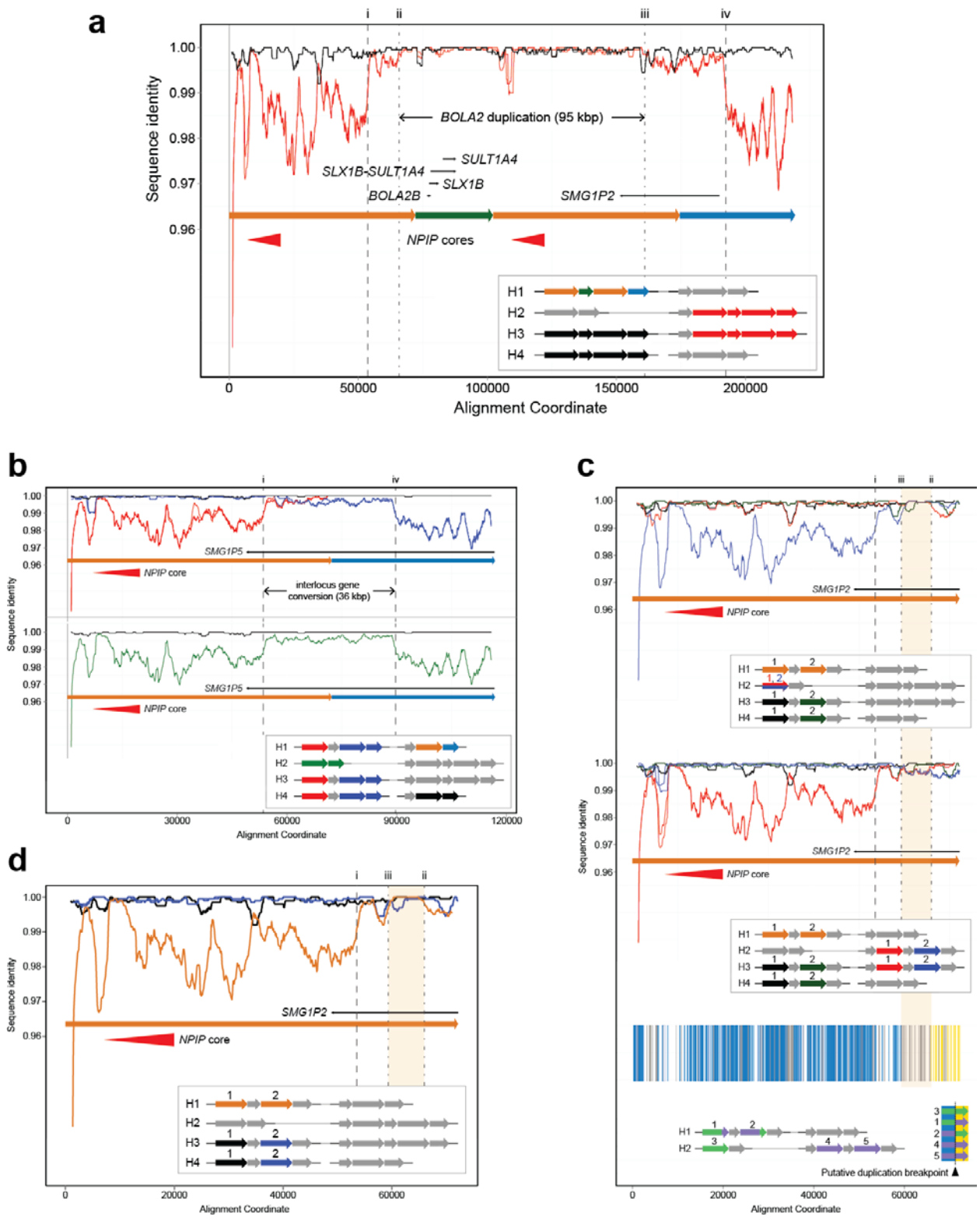


Figure S12. Sequence refinement of interspersed *BOLA2* duplication breakpoints. a) H1 human BP4 sequence (orange, green, orange, and blue arrows in inset) was aligned to its allelic (black arrows in inset) and paralogous (red arrows in inset) counterparts. The sequence identity for each alignment was computed and plotted over 2 kbp windows, sliding by 100 bp. Black lines indicate sequence identity for allelic comparisons, whereas red lines

correspond to paralogous comparisons. While the allelic comparisons exhibit uniform, near-perfect sequence identity across the entirety of the alignment, paralogous comparisons reveal three distinct levels of sequence identity, with the highest level in the middle. This pattern suggests that the *BOLA2* duplication (highest-identity region, 95 kbp) landed within an evolutionarily older segmental duplication having paralogs at BP4 and BP5 (Step 5). Dashed vertical lines (numbered i-iv) indicate putative breakpoints for events that occurred after this older segmental duplication. Junction sequence from the BP5 102 kbp tandem duplication was clearly included in the 95 kbp duplication from BP5 to BP4. b) Alignment of BP4 sequences containing the putative left (red arrows in inset) and right (dark blue arrows in inset) *BOLA2* duplication breakpoints to the BP5 paralog associated with the evolutionarily older segmental duplication (orange and light blue arrows in inset) and sliding window sequence identity analysis supports the hypothesis outlined above. Sequence identity lines for comparisons involving left and right BP4 sequences intersect in the vicinity of the hypothesized *BOLA2* duplication breakpoints. Comparing this result with the same analysis of the human H2 BP4 sequence lacking *BOLA2* (green arrows in inset and green identity line) suggests this BP4 sequence represents the ancestral state of BP4 before the *BOLA2* duplication arrived. Thus, two levels of sequence identity existed between BP4 and BP5 before the *BOLA2* duplication, consistent with an interlocus gene conversion event (section 3.6). c) Alignment of BP4 sequences (orange arrows in insets) containing the putative *BOLA2* duplication breakpoints to their ancestral BP4 (top plot) and their ancestral BP5 (middle plot) counterparts and sliding window sequence identity analysis reveals an ~7 kbp window (highlighted in orange) defining the *BOLA2* duplication breakpoints. Analysis of the underlying multiple sequence alignment (**Table S4**) identified positions with signatures informative for breakpoint localization (blue vertical lines, left BP4 72 kbp block outside of the *BOLA2* duplication and right BP4 72 kbp block within the *BOLA2* duplication; yellow vertical lines, left BP4 72 kbp block within the *BOLA2* duplication and right BP4 72 kbp block outside of the *BOLA2* duplication). Gray vertical lines indicate positions showing signatures of interlocus gene conversion. As both left and right 72 kbp block BP4 sequences within the ~7 kbp window are more highly identical to ancestral BP4 sequence (20/24 informative positions match the ancestral BP4 sequence) than to ancestral BP5 sequence, it is likely that this interval was involved in the *BOLA2* duplication but duplicated only within BP4. Its boundaries define the most likely *BOLA2* duplication breakpoints, and this pattern of sequence identity suggests a template switching replicative mechanism as most likely underlying the *BOLA2* duplication event (**Fig. S13**). d) Alignment of BP4 left and right 72 kbp sequences to each other (orange arrows in inset) and to their allelic counterparts (black and blue arrows in inset) and sliding window sequence identity analysis supports the template-switching scenario described above. Left and right 72 kbp blocks at BP4 are highly identical to one another primarily over the 7 kbp regions flanking the 95 kbp *BOLA2* duplication. The fact that these regions are immediately adjacent to the *BOLA2* duplication and have the same level of high identity as 95 kbp segments suggests duplication of the 7 kbp region within BP4 likely occurred concurrently with the *BOLA2* duplication from BP5 to BP4.

3.10 Step 10: Duplicative transposition of 95 kbp (including *BOLA2*) from BP5 to BP4 within chromosome 16p11.2

Analyses of sequence identity (**Fig. S12a-c**) and informative sites extracted from a multiple sequence alignment (**Fig. S12c** and **Table S4**) allowed us to resolve the breakpoints of the *BOLA2* duplication from BP5 to BP4 within 1 kbp. We delineate an ~95 kbp region within *BOLA2B*-containing BP4 haplotypes corresponding to *Homo sapiens*-specific duplicate sequence having originated from BP5 (**Fig. S12a**). The data suggest the duplication structure at BP4 in the H2 haplotype corresponds to the ancestral state of the BP4 locus (**Fig. S12b**). The *BOLA2* duplication likely involved template switching between BP4 and BP5 during DNA replication, resulting in the duplication of 95 kbp from BP5 to BP4 along with the duplication of a 7 kbp segment within BP4 (**Fig. S12d** and **Fig. S13**).

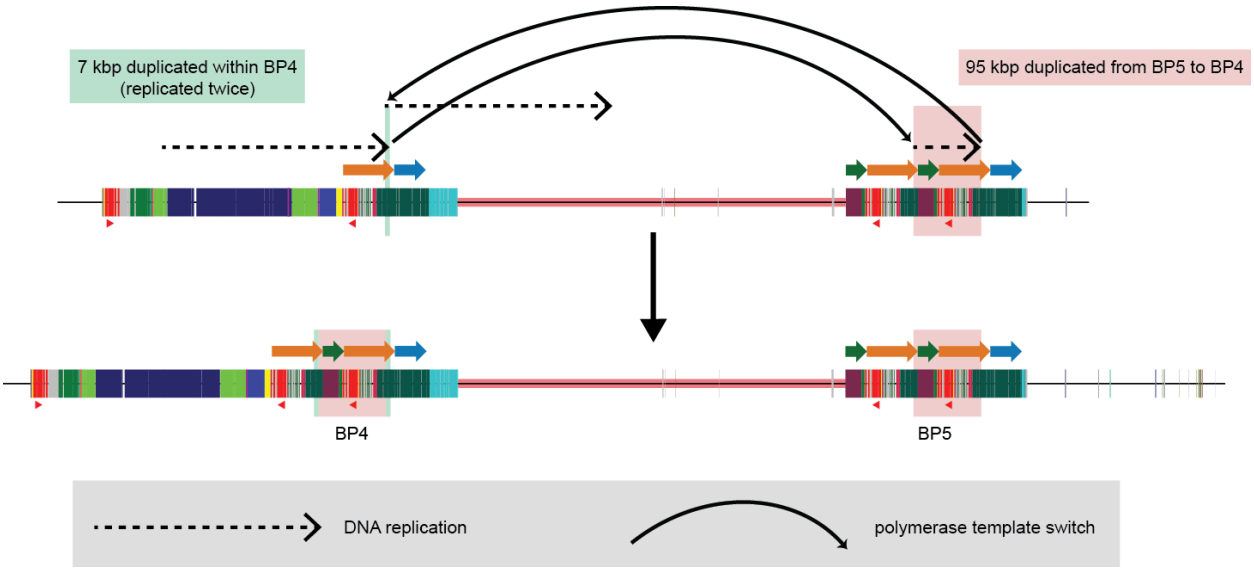


Figure S13. Template-switching model for the formation of *BOLA2B*. This mechanism was inferred from sequence identity analyses (**Fig. S12**) and from analysis of a multiple sequence alignment (**Table S4**).

To estimate the evolutionary timing of when the *BOLA2* duplication (**Fig. S14a**) occurred, we generated a multiple sequence alignment spanning an ~21 kbp region within the 30 kbp block including *BOLA2*, *SLX1*, and *SULT1A3* using sequences from our contigs and from a gorilla clone containing orthologous sequence (CH277-206A9). Estimates of molecular divergences between human sequences were very low (**Table S5**), and phylogenetic analysis did not show human sequences at BP4 and BP5 forming distinct clades (**Fig. 2b** and **Fig. S14b-c**). This result suggested either *BOLA2* duplicated very recently, such that the common ancestor of alleles at BP5 from our haplotype sequences is older than the *BOLA2* duplication event, or that interlocus gene conversion occurred between BP4 and BP5. To distinguish between these possibilities, we constructed a larger ~88 kbp alignment and phylogenetic tree using the full extent of sequence shared between all human paralogs. Here we observed sequences at BP4 and BP5 forming distinct clades (**Fig. S14d**), suggesting interlocus gene conversion underlies the branching pattern in the original tree. Assuming a human-chimpanzee divergence time of 6 mya [18] and a constant substitution rate (**Table S6**), we estimate that *BOLA2* duplicated across the critical region ~282 kya, around the time when contemporary *Homo sapiens* emerged as a species [23] (**Fig. 4.2b** and **Fig. S14b-c**). This estimate is consistent with our *BOLA2* copy number estimates in humans and archaic hominins (**Fig. 4.3a** and **Fig. 4.3c**).

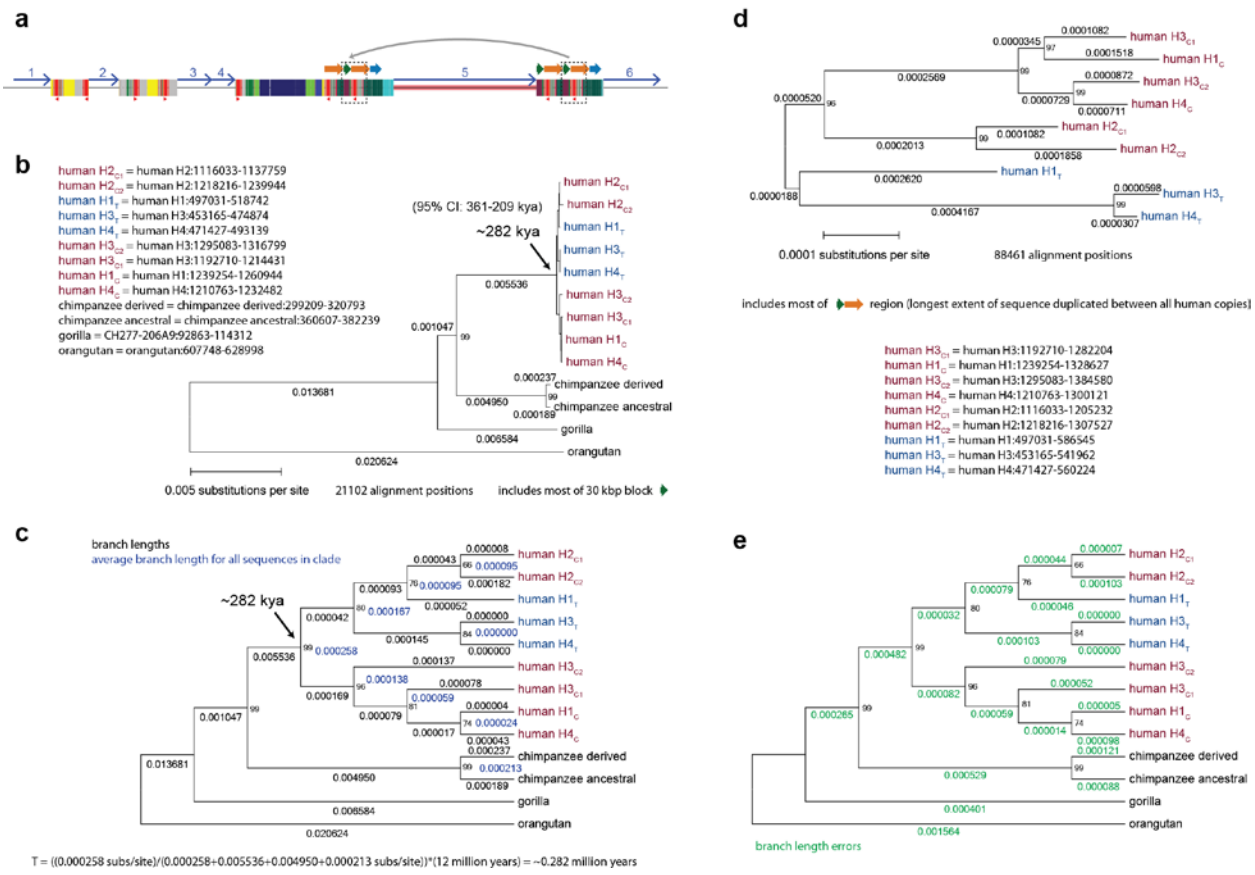


Figure S14. Phylogenetic characterization of the 95 kbp duplication including *BOLA2* from BP5 to BP4. a) Schematic shows regions involved in the duplication event (dashed boxes) and indicates direction of duplication. b) An unrooted neighbor-joining phylogenetic tree based on a 21,102 bp multiple sequence alignment spanning *BOLA2* and most of the 30 kbp block including human sequences from BP4 and BP5 and single-copy orthologous sequences from chimpanzee, gorilla, and orangutan. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified. c) Cladogram representation of the phylogenetic tree shown in panel b. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades. Branch lengths were used to estimate the time corresponding to the 95 kbp duplication including *BOLA2* from BP5 to BP4 as shown. d) An unrooted neighbor-joining phylogenetic tree based on an 88,461 bp multiple sequence alignment spanning the 30 kbp block and most of the 72 kbp block including all human 102 kbp block sequences from BP4 and BP5. This ~88 kbp region corresponds to the longest extent of sequence involved in both tandem and interspersed *BOLA2* duplication events and thus represents the longest segment shared between all human copies. Nonhuman primate species do not have contiguous 102 kbp blocks, so this 88 kbp segment cannot be used to estimate the timing of the interspersed *BOLA2* duplication event. However, it does clearly show BP4 and BP5 human sequences clustering separately, suggesting the branching pattern of the tree in panel b reflects interlocus gene conversion between BP4 and BP5 rather than the duplication history of *BOLA2* sequences. e) Cladogram representation of the phylogenetic tree shown in panel b, with green decimal numbers showing branch length error estimates used in determining a 95% bootstrap-based confidence interval for the timing estimate.

We computed a 95% confidence interval for our *BOLA2* duplication timing estimate using branch length error estimates (**Fig. S14e**) and the following procedure. First, for each branch in the tree, we set the branch length to a randomly selected value between the actual branch length minus the branch length error (or zero if that value is negative) and the actual branch length plus the branch length error, inclusive.

Second, we calculated a timing estimate using the same calculations as for the original tree except with modified branch length values. Third, we repeated the above two steps until one million modified trees and corresponding timing estimates were obtained. Fourth, we sorted the timing estimates and reported the 25,000th and 975,000th sorted timing estimate values as the 95% confidence interval: 361-209 kya.

3.11 Step 11: Polymorphic 102 kbp expansions and contractions at BP4 and BP5

Comparative sequence analyses of distinct human haplotypes delineate the nature and spatial extent of copy number variation within the 16p11.2 locus in humans and provide insight into the mechanism by which it occurs (**Fig. 4.1b**, **Fig. S3**, and section 2.1). Variation in human genomes is largely restricted to the 102 kbp segmental duplication including *BOLA2* at both BP5 (*BOLA2A*) and BP4 (*BOLA2B*) (**Fig. 4.3c** and section 4). Thus, we incorporate this knowledge into the final step of our evolutionary model for humans, highlighting a likely ongoing process of tandem expansions and contractions including *BOLA2* at BP4 and BP5 via NAHR.

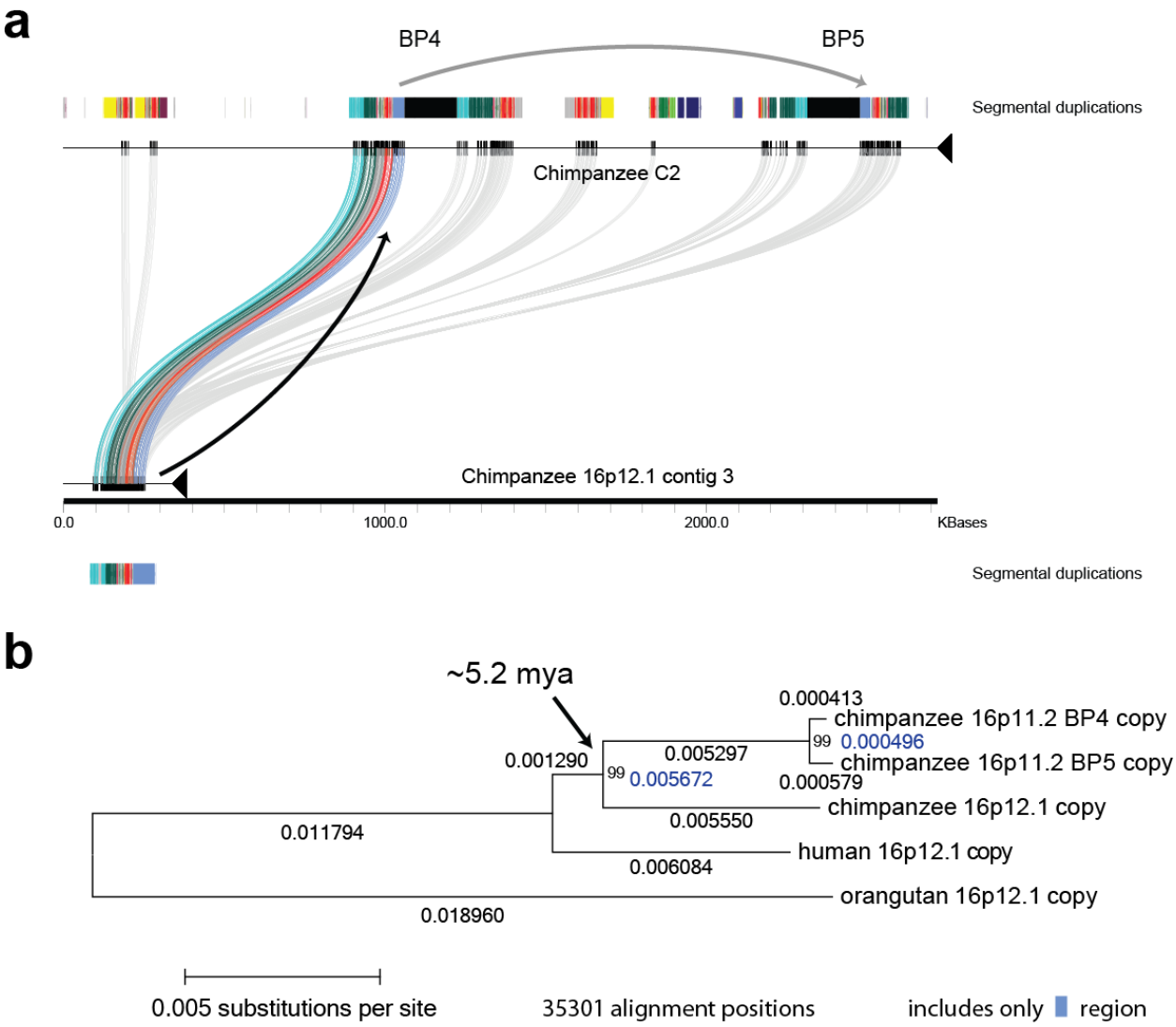
Chimpanzee-specific evolution of chromosome 16p11.2 (Steps 12-20)

3.12 Step 12: Chimpanzee-specific ~1.5 Mbp inversion

Our inversion analysis (section 3.2) suggests three inversions occurred specifically along the chimpanzee lineage. The largest such inversion included unique regions 3-5. The timing and order of this inversion relative to other chimpanzee-specific inversions cannot be inferred.

3.13 Step 13: Duplicative transposition from chromosome 16p12.1 to 16p11.2

The chimpanzee orthologs to human 16p11.2 BP4 and BP5 both contain chimpanzee-specific duplications originating from two separate locations at chromosome 16p12.1 (regions corresponding to colored lines and black boxes in **Fig. S15a**). The juxtapositions of these two 16p12.1 segments only at BP4 and BP5 implies that this duplication architecture first evolved at either BP4 or BP5, followed by the larger cassette duplicating to the other locus. The extent of contiguous sequence shared between BP4 and 16p12.1 is longer than that shared between BP5 and 16p12.1, and this extent includes junction sequence (between duplicons) not present at BP5 or at human BP4 (a proxy for chimpanzee BP4 prior to chimpanzee-specific duplications). These observations imply that the complex chimpanzee-specific duplication architecture first evolved at BP4. Phylogenetic analysis suggests the first chimpanzee-specific duplication (~170 kbp) from 16p12.1 to 16p11.2 BP4 occurred ~5.2 mya (**Fig. S15b**).



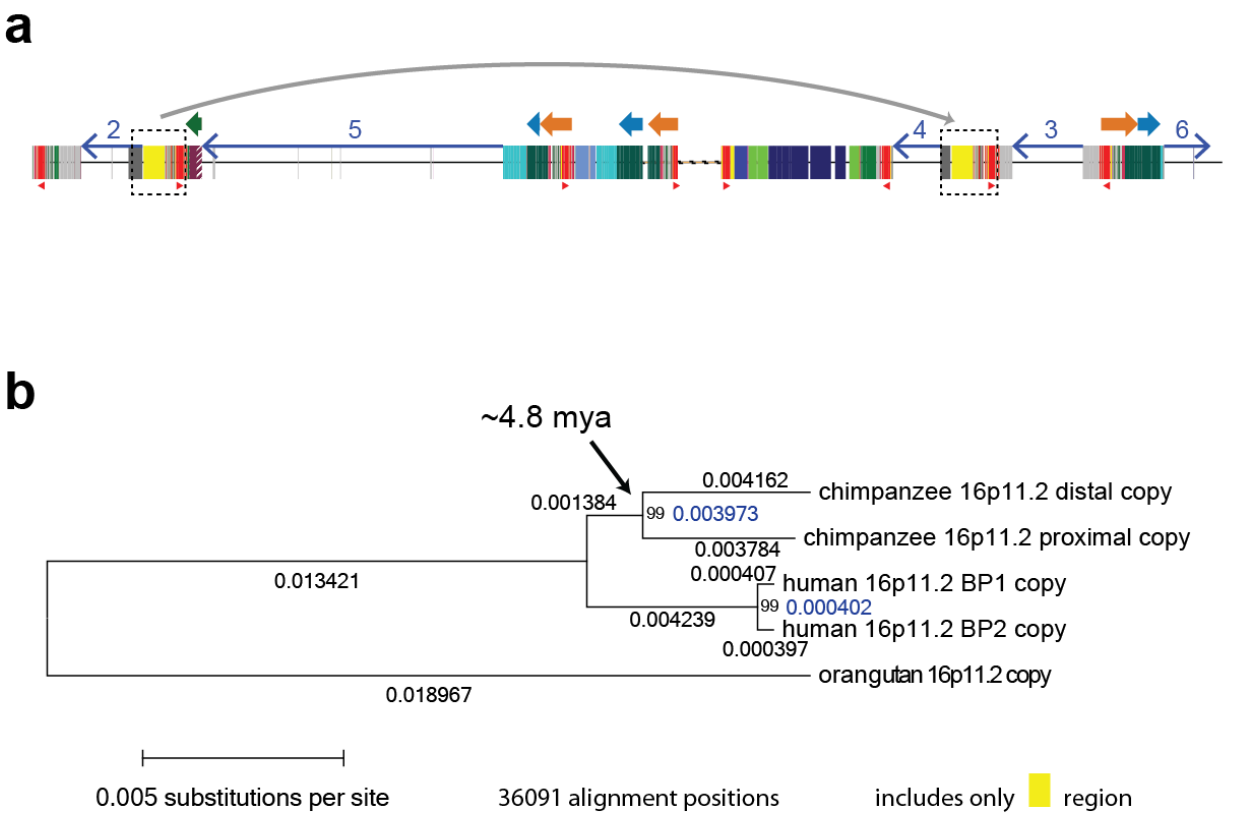
T = ((0.005672 subs/site)/(0.005672+0.001290+0.006084 subs/site))*(12 million years) = ~5.2 million years

Figure S15. Characterization of the chimpanzee-specific ~170 kbp duplicative transposition from 16p12.1 to BP4. a) Sequences (thin horizontal lines) from chimpanzee 16p11.2 (top, chimpanzee C2 contig) and chimpanzee 16p12.1 (bottom, chimpanzee 16p12.1 contig 3) are compared using Miropeats [12] (s = 1,000). Lines connecting the sequences show regions of homology, and colored lines highlight sequence included in the duplication event from 16p12.1 to 16p11.2 BP4. A later duplication copied much of this sequence from BP4 to BP5. b) An unrooted

neighbor-joining phylogenetic tree based on a 35,301 bp multiple sequence alignment including sequences from chromosome16p11.2 and 16p12.1. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades. Branch lengths were used to estimate the time corresponding to the duplication from 16p12.1 to 16p11.2 BP4 as shown. Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified.

3.14 Step 14: Duplicative transposition within 16p11.2 into unique sequence

Comparison of our sequenced haplotype contigs (**Fig. 4.1a**) reveals that a chimpanzee-specific ~130 kbp segment originating from BP2 duplicated to a region of unique sequence, separating regions 3 and 4 (**Fig. S16a**). Phylogenetic analysis suggests this chimpanzee-specific duplication occurred ~4.8 mya (**Fig. S16b**). Because not all sequence at the duplicate locus is also present at BP2, it is likely the duplication block between unique regions 3 and 4 formed via multiple events.



chimpanzee 16p11.2 distal copy = chimpanzee ancestral:254667-294125
chimpanzee 16p11.2 proximal copy = chimpanzee ancestral:1837639-1877168
human 16p11.2 BP1 copy = GRCh37/hg19 chr16:28387685-28427110
human 16p11.2 BP2 copy = GRCh37/hg19 chr16:28710844-28750237
orangutan 16p11.2 copy = orangutan:405006-445215

branch lengths
average branch length for all sequences in clade

T = ((0.003973 subs/site)/(0.003973+0.001384+0.004239+0.000402 subs/site))*(12 million years) = ~4.8 million years

Figure S16. Characterization of the chimpanzee-specific ~130 kbp duplicative transposition from BP2 to unique sequence. a) Schematic shows 16p11.2 regions involved (dashed boxes) in the duplication event that formed most of the duplication block between unique regions 3 and 4. b) An unrooted neighbor-joining phylogenetic tree based on a 36,091 bp multiple sequence alignment, including sequences from BP2 and the duplication block between unique regions 3 and 4. Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades. Branch lengths were used to estimate the time corresponding to the duplication as shown. Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified.

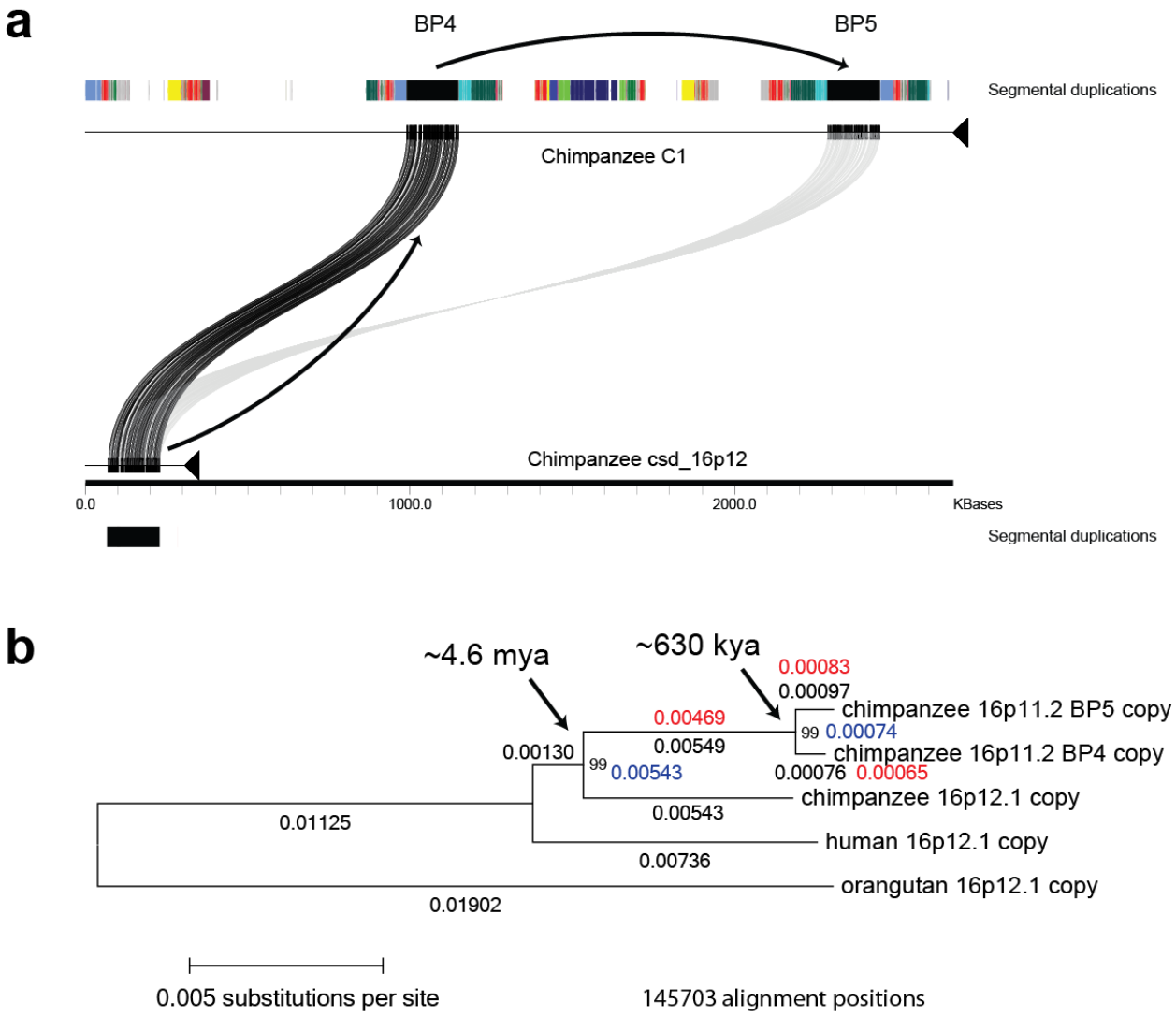
3.15 Step 15: Chimpanzee-specific ~215 kbp inversion

The chimpanzee-specific duplication that separated unique regions 3 and 4 included *NPIP*, resulting in unique region 3 becoming flanked by inversely oriented *NPIP* repeats. NAHR between these inverted

NPIP copies likely mediated the ~215 kbp inversion of unique region 3 found exclusively in chimpanzees.

3.16 Step 16: Duplicative transposition from chromosome 16p12.1 to 16p11.2

The second chimpanzee-specific duplication from 16p12.1 to 16p11.2 BP4 included ~160 kbp of unique sequence (**Fig. S17a**). Tajima's relative rate test indicated that the 16p11.2 copies did not evolve at the same relative rate as the 16p12.1 copy. Adjusting the 16p11.2 branch lengths accordingly, we estimate this duplication event occurred ~4.6 mya (**Fig. S17b**).



T1 = ((0.00543 subs/site)/(0.00543+0.00130+0.00736 subs/site))*(12 million years) = ~4.6 million years

T2 = ((0.00074 subs/site)/(0.00074+0.00469+0.00130+0.00736 subs/site))*(12 million years) = ~0.63 million years

Figure S17. Characterization of the chimpanzee-specific ~160 kbp duplicative transposition from 16p12.1 to BP4. a) Sequences (thin horizontal lines) from chimpanzee 16p11.2 (top, chimpanzee C1 contig) and chimpanzee 16p12.1 (bottom, chimpanzee csd_16p12 contig) are compared using Miropeats [12] (s = 1,000). Lines connecting

the sequences show regions of homology, and black lines highlight sequence included in the duplication event from 16p12.1 to 16p11.2 BP4. A later duplication copied all of this sequence from BP4 to BP5. b) Branch lengths (substitutions per site) are shown on each branch (black decimal numbers), and bootstrap support is indicated (black integers at nodes). Blue numbers correspond to nodes and indicate average branch lengths for all sequences in corresponding clades, while red numbers show either raw or clade-averaged branch lengths corrected to account for the difference in substitution rate between ancestral and duplicate loci. Branch lengths were used to estimate times corresponding to duplications into and within 16p11.2 as shown (see equations for T1 and T2, respectively). Coordinates (in base 1) of all sequences used for the alignment underlying the phylogenetic tree are also specified.

3.17 Step 17: Duplicative transposition of sequence to 16p11.2

The start of the chimpanzee C1 haplotype contig consists of >50 kbp of duplicated sequence not found in orthologous locations in human or orangutan, implying that this sequence resulted from a duplicative transposition event specific to the chimpanzee lineage.

3.18 Step 18: Duplicative transposition from BP4 to BP5 within 16p11.2

The largest (>420 kbp) chimpanzee-specific duplication transposed sequence from BP4 to BP5, resulting in large blocks of inversely oriented duplicated sequences flanking unique regions 3 and 4 in chimpanzee (**Fig. 4.1a**). Phylogenetic analysis suggests this duplication event occurred ~630 kya (**Fig. S17b**).

3.19 Step 19: Chimpanzee >1 Mbp inversion polymorphism

We discovered a large (>1 Mbp) inversion polymorphism in chimpanzee from our haplotype sequencing (**Fig. 4.1a**) and subsequently confirmed this inversion in a different chimpanzee individual via FISH (**Fig. S5**). We mapped the breakpoints of this inversion to the second chimpanzee duplication from 16p12.1 (black boxes in **Fig. 4.1a**, data not shown), implying the inversion occurred via NAHR between the large chimpanzee-specific inversely oriented duplication blocks at BP4 and BP5.

3.20 Step 20: Polymorphic tandem expansions including *NPIP*

Duplication analyses of our chimpanzee haplotype contigs revealed a polymorphic ~80 kbp tandem duplication at BP2 including *NPIP* and *EIF3C*, as well as several ~20 kbp duplications including only *NPIP*.

4. Copy number genotyping

4.1 Overview

We employed three complementary methods to genotype *BOLA2* copy number in a total of 2,824 humans [24, 25], 3 archaic humans [26, 27], 1 Neanderthal [28], 1 Denisovan [29], 14 bonobos [30], 23 chimpanzees [30], 32 gorillas [30], and 17 orangutans [30]. Genotypes for all samples are provided in **Table S7**, and **Table S8** shows population summary statistics. First, we estimated aggregate *BOLA2* copy number, along with *SLX1* copy number and *SULT1A3* copy number, using a previously described approach based on WGS read depth [31]. Second, we inferred aggregate and paralog-specific *BOLA2* copy number by examining relative WGS read depth over genetic markers distinguishing regions of interest. Third, we targeted a subset of these genetic markers as well as polymorphisms for molecular inversion probe (MIP) capture. Using massively parallel sequencing, we determined aggregate and paralog-specific *BOLA2* copy number using relative read-depth analysis [32] as above. Although only a subset of samples were genotyped using multiple methods (**Table S7**), distributions of aggregate *BOLA2* copy number estimates were similar among the methods (**Fig. S18**). This suggests that results from each method reliably reflect the distribution of aggregate *BOLA2* copy number genotypes in the human population, except for high-copy estimates based on WGS read depth which showed the lowest validation rate. We detail each method below, with particular focus on the second approach which has not been previously described.

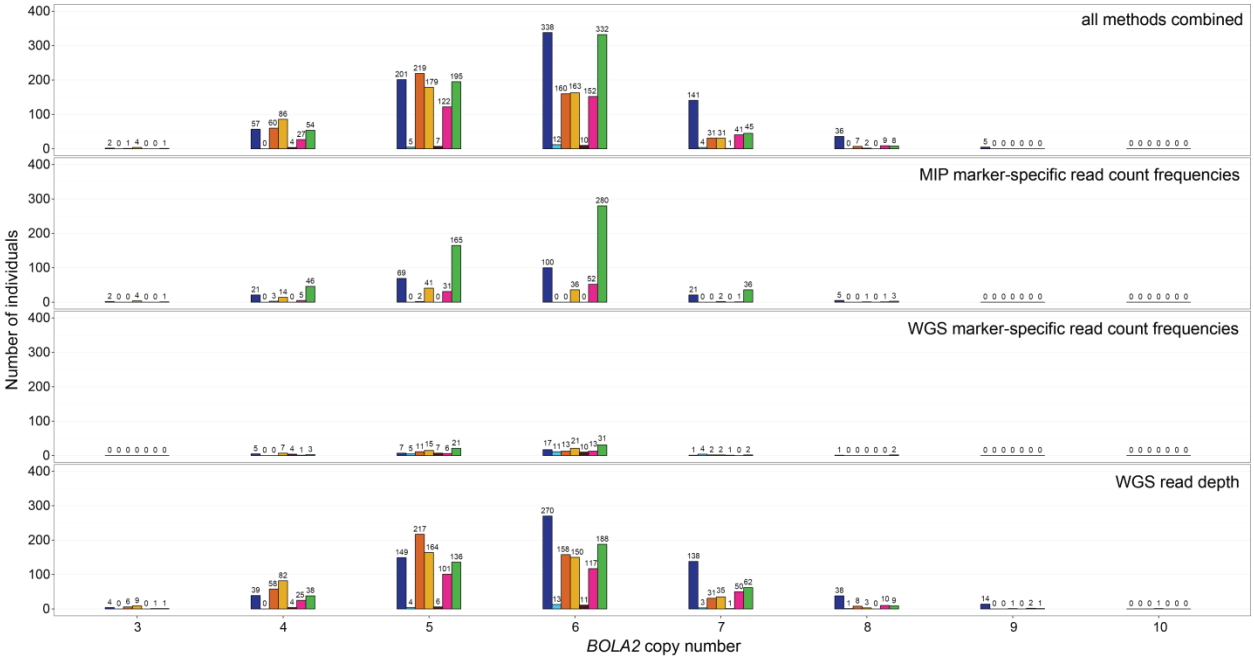
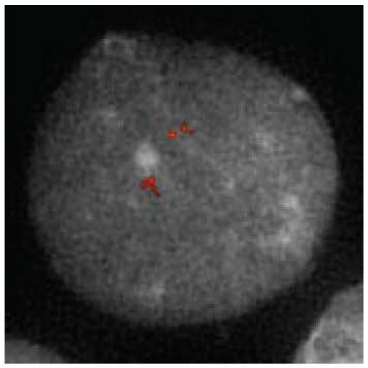


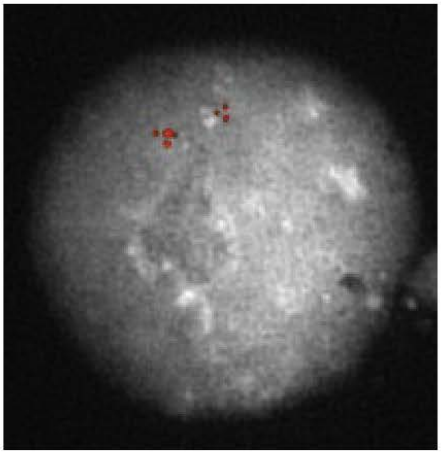
Figure S18. Comparison of aggregate *BOLA2* copy number estimates among the different genotyping methods. Histograms show counts of individuals having different *BOLA2* copy number estimates including results from all methods combined (top panel) or results from only a single method (bottom three panels). No method predicts any human having two copies of *BOLA2* in aggregate. Colors correspond to different populations as in **Fig. 4.3a and c**.

4.2 Aggregate *BOLA2*, *SLX1*, and *SULT1A3* copy number genotyping using WGS read depth

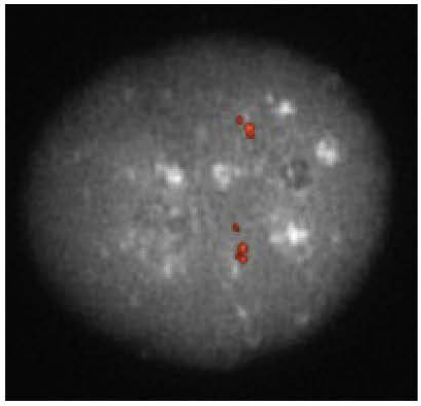
We genotyped aggregate copy number for all genes within the 30 kbp segment duplicated between BP5 and BP4 using a WGS read-depth method as previously described [31]. We measured sequence read depth over an ~5 kbp region that extends from just beyond *CORO1A* to *BOLA2* (GRCh37 chr16:30200398-30205627). To test the accuracy of these estimates, we performed FISH on cell lines derived from individuals having predicted *BOLA2* copy numbers from three to nine using fosmid probes overlapping part of the 30 kbp segment (including *BOLA2*) as well as unique sequence adjacent to this block at BP5 (**Table S2**). FISH analysis validated individuals having *BOLA2* copy number as low as three and as high as eight (**Fig. 4.3b**) but also showed some discrepancies (3 of 7 genotypes were discordant for higher copy number; **Fig. 4.3a and b** and **Fig. S19**). Higher copy numbers are more difficult to discern because tandem *BOLA2* copies are only 102 kbp apart and such FISH signals cannot always be discriminated on interphase nuclei. Low sequence coverage for samples from the 1000 Genomes Project [24] is another source of error. We have previously shown that the accuracy of copy number estimates based on WGS read depth correlates positively with increasing sequencing coverage [31]. As a result, we consider copy number estimates from 236 genomes from the Simons Genome Diversity Project [25] to be more accurate than those from 1000 Genomes Project genomes. With one exception, all FISH estimates were concordant with WGS read-depth estimates within 1 diploid copy number.



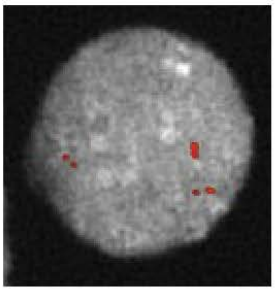
HG02314, *BOLA2* copy number = 7
(copy number 7 predicted by WGS read depth)



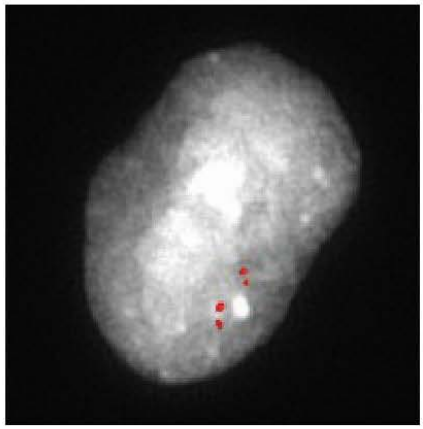
NA19041, *BOLA2* copy number = 7
(copy number 8 predicted by WGS read depth)



NA12275, *BOLA2* copy number = 7
(copy number 9 predicted by WGS read depth)



NA20127, *BOLA2* copy number = 6
(copy number 6 predicted by WGS read depth)



NA12878, *BOLA2* copy number = 4
(copy number 4 predicted by H1 and H2 contig sequences)

Figure S19. Interphase FISH experiments for individuals having a range of *BOLA2* copy number estimates based on the WGS read-depth method or our high-quality haplotype sequences.

Since *SLX1* and *SULT1A3* were also part of the 102 kbp *Homo sapiens*-specific duplication involving *BOLA2*, we reasoned that they may also be genes duplicated specifically in our species. We assessed their copy number using a genomic segment (~4 kbp) corresponding to *SLX1* (GRCh37 chr16:30205164-30208887) and an ~11 kbp segment including *SULT1A3* (GRCh37 chr16:30210549-30221660). Only *BOLA2* shows convincing evidence of being a *Homo sapiens*-specific duplicated gene (**Fig. 4.3a** and **Fig. S20**). We identify a few humans with potentially two copies of *SLX1*, and multiple nonhuman primates potentially have three *SLX1* copies. We cannot definitively exclude the possibility that *SLX1* is duplicated specifically in *Homo sapiens* due to the imprecision associated with genotyping such a small genomic segment using this method. In contrast, all nonhuman primates were estimated to have three or more copies of *SULT1A3*. *SULT1A3* is unambiguously duplicated in nonhuman primates—a finding confirmed by the identification of a chimpanzee BAC clone from chromosome 17 harboring a duplicate *SULT1A3* paralog (RP43-175I2).

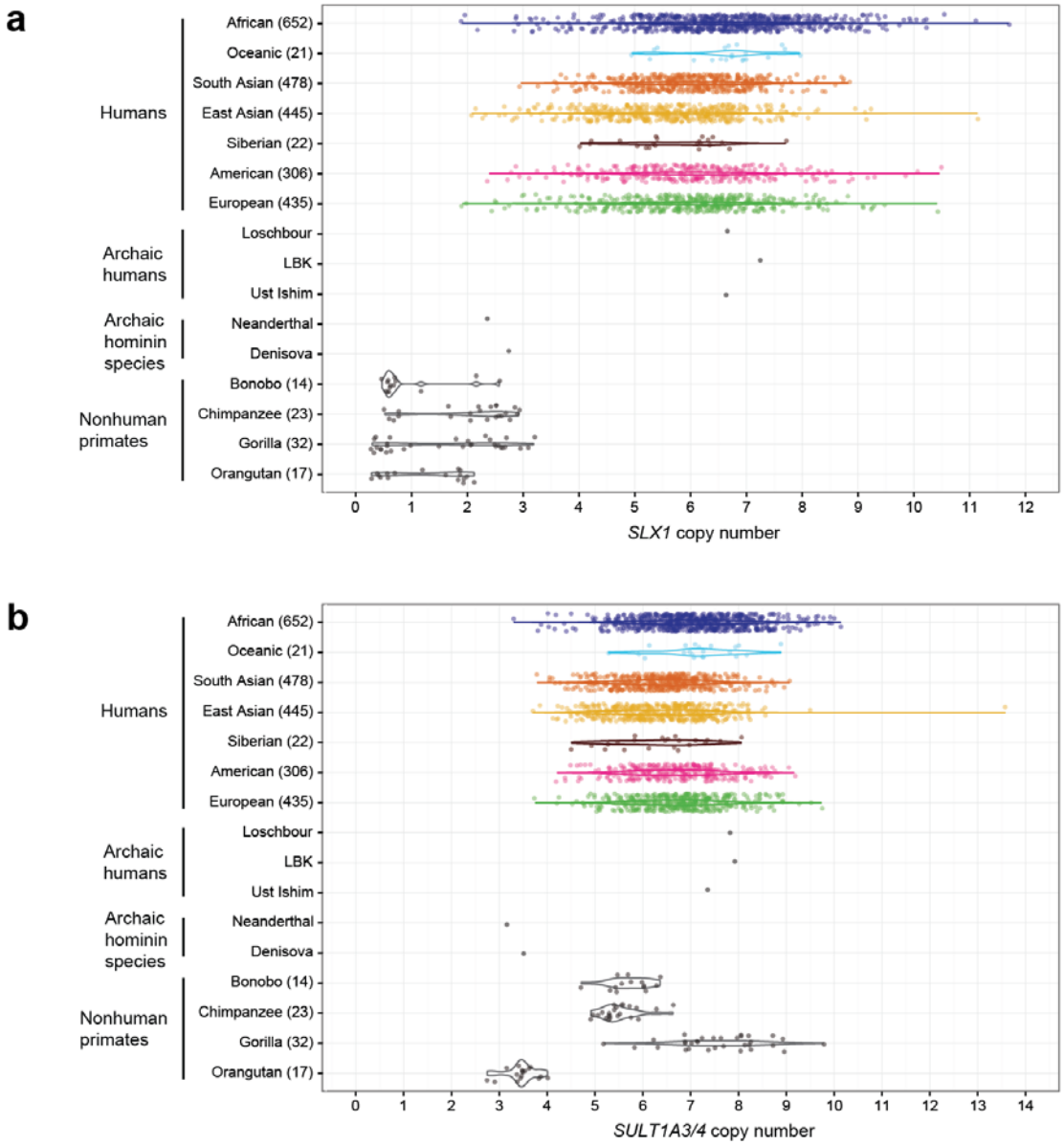


Figure S20. *SLX1* and *SULT1A3* copy number diversity. Diploid copy number estimates (points) for *SLX1* (panel a) and *SULT1A3* (panel b) based on WGS read depth are shown for 2,359 humans from seven populations, three archaic humans, a Neanderthal, a Denisovan, and 86 nonhuman primates, with violin plots overlaid.

4.3 *BOLA2* paralog-specific copy number (PSCN) genotyping

We also genotyped *BOLA2* paralog-specific copy number (PSCN) by adapting previously described strategies [31, 32] based on sequencing read depth over informative paralogous sequence variants. Informative genetic markers have three properties: i) they distinguish sequences of interest (16p11.2 BP4 and BP5) from paralogous sequences elsewhere in the genome (**Fig. S1**); ii) they distinguish different copies (e.g., BP4 from BP5 copies); and iii) they vary predictably with differences in paralog-specific *BOLA2* copy number. Sequencing reads containing such markers can be unambiguously assigned to particular copies of duplication blocks at 16p11.2, allowing quantification and inference of *BOLA2* PSCN.

Using our high-quality human haplotype sequences, we identified 70-mer markers strictly meeting the criteria above only within the 72 kbp block (**Figs. S1-S2**). Copy number of the 72 kbp block varies in concert with *BOLA2* copy number. The 72 kbp block and the 30 kbp block (which includes *BOLA2*) together constitute the 102 kbp variable unit. Thus, aggregate *BOLA2*, *BOLA2A* (BP5), and *BOLA2B* (BP4) copy number can be deduced from knowledge of total and BP5 72 kbp block copy number.

BP4 72 kbp blocks from *BOLA2B*-containing haplotypes are comprised of both ancestral BP4 sequence and sequence derived from the 95 kbp duplication including *BOLA2* (**Fig. S12**). This hybrid architecture effectively divides the 72 kbp block into three sectors (**Fig. S21a**): i) a 59 kbp sector with markers distinguishing the most telomeric BP4 copy (B markers, blue box) from all other copies (R markers, red boxes); ii) a 7 kbp sector with markers distinguishing all BP4 copies (green boxes) from all BP5 copies (BP5 markers, orange boxes); and iii) a 6 kbp sector with markers distinguishing the most centromeric BP4 copy (purple box) from all other copies (yellow boxes). Analyses of our sequenced human haplotypes reveal that B markers never vary in copy number—each haplotype had exactly one complete set of these markers. Thus, diploid genomes contain two complete sets of B markers.

We leveraged WGS sequencing data and all 59 kbp sector markers to determine aggregate 72 kbp block copy number as follows. We extracted all reads containing a 70-mer perfectly matching a 59 kbp sector marker and counted the number of reads corresponding to each such marker. For every marker pair (a B marker and its R marker counterpart), we plotted marker-specific read count frequencies, calculated as the number of reads corresponding to the marker of interest (B or R) divided by the number or reads corresponding to both markers. Because there are two complete sets of B markers in a diploid genome, B marker-specific read count frequencies have values around 2/N, where N is the total number of 72 kbp blocks. Thus, the total number of 72 kbp blocks can be accurately inferred from B marker-specific read count frequency data. Once the total number of 72 kbp blocks is known, performing the same analysis of WGS data using markers within the 7 kbp sector allows inference of BP5 72 kbp block copy number. BP5 marker-specific read count frequencies have values around C/N, where C is the number of BP5 72 kbp blocks and N is the total number of 72 kbp blocks, as above.

We evaluated this inferential strategy by applying it to genotype *BOLA2* PSCN for NA12878 using high-coverage WGS data [33, 34]. Importantly, we know this individual has three copies of *BOLA2A* and one copy of *BOLA2B* because this individual was the source of genomic material for the BAC and fosmid libraries utilized in generating our H1 and H2 human haplotype sequences. The results show that using this approach, we can accurately infer paralog-specific *BOLA2* copy number as well as PSCN for the 45 kbp block (**Fig. S21**).

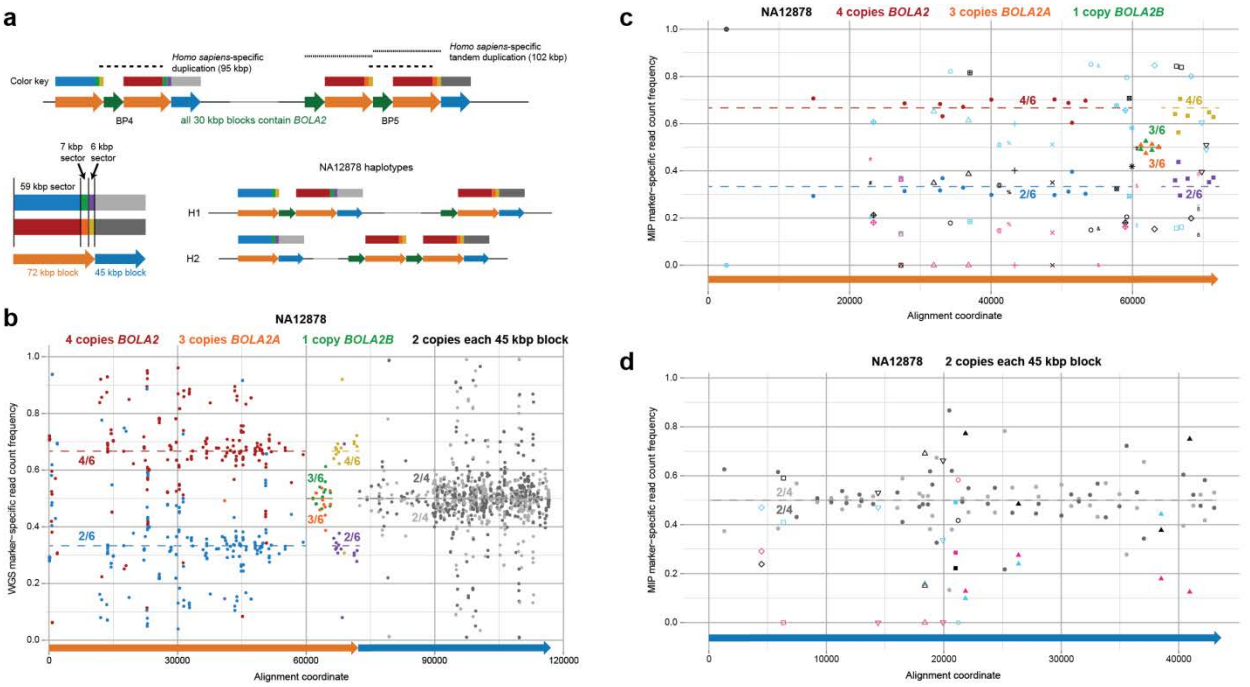


Figure S21. Accurate paralog-specific *BOLA2* copy number inference. a) Top and left schematics detail three distinct sectors of the 72 kbp blocks (orange arrows). Each block has paralogous sequence variants that are informative for particular region(s) when compared to others in chromosome 16p11.2. These markers are color-coded into three sectors within the 72 kbp block of paralogy (a 59 kbp sector, blue and red boxes; a 7 kbp sector, green and orange boxes; and a 6 kbp sector, purple and yellow boxes), indicating which particular regions they distinguish. Bottom-right schematic shows known haplotype structures for individual NA12878. b) Analyzing WGS data from NA12878 yields copy number estimates for *BOLA2A* and *BOLA2B* that match the known *BOLA2* PSCN for this individual. Each point shows a relative marker-specific read count frequency (y-axis) and its position within the duplication blocks. Point colors correspond to different marker sets for each sector, as diagramed in panel a. Fractions indicate the relative copy number of each marker set. Estimates of 4/6 (red marker set) vs. 2/6 (blue marker set) for the 59 kbp sector confirms the sequenced architecture (panel a) with an aggregate of 4 *BOLA2* copies, and the estimate of 3/6 (orange marker set) confirms three copies of *BOLA2A*. c) Using MIPs, we employed the same relative read-depth strategy. Genotyping results for the same sample as in b are shown, with additional markers (points not color-coded as in panels a-b) added based on polymorphic variants. MIP genotypes confirm WGS estimates (in panel b). d) MIP genotyping also yields accurate PSCN estimates for the 45 kbp block. Points colored as in panel b distinguish BP4 and BP5 45 kbp blocks (like same-colored markers in panel b), and points having other colors correspond to polymorphic variants.

We employed this assay to estimate *BOLA2* PSCN for 236 humans from the Simons Genome Diversity Project sequenced to high coverage [25], three archaic humans [26, 27], a Neanderthal [28], and a Denisovan [29]. We excluded all WGS data from the 1000 Genomes Project [24], as the sequencing coverage for corresponding genomes is too low for accurate PSCN estimation using this method. For 14 humans (a subset of individuals of African ethnicity) and for all three archaic human genome sequence datasets, we could not confidently infer *BOLA2* PSCN, though the data clearly show *BOLA2* is duplicated in each of these individuals. Nevertheless, using this approach, we were able to estimate high-confidence *BOLA2* PSCN genotypes for 94% of humans (222 of 236, **Fig. S22a**) as well as for the Neanderthal and the Denisovan (**Fig. S23**).

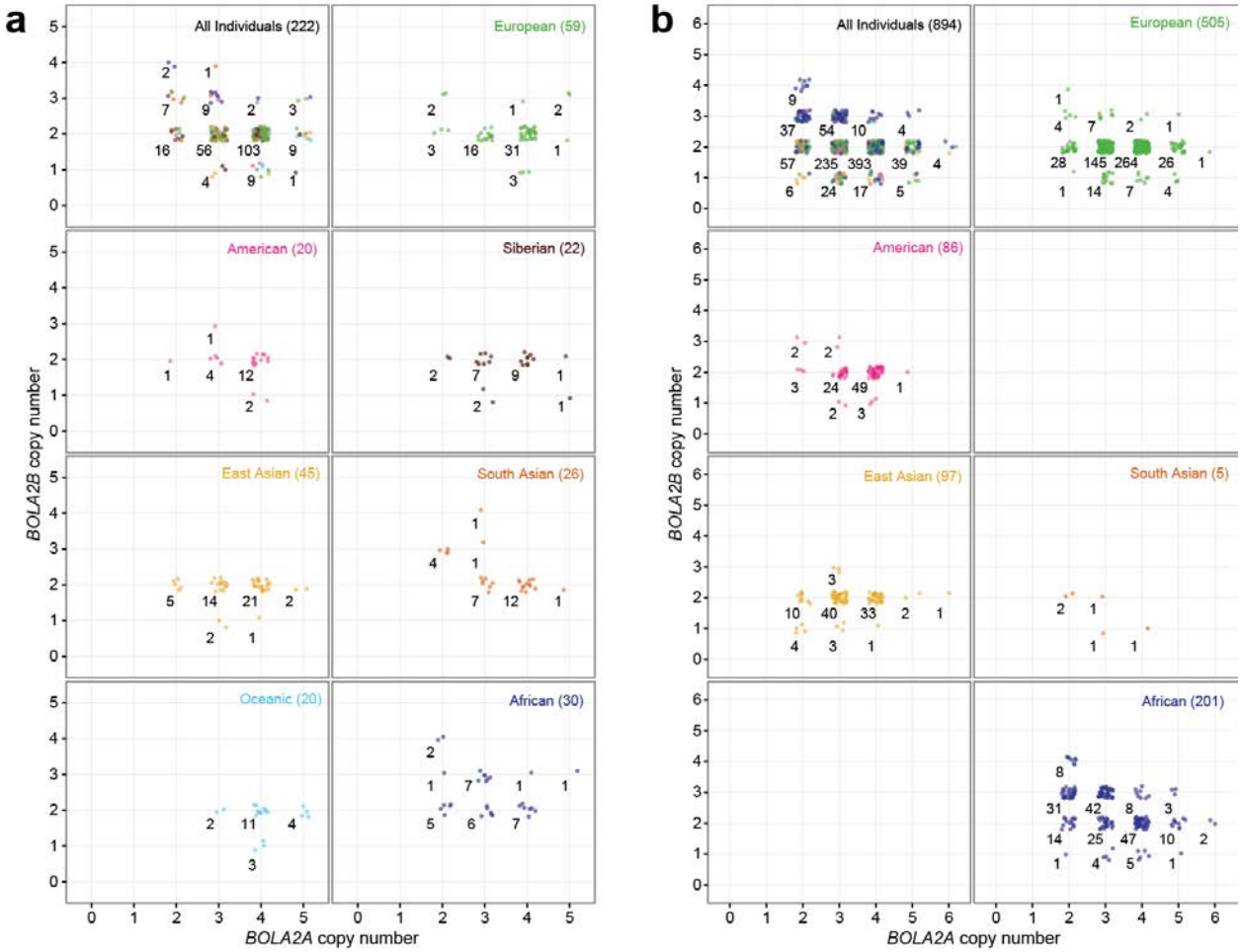


Figure S22. *BOLA2* PSCN distributions in different human populations. a) *BOLA2* PSCN genotypes (points, jittered around their integer values for clarity) were inferred from WGS read depth over informative markers for 222 individuals. Numbers indicate total counts of individuals in each population having a particular *BOLA2* PSCN genotype. b) Plots show *BOLA2* PSCN genotypes inferred from MIP sequence data from 894 humans. Low-confidence estimates were excluded. Points and numbers follow the same convention as in panel a.

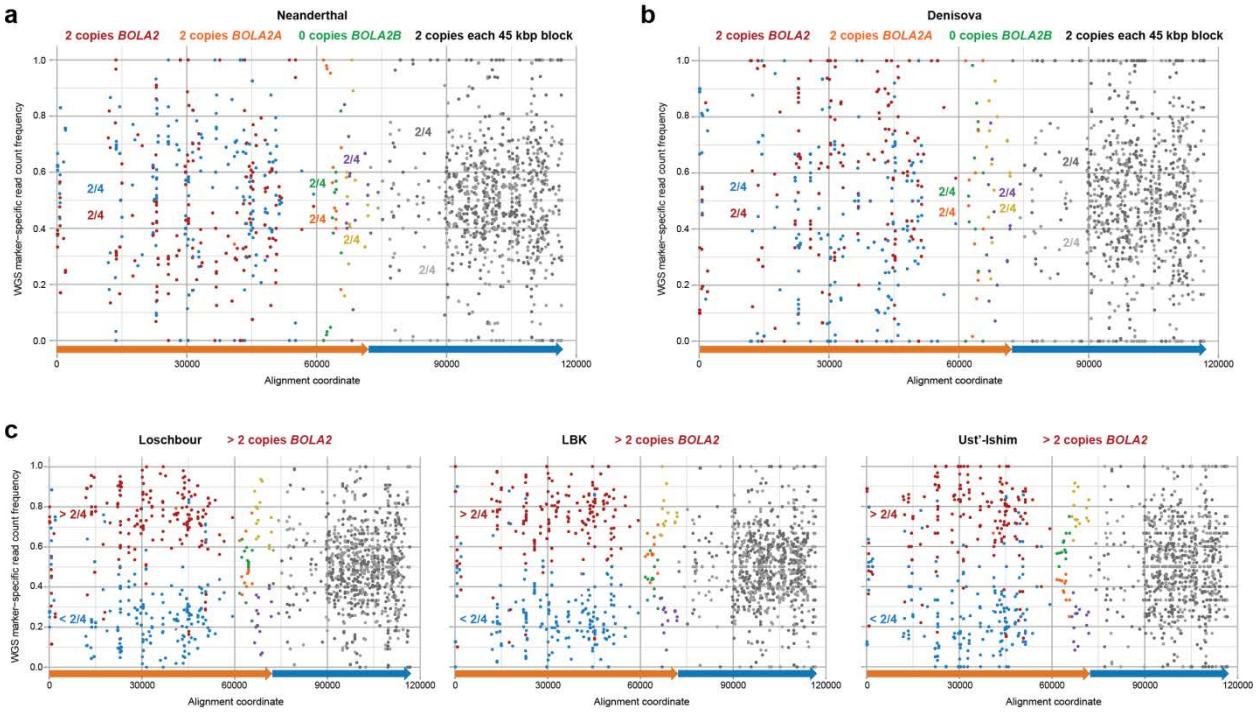


Figure S23. *BOLA2* PSCN estimates from archaic genomes. Plots follow the same convention as in **Fig. S21b**. The complete intermixing of marker-specific read count frequencies for markers over the 59 kbp sector (blue and red points) in Neanderthal and Denisova is consistent with these species having only four total 72 kbp blocks and two copies of *BOLA2*, confirming our copy number estimates based on WGS read depth. In contrast, all archaic humans have >4 total 72 kbp blocks and >2 *BOLA2* copies. Data from markers over the 7 kbp sector indicate that both Neanderthal and Denisova had two copies of *BOLA2A*, as expected based on our phylogenetic estimate of *BOLA2B* age. This same data indicate all archaic humans examined had at least one copy of *BOLA2B*.

Our analysis indicates that *BOLA2A* copy number (standard deviation = 0.77) is more variable than *BOLA2B* copy number (standard deviation = 0.46). This differential variability is consistent with the nearly identical 102 kbp tandem duplications within BP5 being more likely to mediate NAHR than the directly oriented duplications of the 72 kbp block within BP4, which contain only a few small regions of near-perfect sequence identity (**Fig. S24**). Our genotyping results also suggest that some individuals have more than two copies of *BOLA2B*. Prior to our WGS analyses, we developed a FISH assay for genotyping paralog-specific *BOLA2* copy number (**Table S2**). Interphase FISH on the first individual tested revealed both chromosome 16 homologs having *BOLA2* paralogs at BP4 and BP5. The fluorescence intensities showed evidence of one homolog harboring multiple copies of *BOLA2A* and the other including multiple copies of *BOLA2B* (**Fig. S25**). This finding provides experimental confirmation that individuals exist in the human population having more than two copies of *BOLA2B*.

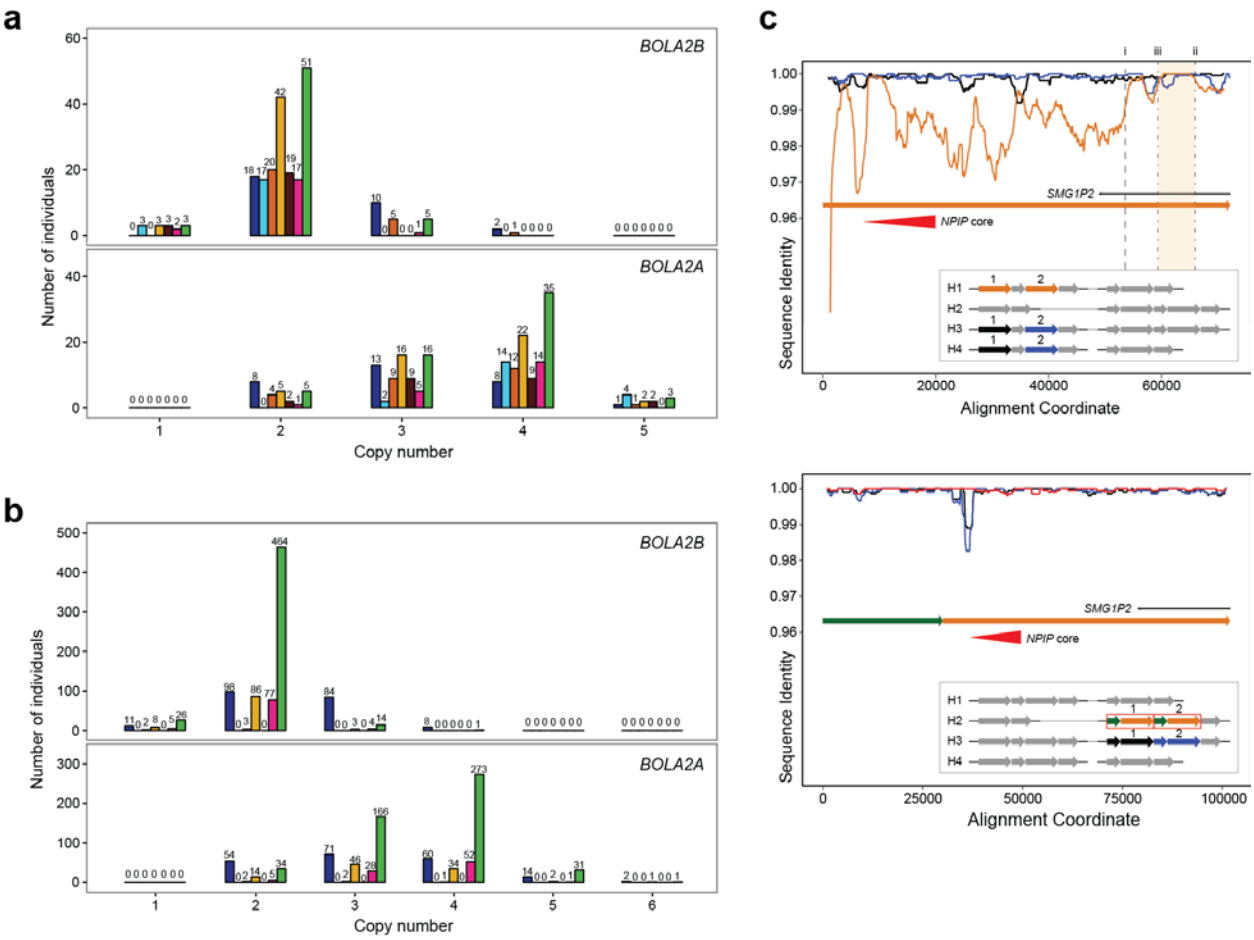


Figure S24. Differential paralog-specific *BOLA2* copy number variability and its likely basis in different sequence identity profiles for tandem duplications within BP4 compared to those within BP5. a) Histograms show counts of individuals having each paralog-specific *BOLA2* copy number genotype inferred from WGS marker-specific read count frequency data, with different-colored bars corresponding to different populations as in **Fig. 4.3a and c**. b) Histograms follow the same convention as in panel a but show genotypes inferred from MIP analysis. c) The near-perfect sequence identity (red line, bottom plot) of tandemly duplicated 102 kbp segments at BP5 renders them highly prone to tandem expansions and contractions via NAHR. In contrast, directly oriented duplications of the 72 kbp block within BP4 exhibit only small patches of near-perfect sequence identity (orange line, top plot), with the largest such patch 7 kbp in length (between iii and ii, highlighted in red). Accordingly, these duplications are less likely than those within BP5 to participate in NAHR and mediate expansions and contractions of *BOLA2B*. Black and blue lines in both sequence identity plots show allelic comparisons specified by the insets as in **Fig. S12**.

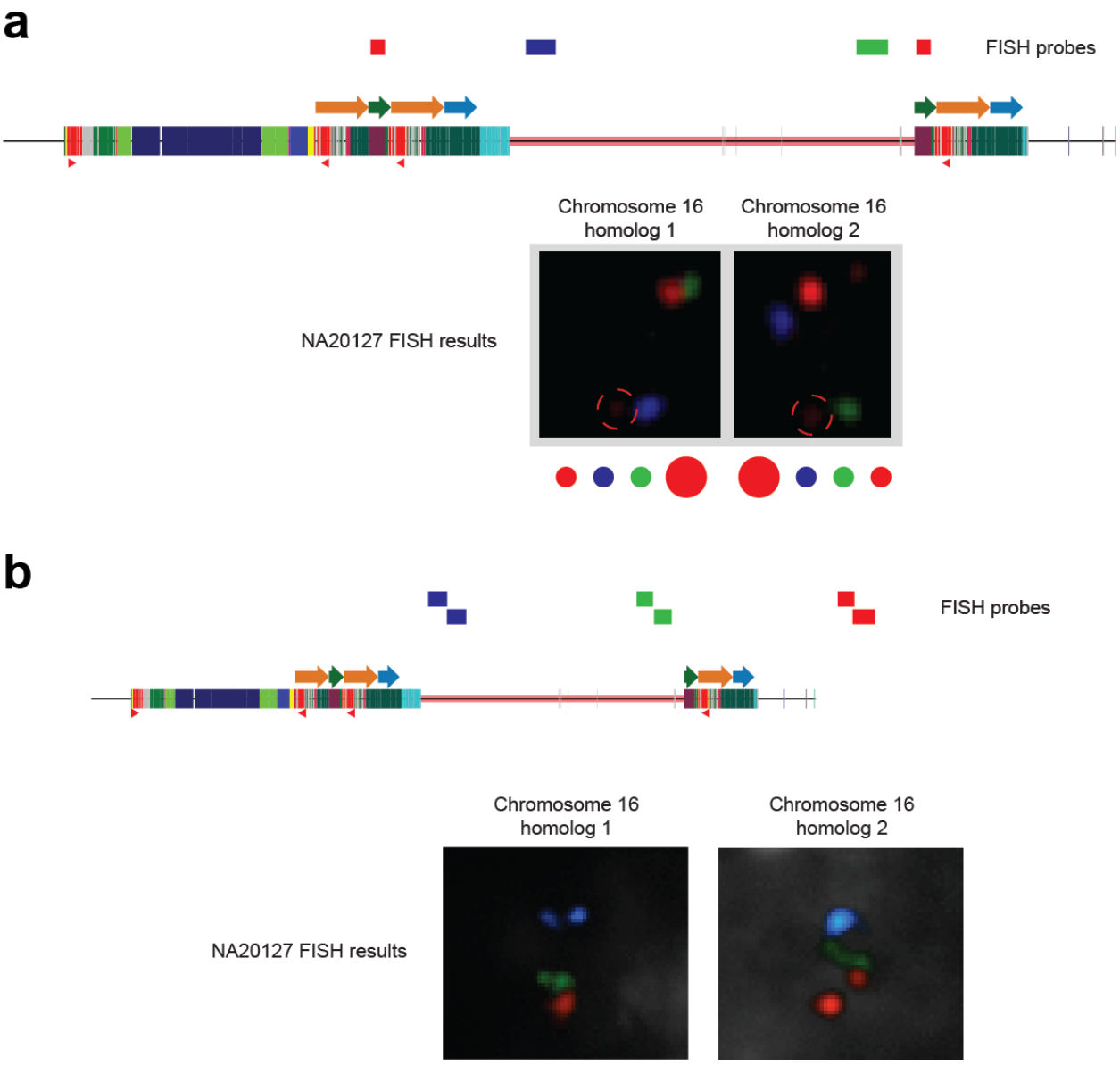


Figure S25. Interphase FISH confirms NA20127 has three copies each of *BOLA2A* and *BOLA2B*. a) Previous FISH analysis (**Fig. S19**) suggests the individual NA20127 has six total copies of *BOLA2*. Diagram outlines a three-color FISH assay including two probes (blue, green) targeting sequences within the autism critical region and one probe (red) targeting ~18 kbp of sequence (including *BOLA2*) over the 30 kbp duplication block. Signals from the red probe are detected on the telomeric (BP4) and centromeric (BP5) sides of the critical region (adjacent to the blue and green probes, respectively) on both chromosome 16 homologs. However, the red probe signal intensity is strongest adjacent to the green probe for one homolog but, in contrast, is strongest adjacent to the blue probe for the other chromosome 16 homolog, consistent with higher *BOLA2A* copy number in the first case and higher *BOLA2B* copy number in the second case. b) This differential signal intensity pattern does not result from an inversion of the 16p11.2 critical region in this individual, as data from the FISH experiment detailed here refute this possibility. Information on probes used in these FISH experiments is provided in **Table S2**.

The PSCN genotyping method has some limitations. It requires high sequencing coverage and interlocus gene conversion events over markers critical for copy number inference may yield inaccurate PSCN genotypes if not detected and taken into account. Small focal deletions or duplications affecting *BOLA2*

but not the 72 kbp block would be completely missed using our approach. Finally, marker-specific read count frequencies become more difficult to distinguish as the aggregate copy number of the 72 kbp block increases. For example, distinguishing 2/9 from 2/10 is much more difficult than differentiating 2/6 from 2/7. Thus, higher *BOLA2* copy number estimates using this method are inherently less confident than lower *BOLA2* copy number estimates. All these considerations also apply to the MIP method detailed in section 4.4 below, though the higher sequence coverage afforded by MIPs (>20-fold) increases genotyping precision [32].

4.4 *BOLA2* PSCN genotyping using MIPs

To enable large-scale genotyping of *BOLA2* PSCN, we designed MIPs targeting 112-mer markers differentiating the same groups of 72 kbp blocks as in our WGS marker-specific read count frequency analyses (section 4.3). 112 bp informative k-mers were specifically selected because this represents the distance between MIP targeting arms. We used the same general parameters for MIP design as previously described [32], except the copy count threshold for arm hybridization sequences was 30 rather than 8 and no filtering was performed based on target G+C content. Only 20 MIPs were successfully designed meeting the above criteria. We also designed an additional 27 MIPs corresponding to polymorphic single-nucleotide variants to improve genotyping precision. We performed MIP capture, sequencing, and analysis as previously described [32] using single-molecule MIPs [35] (**Table S9**), mapping sequencing data to a minimal genome consisting of all BP4 and BP5 72 kbp blocks plus all paralogous sequences from GRCh37 outside 16p11.2 (**Fig. S1**). We inferred *BOLA2* PSCN as described above (**Fig. S21**) for 894 humans (**Fig. S22b** and **Fig. S24b**). Importantly, we used this method to evaluate *BOLA2* copy number estimates for individuals at the extremes of the aggregate *BOLA2* copy number estimate distribution based on WGS read depth, confirming *BOLA2* copy number is at least three and at most eight among all humans examined (**Fig. S26**).

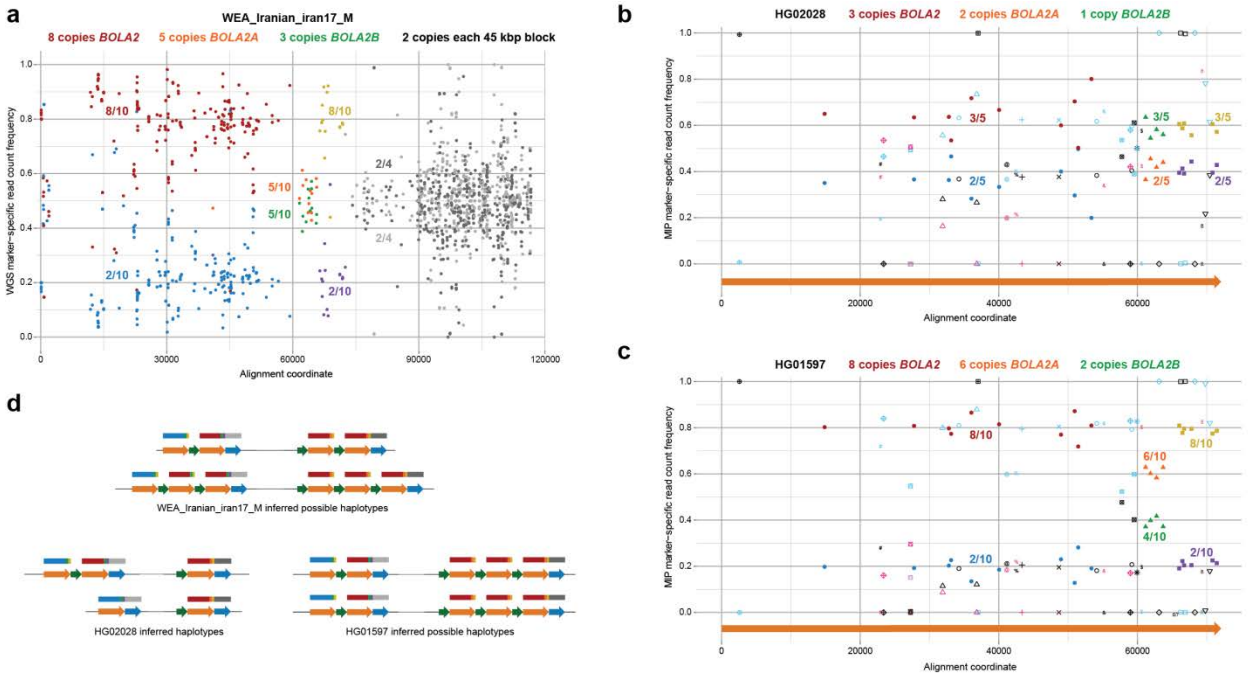


Figure S26. Extreme *BOLA2* PSCN estimates in the human population. a) WGS ratio-based *BOLA2* PSCN estimate for an Iranian confirms the estimate of eight total copies based on the WGS read-depth method. See section 4.3 and **Fig. S21** for inference methodology. b) MIP sequence marker-specific read count ratios confirm three total copies of *BOLA2*, consistent with the WGS read-depth method for this Vietnamese sample. c) MIP ratio data from a different Vietnamese sample indicate this individual has a *BOLA2* copy number of eight, discordant with the WGS

aggregate read-depth estimate of nine copies. d) Schematics show inferred possible haplotypes for these three individuals based on the data in panels a-c.

In summary, we assessed copy number estimates at the extremes of the distribution for humans. We genotyped all individuals predicted to have three *BOLA2* copies or more than eight using our MIP approach. The maximum *BOLA2* copy number validated was eight. The majority of the low-copy estimates (15/21) were experimentally validated as four copies as opposed to three. No humans were found to have fewer than three *BOLA2* copies. Neanderthal and Denisova were found to have two *BOLA2* copies based on examination of WGS marker-specific read count frequencies (section 4.3). Thus, the conclusion that *BOLA2* is duplicated specifically in *Homo sapiens* is robust to uncertainties in human and archaic hominin *BOLA2* copy number estimates based on the WGS read-depth method.

5. Population genetic modeling

5.1 Overview

Because time is necessary for a new mutation to reach high frequency, the near fixation of *BOLA2B* in humans contrasted with its young evolutionary age suggests a rapid increase in allele frequency. We modeled various evolutionary scenarios and assessed the joint probability of our observed genotype data for a single Neanderthal, a single Denisovan, 4 San individuals, and 55 Yorubans. We performed four sets of simulations, each based on the same underlying model of human demographic history [36]: i) neutral evolution without specifying *BOLA2B* age, ii) neutral evolution assuming *BOLA2B* formed 282 kya (our point-estimate for *BOLA2B* age based on phylogenetic analysis), iii) neutral evolution assuming *BOLA2B* formed at a specified age varied from 200 kya to 2 mya, and iv) evolution under positive selection assuming *BOLA2B* formed 282 kya. All scripts used to perform these simulations are freely available via GitHub at https://github.com/Schraiber/BOLA2. These analyses suggest *BOLA2B* likely did not rise to high frequency in *Homo sapiens* under a simple model of neutrality.

5.2 Coalescent simulations

We adapted a published demographic model [36] for the *Homo* lineage to simulate the evolution of *BOLA2B*. Our adaptation includes the Neanderthal and Denisova species as well as the Yoruban and San human populations. We modeled the duplication that formed *BOLA2B* as a point mutation that occurred once in history with no recurrent mutation or reversion, considering individuals having two or more *BOLA2B* copies as homozygous for the derived state, individuals having a single copy as heterozygous, and individuals lacking *BOLA2B* as homozygous for the ancestral state. Although it is possible that individuals having two *BOLA2B* copies could be effectively heterozygous by having both copies on the same homolog, never observing a human lacking *BOLA2B* and only rarely observing individuals with four *BOLA2B* copies suggests this possibility is remote. Under these assumptions, our PSCN genotype data indicate that 53 Yorubans are homozygotes for the derived state, 2 Yorubans are heterozygotes, 4 San individuals are homozygotes for the derived state, and both Neanderthal and Denisova are homozygotes for the ancestral state.

In our first set of simulations, we used the coalescent simulator *ms* [37] to assess the joint probability of observing *BOLA2B* on at least 108 of 110 sampled Yoruban haplotypes and 8 of 8 San haplotypes, conditional on its presence in at least one sampled human haplotype and its absence in archaic hominins. We performed 793,683 simulations meeting the requirements and generated a heat map (**Fig. S27**) to show how often each possible genotype outcome occurred in simulated data, with each genotype outcome defined by the numbers of simulated Yoruban and San haplotypes having *BOLA2B*. Under this scenario modeling neutral evolution without accounting for *BOLA2B* age, we find the observed genotype data

improbable: only 12,232 simulations yielded *BOLA2B* allele frequencies as high as or higher than those inferred from our genotyping data given the assumptions above ($p < 0.012$).

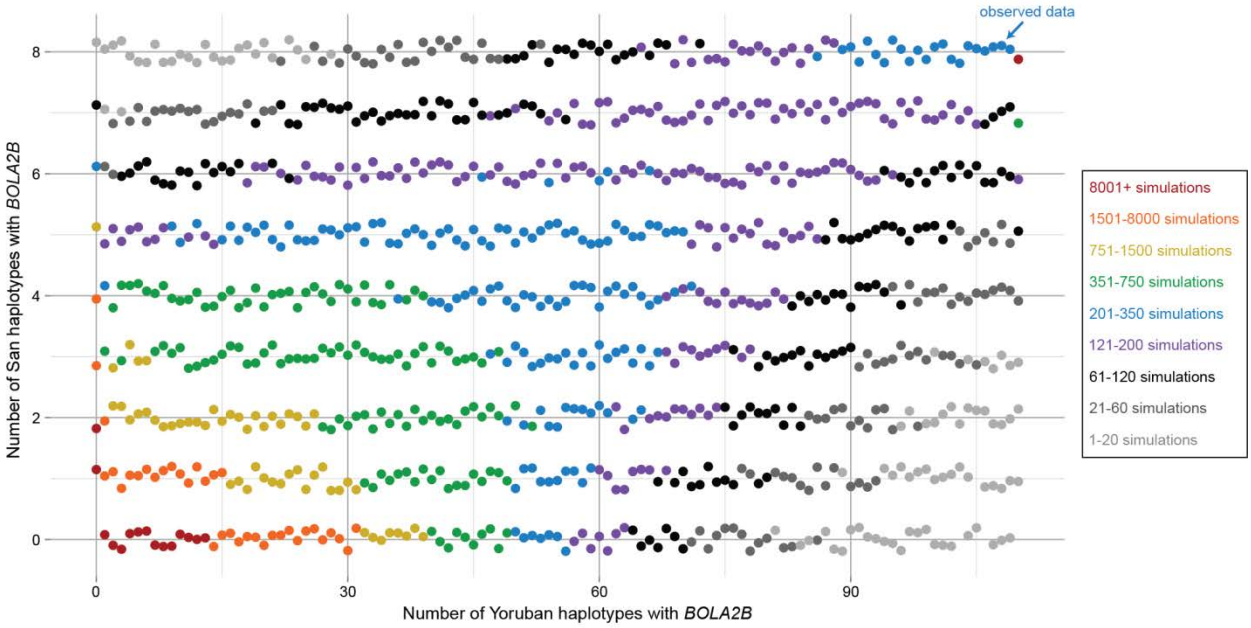


Figure S27. Frequencies of different genotype outcomes from simulations of *BOLA2B* evolution under neutrality. Each point shows a genotype outcome that was attained in at least one simulation (n = 793,683), with colors indicating how frequently each genotype outcome occurred in simulated data.

5.3 Assessing different evolutionary ages of *BOLA2B*

Our initial simulations did not make use of additional information we have regarding the evolution of *BOLA2B*, namely, its age estimated from our phylogenetic analyses. To condition on *BOLA2B* age, we performed simulations with *msms* [38], assuming that *BOLA2B* originated 282 kya and was not subject to selection. In no simulation out of 1,000,000 did *BOLA2B* reach even close to the observed frequencies in Yoruban and San populations ($p = 0$, **Fig. 4b**). In most simulations (999,807), *BOLA2B* was lost from the human species. Because these simulations are not conditioned on *BOLA2B* being observed in at least one individual, the p-value corresponding to these simulations is not directly comparable to the p-value for the above scenario disregarding *BOLA2B* age. Nevertheless, the fact that the highest simulated *BOLA2B* frequencies were not close to the observed frequencies argues that *BOLA2B* rising to high frequency in *Homo sapiens* within the last 282,000 years is very unlikely under neutrality.

To account for uncertainty surrounding the age of *BOLA2B*, we asked how old it would have to be to have risen to high frequency exclusively in *Homo sapiens* under neutral evolution. We again employed *msms* [38] and performed simulations excluding the possibility of selection, exploring a range of possible *BOLA2B* ages. We varied *BOLA2B* age from 200,000 years old to 2 million years old, performing 1,000,000 simulations for each age in this range at all 25,000 year increments. We calculated the relative likelihood of our genotype data for each age value as the number of simulations with that particular age value yielding the observed genotype data divided by the maximum number of simulations yielding the observed data for any single age value, considering all age values examined. Regardless of the age we used, the observed genotype data were always unlikely—for younger ages, there was not sufficient time for *BOLA2B* to rise to high frequency under neutrality, while for older ages, several simulations yielded high *BOLA2B* frequency in humans but also *BOLA2B* presence in Neanderthal and Denisova. The most

likely ages for *BOLA2B* based on simulations alone, assuming neutral evolution, were 1.5 million years old and 1.7 million years old (**Fig. S28**), more than five times older than our age estimate determined from phylogenetic analysis. The noisiness of the simulations is apparent from age values relatively close to one another often having substantially different relative likelihoods, underscoring the improbability of the observed *BOLAB* genotype pattern in any scenario driven by neutral evolution. Importantly, no simulations assuming a *BOLA2B* age within the 95% confidence interval for our phylogenetic age estimate produced the genotypes observed in Yoruban and San populations. Together, these results support our conclusion of non-neutral evolution for *BOLA2B*—a conclusion that is not dependent on the validity and precision of our estimate for the age of *BOLA2B*.

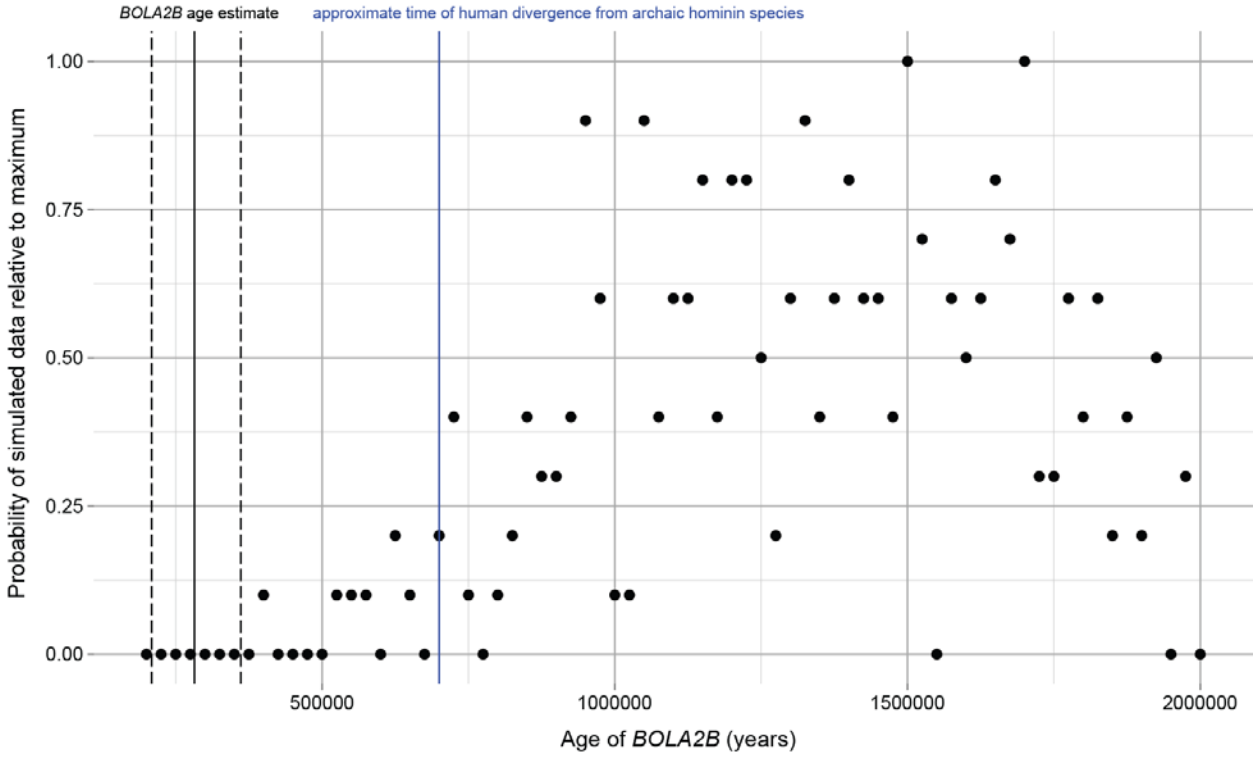


Figure S28. Modeling *BOLA2B* age variation. Simulations were performed with *msms* [38], and the time of the duplication that formed *BOLA2B* was varied from 200 kya to 2 mya in increments of 25,000 years. The probability of the observed data under each scenario was calculated from 1,000,000 simulations, scaled, and plotted. Vertical lines indicate the estimated time of the duplication that formed *BOLA2B* and the 95% confidence interval around this estimate (solid and dashed black lines, respectively), as well as the approximate time of human divergence from archaic hominin species [28] (blue line).

5.4 Estimating positive selection

Given that the observed high frequencies of *BOLA2B* were improbable under neutrality (especially when age is taken into account), we asked if the data were better explained by a model where *BOLA2B* was driven to high frequency by positive selection. We again used msms [38] and incorporated *BOLA2B* age (282 kya), this time assuming a model of genic selection, with *BOLA2B* homozygotes assigned a relative fitness of $(1 + s)$, heterozygotes assigned a relative fitness of $(1 + s/2)$, and homozygotes lacking *BOLA2B* assigned a relative fitness of 1, where s is the selection coefficient. We initially explored a wide range of values for s using a small number of simulations to get a sense of selection strengths most consistent with the data. Eventually, we settled on a narrow range from $s = 0.0009$ to $s = 0.0024$ where we performed 1,000,000 simulations for each value of s and varied s by increments of 0.0001. We calculated the relative likelihood of the observed San and Yoruban *BOLA2B* genotypes for each value of s as the number of

simulations at that particular *s* value yielding the observed genotype data divided by the maximum number of simulations yielding the observed data for any single value of *s*, considering all *s* values examined. We thus obtained a maximum likelihood estimate of the selection coefficient: $s = 0.0016$ (**Fig. 4c**). Importantly, this is an estimate for the net value of *s*, (i.e., the sum of advantageous and deleterious effects). The positive effect of the *BOLA2B* duplication must be even higher than suggested by this *s* value because it also has adverse consequences, conferring susceptibility to microdeletions and microduplications associated with disease.

6. *BOLA2* mRNA and protein characterization and expression

6.1 *BOLA2* RNA expression in human tissues and the discovery of *Homo sapiens*-specific fusion transcripts

We searched for transcripts mapping to *BOLA2* loci by analyzing annotated mRNA and expressed sequence tags (UCSC Genome Browser using the reference genome GRCh37). We identified two distinct sets of transcripts: a set consistent with the canonical *BOLA2* model and a set suggesting of fusion transcripts between *BOLA2* and *SMG1P* (**Fig. S29**). We designed two PCR assays to amplify both canonical and fusion *BOLA2* transcripts (**Fig. S29**). Oligonucleotides for PCR amplification are provided below:

BOLA2_forward: TAGAGCAGGTAGACGCCGAAA
BOLA2_reverse: AATTTAATGGCTGTGCAGATCCC
BOLA2_fusion_forward: GAACAAGCTCTCGGGGACTATC
BOLA2_fusion_reverse: GTGATTCTGCAGACATGTTGACA

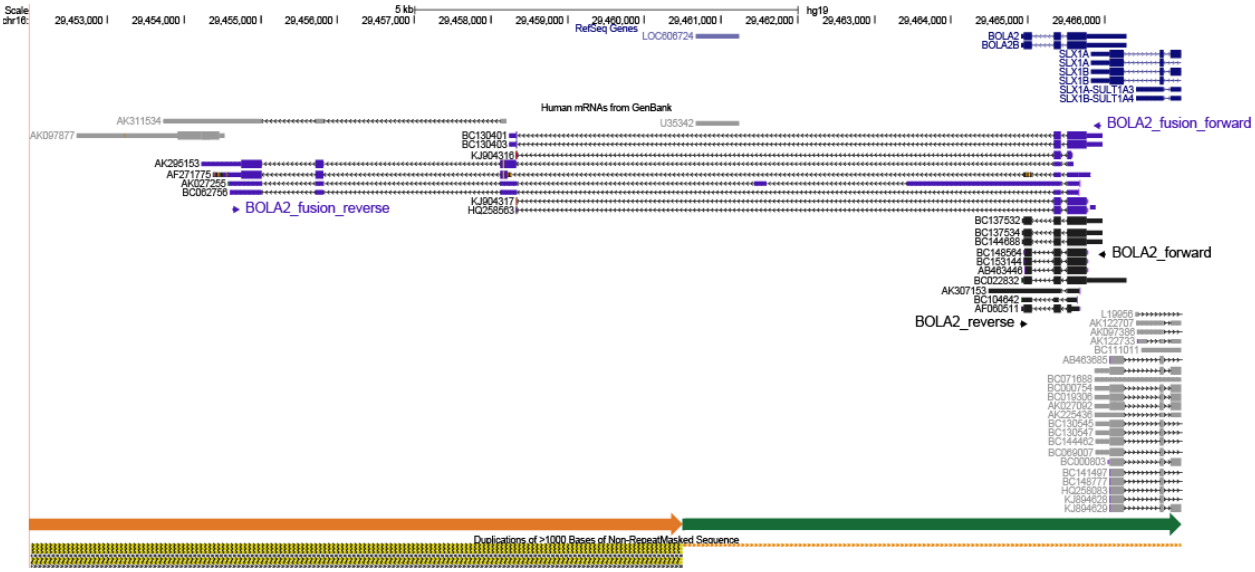


Figure S29. *BOLA2* transcript models. Two classes of *BOLA2* mRNA transcripts are apparent: those mapping to the canonical *BOLA2* model (black) and those spanning the junction between *BOLA2* and *SMG1P* (purple). Arrows indicate locations of primers that were designed to amplify *BOLA2* transcripts.

We prepared cDNA from total RNA (Clontech) from fetal brain (obtained from spontaneously aborted fetuses, ages 20-33 weeks), brain, cerebellum, heart, skeletal muscle, spleen, testis, and thymus using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) and oligo-dT primers, following the

manufacturer's instructions (section 9.2). We then performed long-range PCR with each primer set on cDNA from each tissue using the Expand Long Template PCR System (Roche). All reactions yielded products of the expected sizes (**Figs. S30-S31**) except for the reaction including skeletal muscle cDNA and fusion *BOLA2* primers, indicating that canonical and fusion *BOLA2* isoforms are widely expressed.

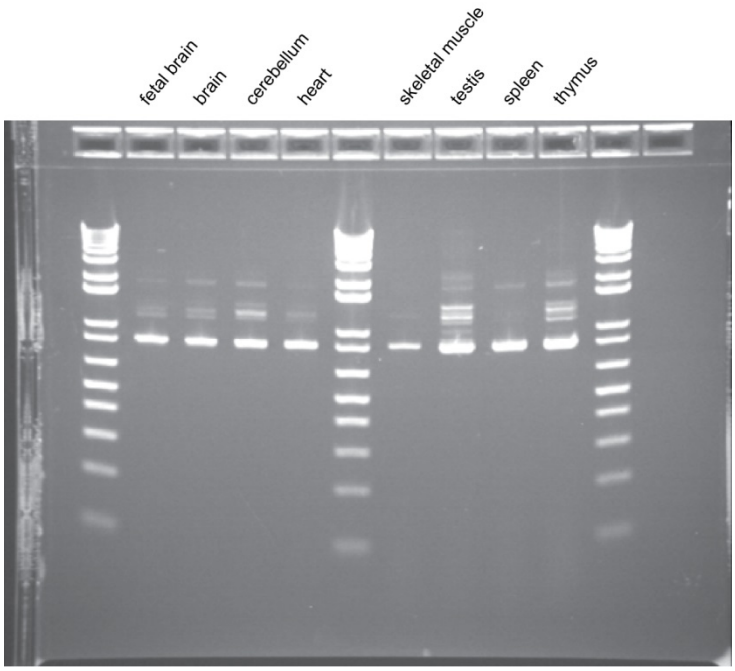


Figure S30. RT-PCR expression profile for canonical *BOLA2*. The expected product size for canonical *BOLA2* (838 bp) was observed in all eight human tissues. 1 kb + DNA ladder (Thermo Fisher).

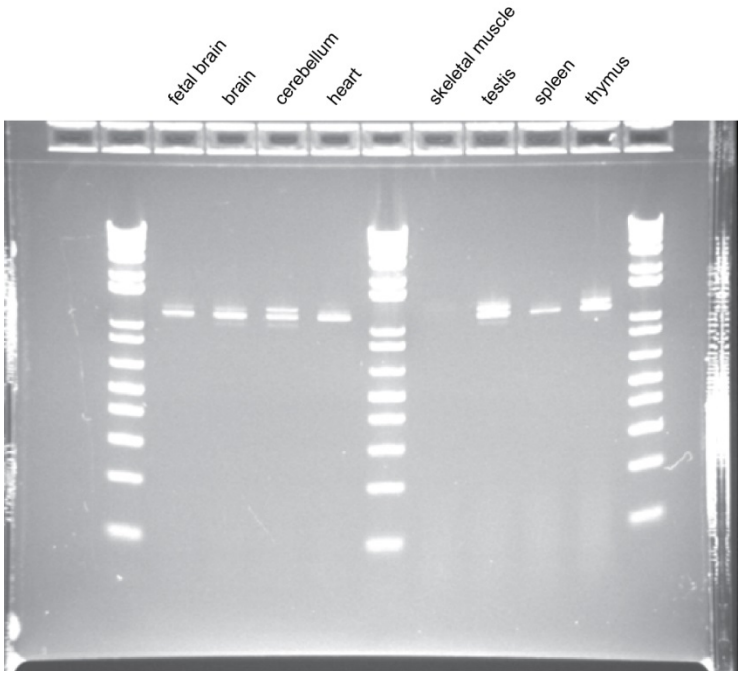


Figure S31. RT-PCR expression profile for *BOLA2-SMG1* fusion product. The expected product size for the *BOLA2* fusion transcript (1,239 bp) was observed as a doublet in all tissues except skeletal muscle. Intensity of upper band differs between tissues. 1 kb + DNA ladder (Thermo Fisher).

We cloned and sequenced PCR products from reactions including brain cDNA. Specifically, we cloned products from both canonical and fusion *BOLA2* reactions into pCR Blunt II TOPO using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher), following the manufacturer's instructions. We sequenced five clones from each reaction using capillary sequencing and M13 primers provided with the cloning kit. All sequenced products derived from *BOLA2* transcripts. The five products from the canonical *BOLA2* reaction matched the predicted *BOLA2* exon structure, including an open reading frame (ORF) spanning 456 bp predicted to encode a 152 amino acid (aa) protein. Sequenced products from the fusion transcript reveal two isoforms both including the first two exons from canonical *BOLA2*, skipping the final canonical exon, and including additional exons from *SMG1P* (**Fig. 4.5c**). One of these fusion isoforms (two sequenced products) terminates the *BOLA2* ORF shortly after the fusion junction (*BOLA2T*), such that if translated, only 5 aa would derive from SMG1P. Intriguingly, the other fusion isoform (three sequenced products) maintains the *BOLA2* ORF across the junction (*BOLA2F*), putatively encoding a predicted protein with 164 amino acids from SMG1P. Note that *SMG1P* is a paralog of *SMG1*—a phosphatidylinositol 3-kinase-related kinase involved in nonsense mediated decay and thought to be important for early embryogenesis [39].

To validate these transcripts and assess their expression in a wider panel of human tissues, we analyzed RNA-seq data from GTEx [40]. Specifically, we quantified expression for all three *BOLA2* transcripts in human tissues by running Sailfish [41] (version 0.6.3) with the default parameters and k = 20 on each of the GTEx RNA-seq samples (dbGaP version phs000424.v3.p1). We mapped reads to all GRCh38/hg38 RefSeq transcripts, replacing RefSeq transcripts for *BOLA2* with our three *BOLA2* transcript models, and quantified expression in units of transcripts per million (TPM, **Fig. S32**). These results corroborate our RT-PCR data and confirm widespread mRNA expression for all *BOLA2* isoforms in human tissues including muscle.

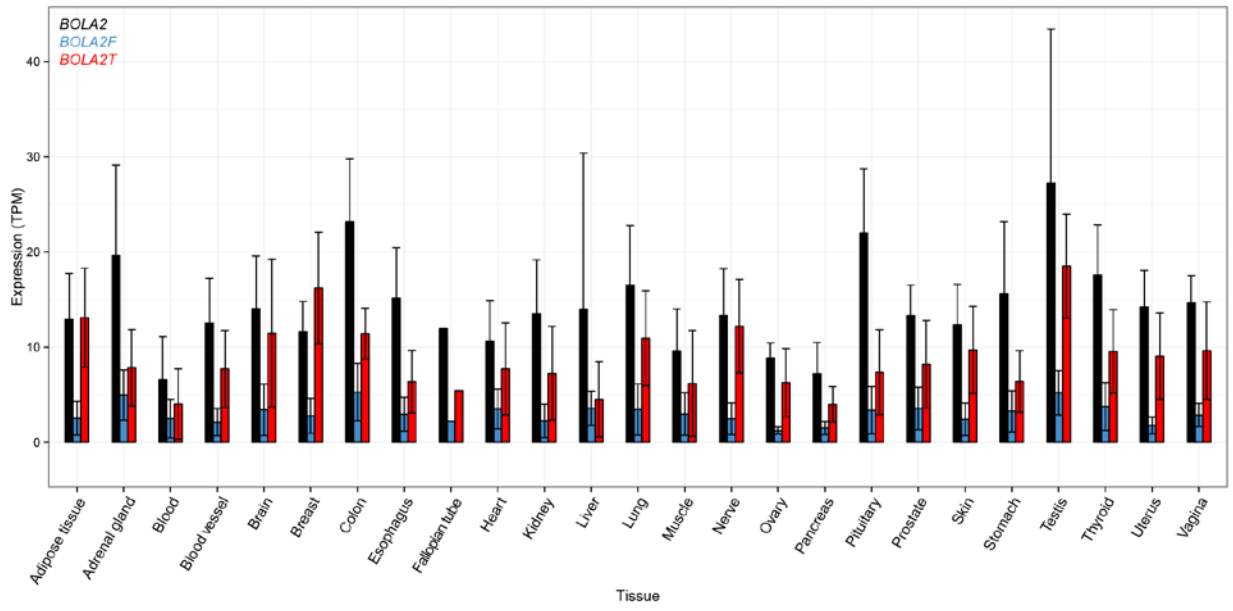


Figure S32. *BOLA2* RNA-seq expression analysis. Canonical (*BOLA2*) and fusion transcripts (*BOLA2F, BOLA2T*) were assessed across 25 humans from GTEx RNA-seq data [40]. Bar heights indicate mean expression levels for

each tissue in TPM with standard errors shown (error bars). Colors correspond to different *BOLA2* isoforms as indicated.

The genomic architecture juxtaposing *BOLA2* and *SMG1P* exists at two locations at chromosome 16p11.2: at the *BOLA2B* locus at BP4 and at junctions between tandem 102 kbp duplications at BP5. In both cases, this architecture evolved as a result of duplication events that occurred specifically in *Homo sapiens*. Thus, we predict that fusion *BOLA2* transcripts are *Homo sapiens*-specific. To test this hypothesis, we analyzed RNA-seq data from induced pluripotent stem cells (iPSCs) from human, chimpanzee, and bonobo [42]. We leveraged unique k-mers [31] distinguishing the exon 2-3 junction, which differs between each *BOLA2* isoform (**Fig. 4.5c**). For each isoform, we generated a list of all distinguishing 70-mers over the exon 2-3 junction including at least 15 bp from each exon. We quantified the relative expression of *BOLA2* isoforms in iPSCs from different species (**Fig. S33**) by assessing the reads assigned to each *BOLA2* isoform for each RNA-seq dataset.

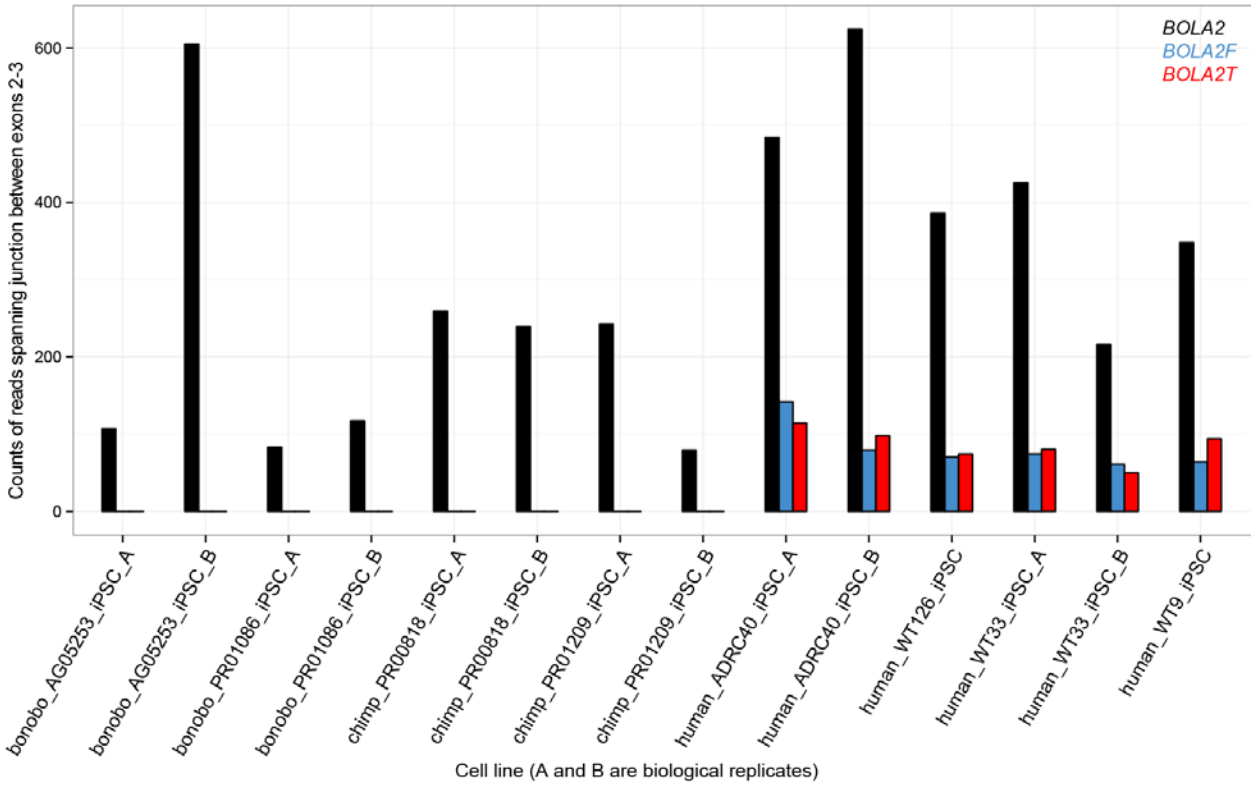
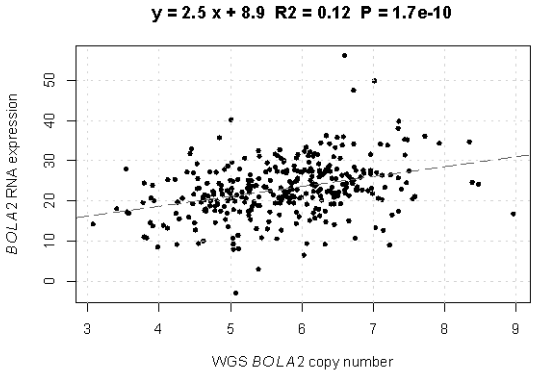


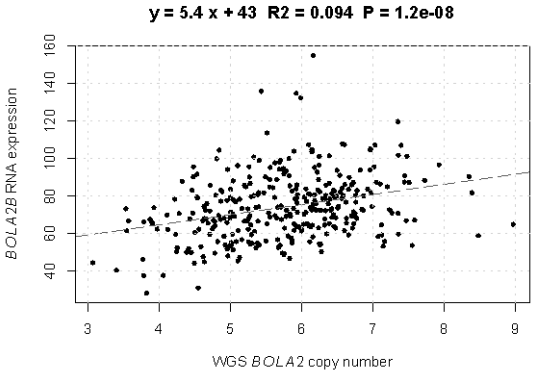
Figure S33. Expression of different *BOLA2* isoforms in iPSCs from human, chimpanzee, and bonobo. Bar heights indicate counts of reads containing a 70-mer perfectly matching exon 2-3 junction sequence corresponding to a particular *BOLA2* isoform. Colors correspond to different *BOLA2* isoforms as defined above.

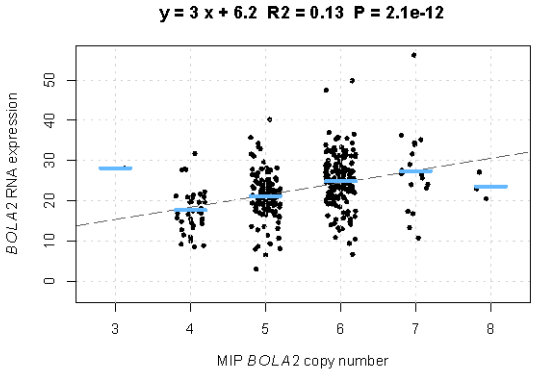
We did not observe fusion isoforms in chimpanzee or bonobo, confirming the *Homo sapiens*-specific nature of fusion transcripts. No substitutions exist between human and chimpanzee or between human and bonobo over the last 55 bp of exon 2, so this result is not an artifact of generating our lists of 70-mers using human transcript sequence data. Read counts are not normalized, so comparisons between experiments only shed light on the presence or absence of different isoforms in different cell lines. However, read counts for different isoforms within each experiment provide insight into the relative expression of different isoforms. These data suggest that the expression of both fusion isoforms in human is approximately equal but ~3- to 10-fold lower than the expression of the canonical *BOLA2* transcript.

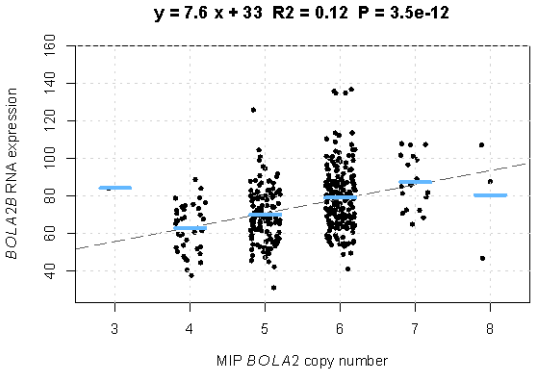
6.2 Correlation of *BOLA2* copy number with *BOLA2* RNA expression

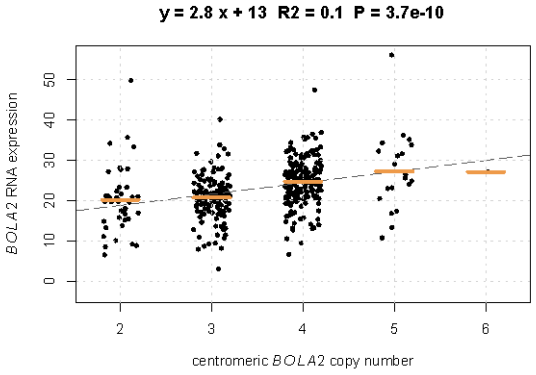
We assessed whether the copy number of *BOLA2* is correlated with its expression at the mRNA level. We used transcripts designated as *BOLA2* (ENSG00000183336.7) and *BOLA2B* (ENSG00000169627.7) and expression quantifications [43] (PEER-normalized RPKM) from lymphoblastoid cell lines (LCLs) derived from 366 individuals having *BOLA2* copy number estimates (**Table S10**). Note that for this section and section 6.3, *BOLA2B* refers to the specific transcript above and not to the *BOLA2* paralog at BP4. We computed Pearson's correlations between copy number and expression values using the R software environment [44]. We found significant correlations between *BOLA2* copy number and both *BOLA2* mRNA expression and *BOLA2B* mRNA expression, regardless of which method was used for genotyping copy number (**Fig. S34**; for 330 WGS read-depth-based estimates, R = 0.34, p = 1.7e-10 and R = 0.31, p = 1.2e-08, for *BOLA2* and *BOLA2B*, respectively; for 366 MIP-based estimates, R = 0.36, p = 2.1e-12 and R = 0.35, p = 3.5e-12, for *BOLA2* and *BOLA2B*, respectively). Better correlations were observed with MIP-based copy number estimates, consistent with their generally higher accuracy compared to WGS read-depth-based estimates likely due to low sequencing coverage for 1000 Genomes Project samples [24] (section 4.2).

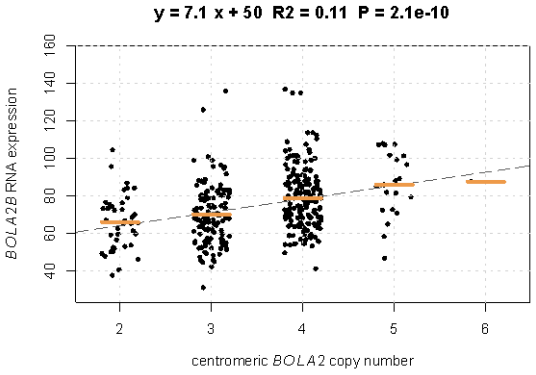

y = 2.5 x + 8.9 R2 = 0.12 P = 1.7e-10
BOLA2 RNA expression
WGS BOLA2 copy number

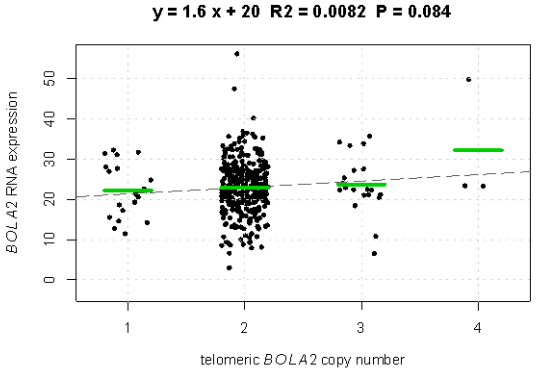

y = 5.4 x + 43 R2 = 0.094 P = 1.2e-08
BOLA2B RNA expression
WGS BOLA2 copy number


y = 3 x + 6.2 R2 = 0.13 P = 2.1e-12
BOLA2 RNA expression
MIP BOLA2 copy number


y = 7.6 x + 33 R2 = 0.12 P = 3.5e-12
BOLA2B RNA expression
MIP BOLA2 copy number


y = 2.8 x + 13 R2 = 0.1 P = 3.7e-10
BOLA2 RNA expression
centromeric BOLA2 copy number


y = 7.1 x + 50 R2 = 0.11 P = 2.1e-10
BOLA2B RNA expression
centromeric BOLA2 copy number


y = 1.6 x + 20 R2 = 0.0082 P = 0.084
BOLA2 RNA expression
telomeric BOLA2 copy number

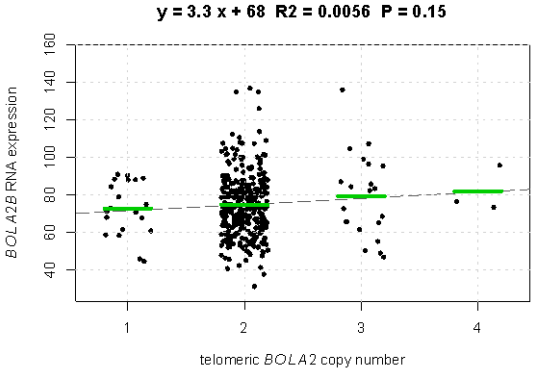

y = 3.3 x + 68 R2 = 0.0056 P = 0.15
BOLA2B RNA expression
telomeric BOLA2 copy number

Figure S34. *BOLA2* and *BOLA2B* RNA expression in LCLs correlates with genomic *BOLA2* copy number. The top four panels show *BOLA2* (left) and *BOLA2B* (right) RNA expression (Geuvadis dataset [43], PEER-normalized RPKM values, y-axis) correlates with *BOLA2* copy number (x-axis) estimated by WGS read depth (N = 330) and our MIP assay (N = 366, copy number values are jittered around their integer values for clarity). *BOLA2* copy number explains >10% of the variability in *BOLA2* RNA expression. The bottom four panels show *BOLA2* (left) and *BOLA2B* (right) RNA expression as a function of *BOLA2* copy number at BP5 and BP4. The majority of the RNA expression variation is driven by BP5 *BOLA2* copy number, reflecting the higher copy number variation at this locus. Regression lines are shown as dashed lines. Colored horizontal lines represent the mean expression value of samples grouped by copy number. Regression line equations, R-squared values, and p-values are shown on top of each plot.

We next assessed whether the copy number change at the *BOLA2* ancestral centromeric locus (BP5) and at the human-specific telomeric locus (BP4) differently contribute to the expression variation. Using the R software environment [44], we compared two linear models that describe *BOLA2* and *BOLA2B* expression: i) accounting only for total *BOLA2* copy number (MIP-based estimates) and ii) accounting for the centromeric (BP5) and telomeric (BP4) *BOLA2* copy numbers (**Table S11**). The models suggest that the overall *BOLA2* copy number drives the variation of *BOLA2*/*2B* RNA expression and both the BP5 and BP4 copy numbers have an effect with a similar magnitude. As noted, the BP5 copy number variable is more statistically robust, likely because BP5 shows greater copy number polymorphism. We compared *BOLA2* copy number and expression within human population groups to control for potential differences in genetic background (**Fig. S35**). The correlation is generally consistent in the different populations, with some, like the Toscani and British showing higher correlations than others such as Yorubans.

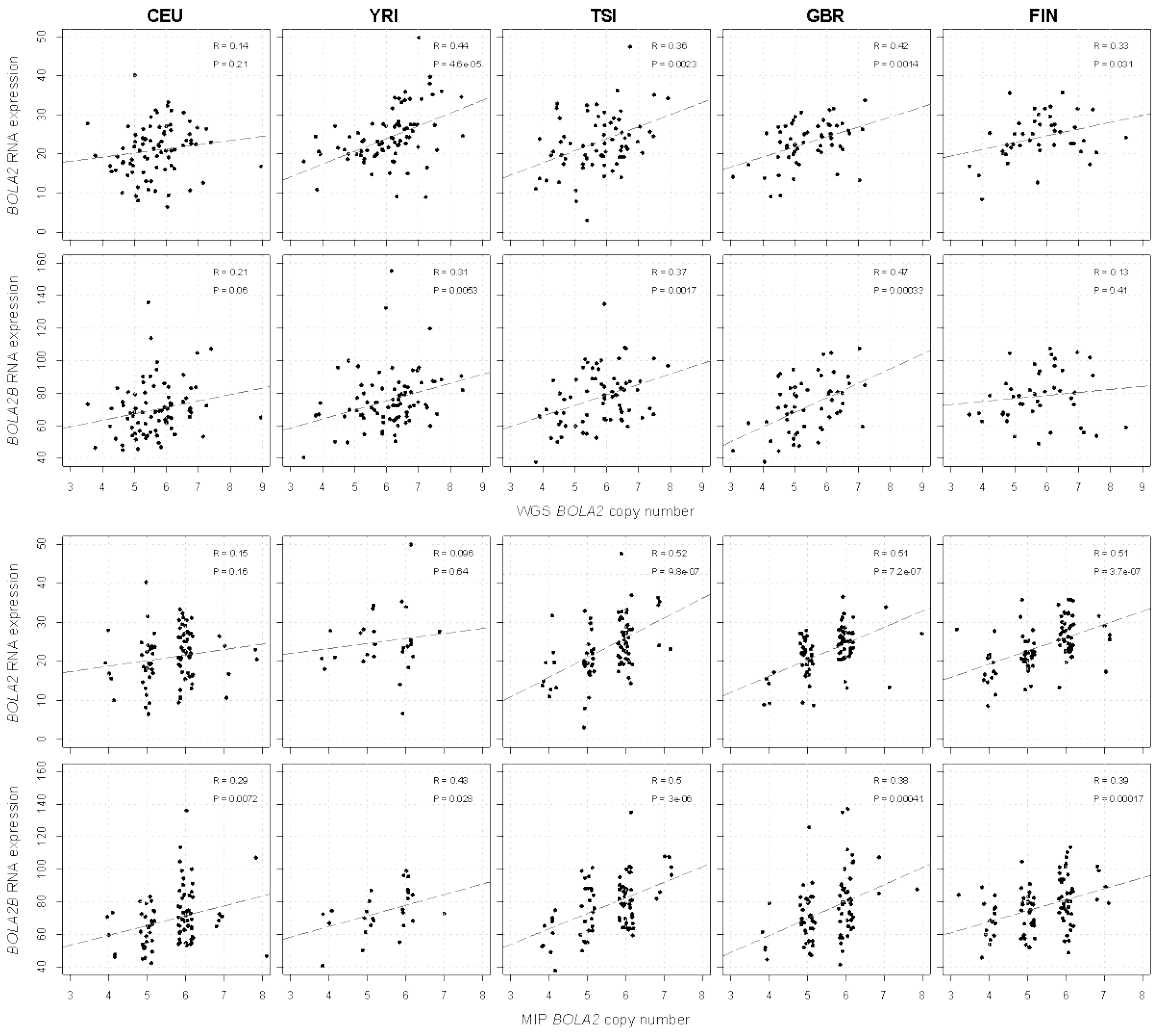


Figure S35. *BOLA2* and *BOLA2B* RNA expression in LCLs correlates with genomic *BOLA2* copy number in all populations examined. *BOLA2* and *BOLA2B* RNA expression correlations with *BOLA2* copy number estimates based on WGS read depth (top panels) and our MIP assay (bottom panels) in the populations CEU (N = 80 and N = 87, respectively), YRI (N = 80 and N = 26), TSI (N = 71 and N = 79), GBR (N = 55 and N = 84), and FIN (N = 44 and N = 90). Regression lines are shown as dashed lines. CEU = Centre d'Etude du Polymorphisme Humain collection (European); FIN = Finnish from Finland; GBR = British from England and Scotland; TSI = Toscani in Italia; YRI = Yoruba from Ibadan, Nigeria.

6.3 Genome-wide correlation of *BOLA2* copy number with gene expression

We also assessed if the expression of other genes besides *BOLA2* and *BOLA2B* correlated with *BOLA2* WGS read-depth-based copy number in LCLs. We computed Pearson's correlations between *BOLA2* copy number and gene expression (Geuvadis dataset [43]) and corrected p-values for multiple testing using the Benjamini-Hochberg method. At a false discovery rate (FDR) <0.05, we identified *BOLA2*, *BOLA2B*, *SLX1A*, *SLX1B*, *SLX1A-SULT1A3*, *SULT1A1*, and *SULT1A2* as having RNA expression correlated with *BOLA2* copy number (**Fig. S36**). Notably, these genes either map to the 102 kbp copy number variant segment at BP4 and BP5 or have paralogs mapping within this unit. This result is consistent with the 102 kbp unit of copy number variation described above (section 2.1) and indicates that the copy number change affects the expression level of genes embedded in the variable segment but does not more broadly perturb the transcriptome, at least in LCLs. We found no correlation between *BOLA2*

copy number and the expression of the 29 genes mapping to the 550 kbp critical region [45]. Similar results were obtained using the MIP-based copy number estimates (data not shown).

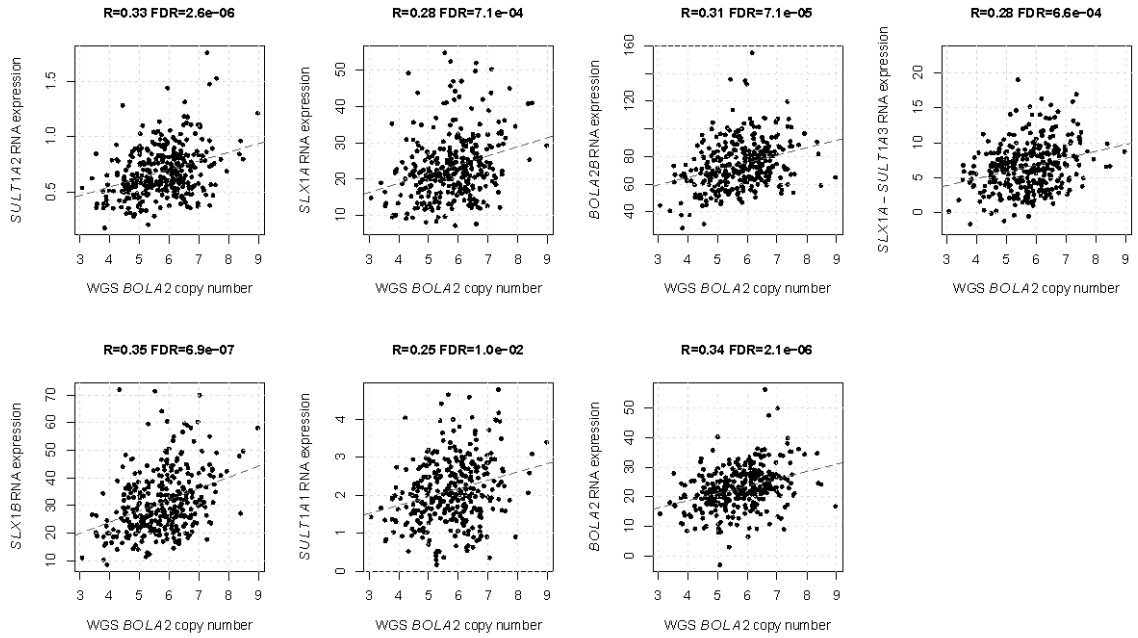


Figure S36. Genes having significantly correlated RNA expression with *BOLA2* copy number in LCLs. *BOLA2* copy number correlates with RNA expression of *SULT1A2*, *SLX1A*, *BOLA2B*, *SLX1A-SULT1A3*, *SLX1B*, *SULT1A1* and *BOLA2*. All these genes either map within the copy number variant 102 kbp unit or are paralogous to genes in this unit. Regression lines are shown as dashed lines.

6.4 *BOLA2* protein definition and anti-BOLA2 antibody validation

As we found correlation between *BOLA2* copy number and expression at the RNA level, we explored if the former also correlated with protein level changes. We tested three different antibodies to detect BOLA2 protein: a goat polyclonal antibody raised against a peptide mapping near the C-terminus of human BOLA2 (Sc-163747, Santa Cruz Biotechnology), a mouse polyclonal antibody raised against the human full-length BOLA2 (amino acids 1-152, ab169481, Abcam), and a rabbit polyclonal antibody raised against the N-terminal amino acids 1-50 of mouse BOLA2 (ab105534, Abcam). We detected a consistent band at a molecular weight of 10 kDa with the three different antibodies in human LCL whole-protein lysates. None of these three antibodies reacted to a band at the predicted molecular weight of 17 kDa.

To further assess the specificity of the antibodies, we cloned and overexpressed the coding sequence (CDS) of the annotated 17 kDa BOLA2 in HeLa cells. We detected two bands at ~10 and 17 kDa in transfected cells and did not detect the same bands in control non-transfected cells (**Fig. S37**). These results show that two forms can be translated from the *BOLA2* CDS and are recognized by the antibodies—a 17 kDa form corresponding to the entire predicted CDS and a shorter one that migrates at the size of the above-mentioned antigenic band. This indicates that the 10 kDa band observed in LCL lysate is specific. These results also demonstrate that the 10 kDa band corresponds to a shorter BOLA2 having a lower apparent molecular weight and higher mobility in the electrophoretic field, and not to the annotated 152 residue protein.

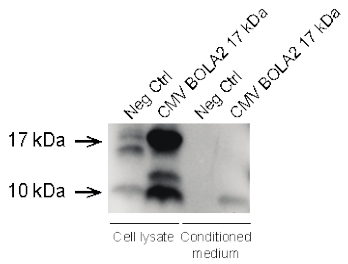


Figure S37. Western blotting of HeLa cells transfected with the human *BOLA2* annotated CDS and probed with an anti-BOLA2 antibody (Sc-163747). Both the whole-cell lysate and the conditioned medium were analyzed. Two bands with a molecular weight of 10 and 17 kDa are identified in transfected cells and correspond to two BOLA2 protein isoforms rising from different translation start sites. The detection of BOLA2 overexpressed protein in the cell medium supports existing data reporting BOLA2 is a secreted protein [46]. Negative control lanes correspond to HeLa cells not transfected with the overexpression construct.

To investigate the origin of the different proteins expressed from the *BOLA2* CDS, we predicted translation start sites (TSS) in the human *BOLA2* mRNA sequence using atgpr [47] and NetStart [48]. Both tools reported two possible TSS at position 514 and 712, corresponding to translated proteins of 152 (17 kDa) and 86 (10 kDa) residues, respectively. We retrieved available CAGE data (5' cap analysis gene expression) [49], ribosome profiling data (GWIPS-viz) [50], and mass spectrometry (MS) peptide data (PeptideAtlas) [51] for human *BOLA2* (**Fig. S38**). CAGE data showed an RNA 5' cap site in the first exon of *BOLA2*, suggesting that the annotated mRNA does not correspond to the most abundant transcript. The Ribo-seq profile of *BOLA2A* (BP5) and *BOLA2B* (BP4) showed few reads mapping to the sequence encoding for the N-terminal additional part of the hypothetical 17 kDa form. Finally, almost all peptides from MS/MS spectra mapped to the sequence common to the two forms (38 of 40 peptides), as we identified only two peptides derived from the additional N-terminal 66 residue segment of the 17 kDa protein, both from placenta. Four different peptides that begin with the start methionine of the 10 kDa BOLA2 and that could not be the products of trypsin digestion demonstrate that the methionine codon 67 (position 712) is likely used as a TSS, consistent with the predicted Kozak sequences.

Overall, these data strongly suggest that the bulk of human *BOLA2* transcription and TSS are different from current annotations. These results also demonstrate that the 10 kDa band observed in LCL lysates using the three antibodies corresponds to the 10 kDa primary protein form of BOLA2.

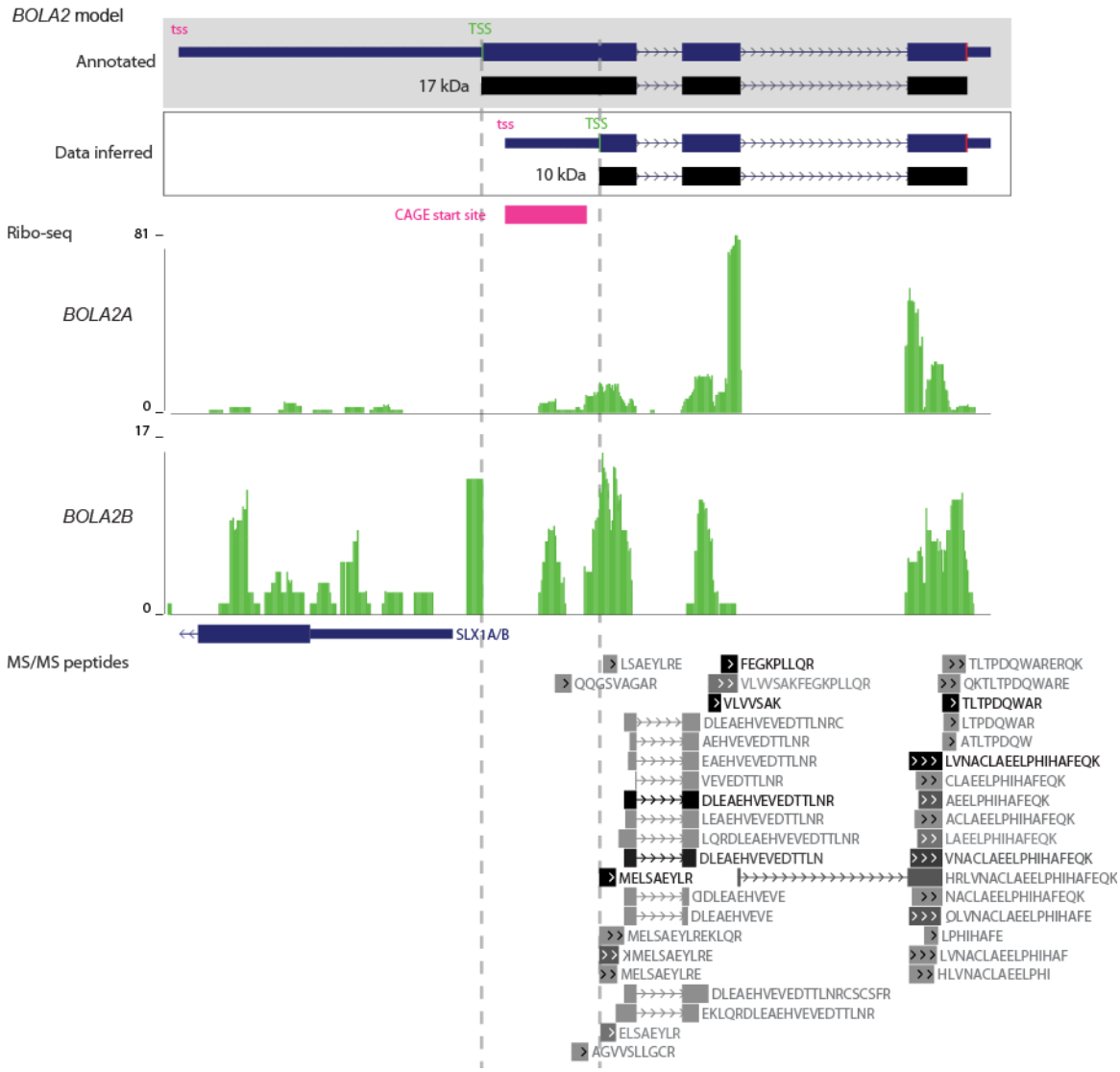


Figure S38. Annotated human *BOLA2* model and a new model inferred from available CAGE, ribosome profiling, and peptide data. The mRNA (blue) and protein (black) models of human *BOLA2* based on current annotations (top) and inferred using CAGE, ribosome profiling and MS peptide data (bottom) are shown. The main *BOLA2* CAGE start site from ENCODE/RIKEN in multiple cell lines, global aggregate Ribo-seq coverage for *BOLA2A* and *BOLA2B* (adapted from GWIPS-viz genome browser), and the peptide sequences and mappings identified from MS/MS spectra by PeptideAtlas (adapted from the UCSC Genome Browser) are shown. CAGE data indicate a different *BOLA2* transcription start site (tss) from the annotated one. Similarly, Ribo-seq and MS/MS peptide data suggest a different TSS. Note that some Ribo-seq coverage may derive from *SLX1* that overlaps the annotated *BOLA2* 5' UTR.

To further assess the possible functional relevance of the annotated 17 kDa BOLA2 form, we constructed a multiple sequence alignment including predicted BOLA2 protein sequences from human, chimpanzee, gorilla, orangutan, macaque, marmoset, mouse, rat, dog, and cow (**Fig. S39**). The first methionine of the 10 kDa form is present in all mammals, whereas the potential first methionine of the 17 kDa form is present only in primates, with the exception of orangutan.

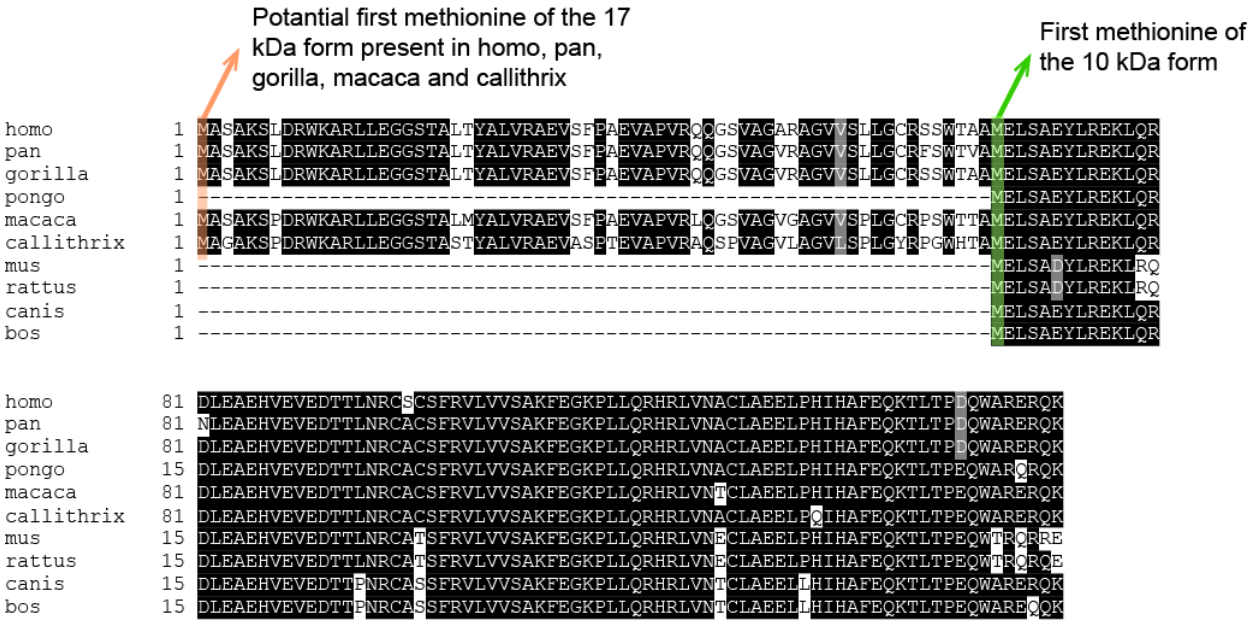


Figure S39. Multiple sequence alignment of predicted BOLA2 protein sequences of human, chimpanzee, gorilla, orangutan, marmoset, mouse, rat, dog, and cow. The two predicted first methionines are highlighted in red and green and marked with arrows.

We analyzed the evolutionary conservation of the sequence encoding for the 10 kDa protein form in primates, as well as the conservation across primates of the additional N-terminal portion of the putative 17 kDa form. We leveraged our multiple sequence protein alignment and employed a maximum likelihood framework to model different evolutionary scenarios using PAML [52]. The likelihood ratio test was used to assess the significance of different values of omega between the single parameter model (where omega is free to vary but remains constant across all branches of the phylogenetic tree) and the neutral model (where omega is set to 1 for all branches). The single parameter model provides a significantly better fit to the data than the neutral model for the 10 kDa protein sequence, suggesting it is evolving under negative constraint (omega = 0.1899, p-value = 0.007). In contrast, the additional N-terminal portion seems to be neutrally evolving, as the best single parameter model (omega = 0.8975) does not provide a significantly better fit to the data than the neutral model (p-value = 0.48). This result suggests the 17 kDa BOLA2 form is most likely evolving neutrally, and even if present probably lacks functional importance.

6.5 *BOLA2* phylogeny

We constructed multiple sequence alignments for both *BOLA2* CDS and BOLA2 protein sequence from different mammals using Clustal Omega [53] (**Fig. S40**). We leveraged the former alignment to build a maximum likelihood phylogenetic tree using MEGA6 [54] (**Fig. S41**).

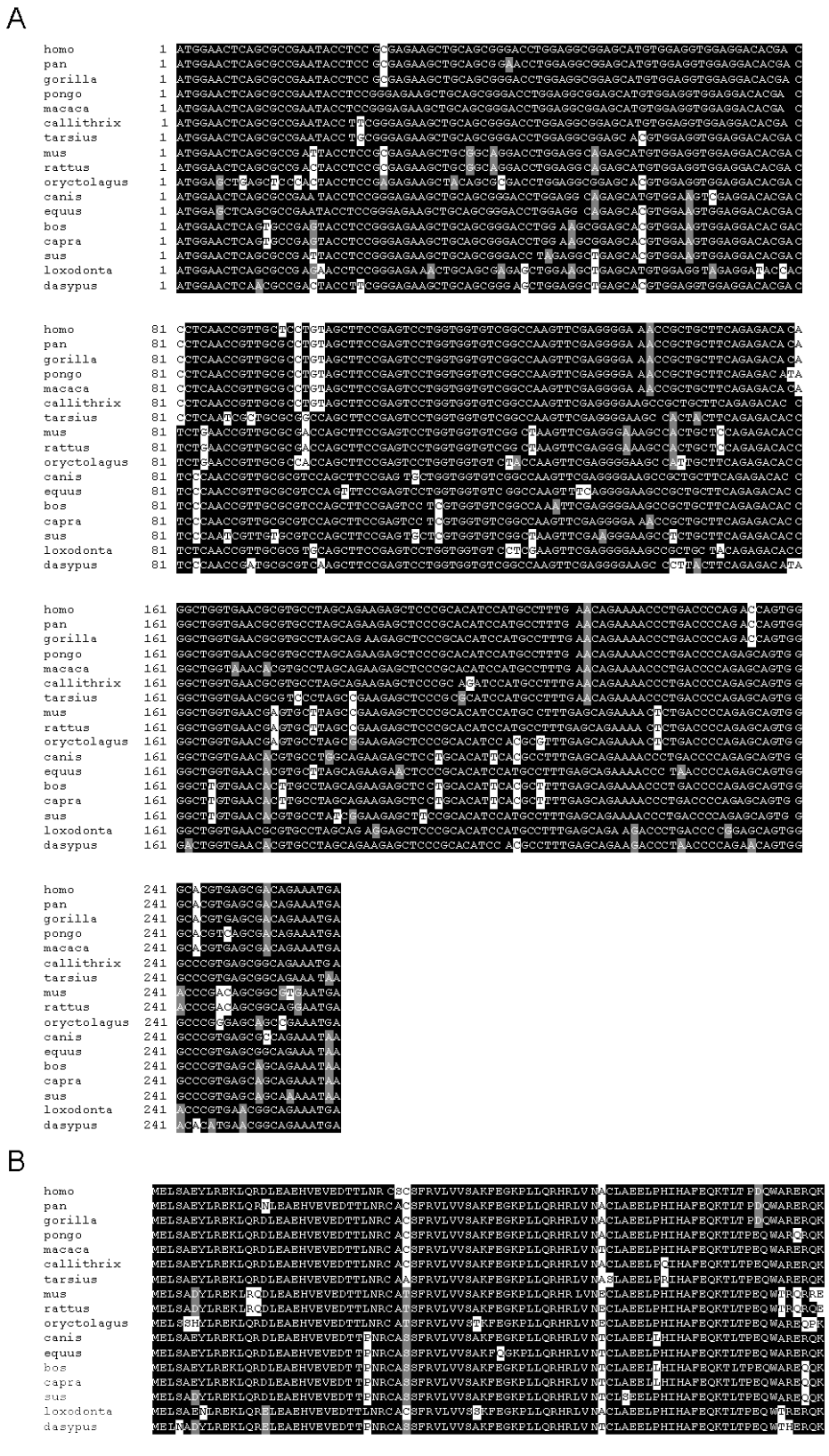


Figure S40. Multiple sequence alignment of *BOLA2* CDS (panel A) and protein sequence (panel B) from different mammals.

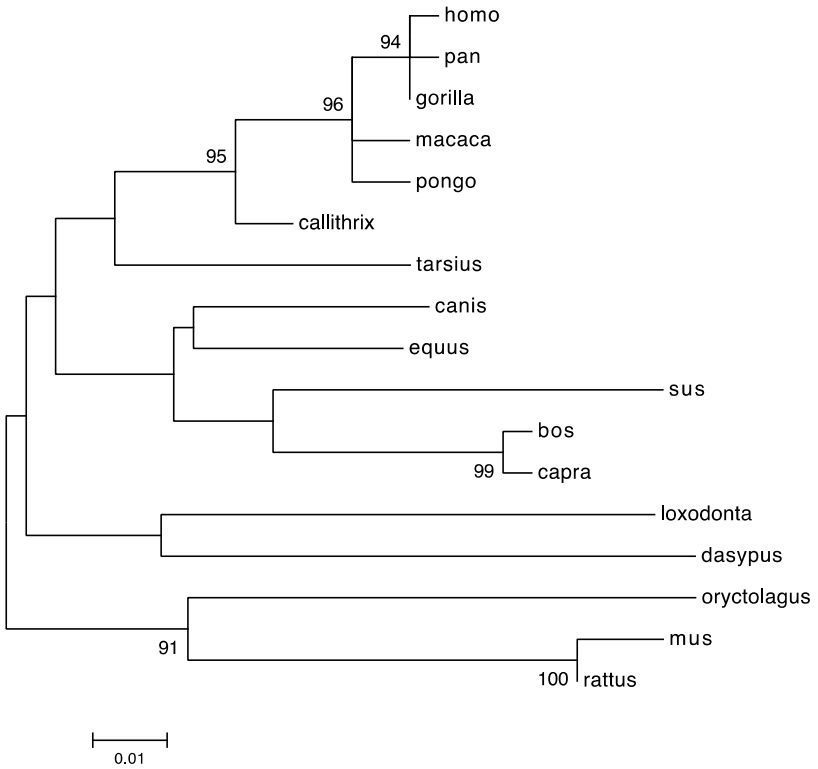


Figure S41. Maximum likelihood *BOLA2* phylogeny based on coding sequence. Maximum likelihood phylogenetic tree was constructed using *BOLA2* CDS. Tree is drawn to scale, with branch lengths measured in the number of substitutions per site and bootstrap values (1000 replicates) shown at nodes.

6.6 Correlation of *BOLA2* copy number with BOLA2 protein expression

We assessed whether *BOLA2* copy number correlates with its protein expression level (10 kDa band). We analyzed whole-protein lysates from a panel of 38 human LCLs that are part of HapMap and the 1000 Genomes Project (Coriell Institute) from individuals having BOLA2 copy number estimates (**Table S12**). All LCLs were from individuals of European ancestry, except NA18907 who is of Yoruban ancestry. The human samples, together with one chimpanzee and one orangutan LCL, were analyzed by western blotting in three parallel SDS-PAGE gels (**Fig. S42**). We quantified the BOLA2 band (antibody from Santa Cruz Biotechnology) using densitometry (Bio1D software) and normalized using actin densities. After removing the variation in BOLA2 actin-normalized densities due to the gel factor, we analyzed the correlation of BOLA2 densities with the copy number estimates (**Fig. S43**). Similar to the RNA expression, we detected significant Pearson's correlations of BOLA2 protein level with copy number (**Fig. S43**), suggesting that *BOLA2* copy number variation affects the quantity of BOLA2 protein in the cell. In chimpanzee and especially in orangutan LCL lysates, we detected a lower amount of BOLA2 compared to human cell lines, regardless of human *BOLA2* copy number (**Fig. S42**).

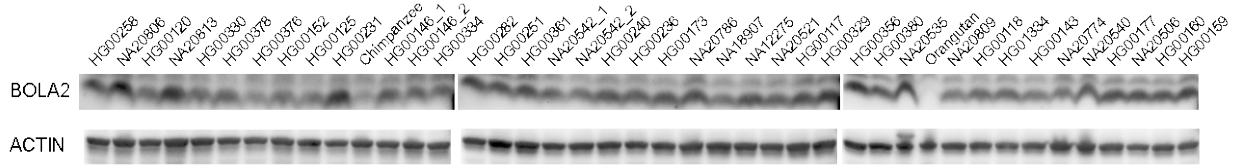


Figure S42. Western blotting of 38 human, one chimpanzee and one orangutan LCL protein lysates probed with the anti-BOLA2 (10 kDa band) and anti-actin antibodies for normalization.

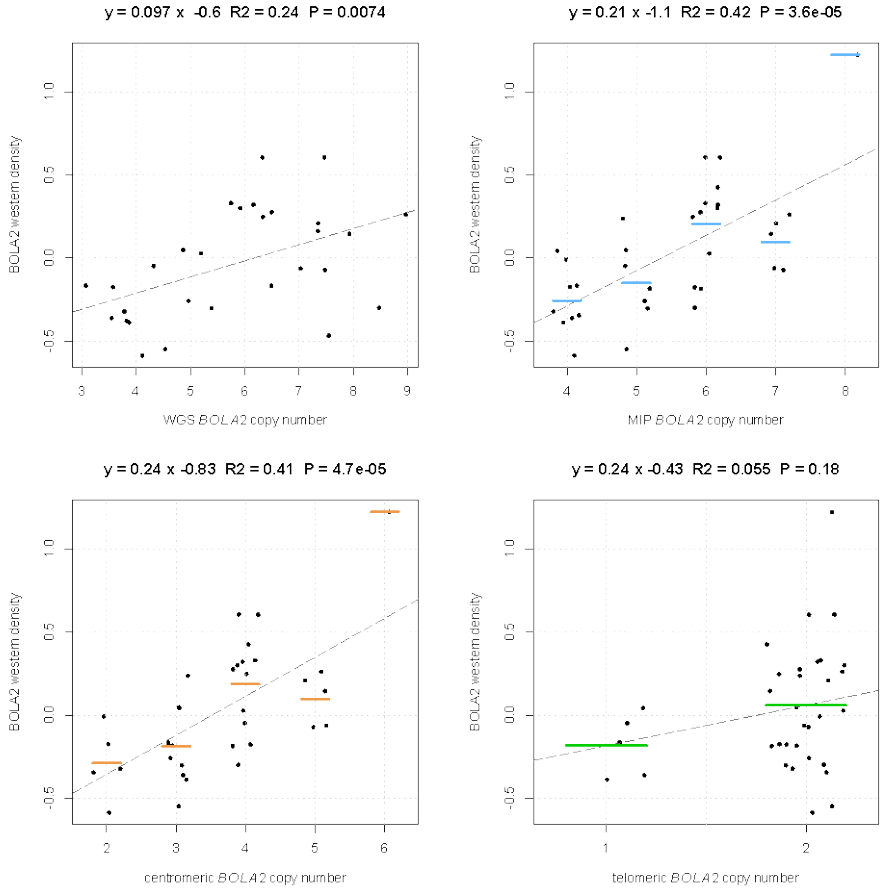


Figure S43. BOLA2 protein expression in LCLs correlates with genomic *BOLA2* copy number. Top panels show BOLA2 actin-normalized western densities correlate with *BOLA2* copy number estimated by WGS read depth (N = 29) and our MIP assay (N = 34, copy numbers jittered around integer values for clarity). *BOLA2* copy number explains >40% of variation in BOLA2 protein expression. The better correlation with MIP-based copy number estimates compared to low-coverage WGS read-depth-based ones again reflects higher accuracy for the former estimates. Bottom panels show that similar to the variation in RNA expression, the majority of protein expression variation is driven by the centromeric *BOLA2* copy number (BP5). Regression lines are shown as dashed lines. Colored horizontal lines represent the mean expression value of samples grouped by copy number. Regression line equations, R-squared values, and p-values are shown on top of each plot.

To better understand the contribution of the copy number change at the centromeric and telomeric sides to the BOLA2 protein expression level, we computed and compared four different linear models that respectively describe BOLA2 protein expression level based on: i) the total *BOLA2* copy number (MIP-based estimates); ii) the centromeric (BP5) copy number; iii) the telomeric (BP4) copy number; and iv)

both the centromeric and telomeric copy numbers as two independent variables (**Table S13**). Different from the RNA expression, the centromeric copy number variation explains 41% of the protein expression variation and the telomeric copy number does not contribute significantly to expression variation. Comparing these models with those that describe the RNA expression level, we note that the lack of contribution from the less variable telomeric (BP4) copy number might relate to a power issue, having here far fewer samples with telomeric copy number differing from 2 than in the RNA dataset.

6.7 Evidence that BOLA2 is a non-classically secreted protein that localizes to the cell cortex

The function of *BOLA2* is not well known. The *BOLA2* yeast ortholog, called *Fra2*, is involved in the negative regulation of the transcription of the iron regulon [55]. Human BOLA2 forms [2Fe-2S] bridge complexes with human glutaredoxin 3 [56], similar to what has been seen in yeast where the complex serves a role in signaling cellular iron status. Interestingly, we searched for the presence of iron responsive elements (IREs) in the 5' and 3' UTRs of human *BOLA2* (SIREs web server 2.0) and found none. BOLA2 was previously reported to be a non-classically secreted protein [46]. To further confirm this finding, we analyzed the human 10 kDa, human 17 kDa, and mouse BOLA2 protein sequences for the presence of the signal peptide (SignalP 4.1 Server) and prediction of secretion (Secretome 2.0 Server). Proteins are predicted to be non-classically secreted if their Secretome NN-score is > 0.5 and they do not contain a signal peptide. The results predicted that BOLA2 is a non-classically secreted protein since all three proteins do not have a predicted signal peptide and the NN-score is equal to 0.646 for the human 17 kDa form, 0.742 for the human 10 kDa form, and 0.829 for the mouse protein.

To experimentally test this prediction, we probed the conditioned medium of cytomegalovirus (CMV) *BOLA2* transfected HeLa cells for presence of secreted BOLA2. The detection of BOLA2 (10 kDa form) in the conditioned medium using the anti-BOLA2 antibody further suggests that BOLA2 is secreted outside of the cell (**Fig. S37**).

We then analyzed the subcellular localization of both the endogenous and transfected BOLA2 10 kDa protein in human LCL and HeLa cells, respectively. Antibody staining showed a cell cortex localization of BOLA2 in both LCL (endogenous BOLA2) and transfected HeLa cells (CMV overexpressed BOLA2) (**Fig. S44**). Of note, glutaredoxin 3 (GLRX3), an interacting partner of BOLA2 [56], was previously shown to similarly localize in the cell cortex [57].

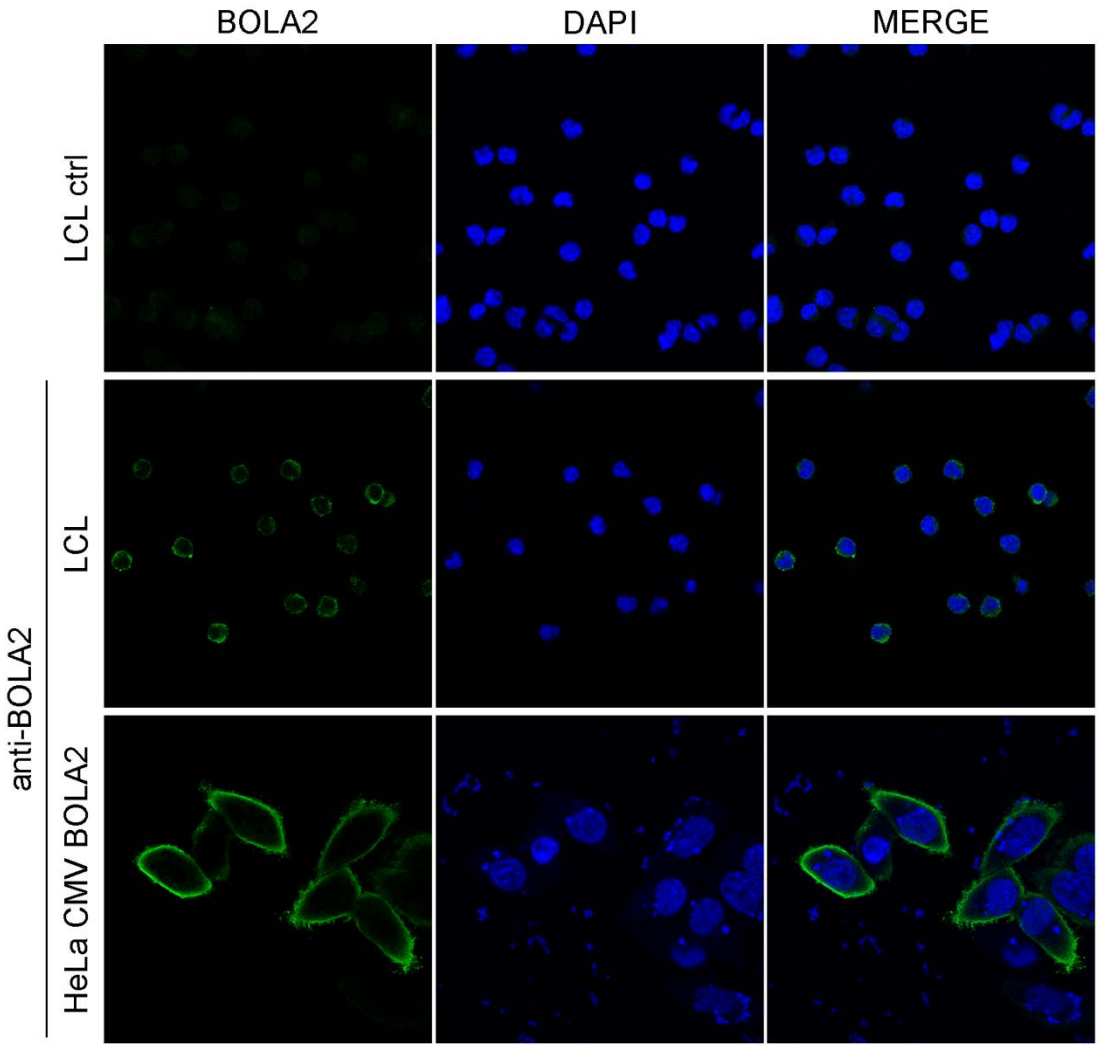


Figure S44. BOLA2 immunofluorescence in human LCL and HeLa cells transfected with CMV *BOLA2* (10 kDa). The antibody staining shows a cell cortex localization of both endogenous BOLA2 (LCLs) and CMV-expressed BOLA2 (HeLa cells). Images were taken with a Zeiss LSM710 confocal microscope with a 63x (LCL) and 100x (HeLa) objective. The negative control corresponds to LCLs with no anti-BOLA2 primary antibody incubation.

6.8 Chimpanzee and human iPSC and RNA sequencing analysis

6.8.1 Overview

Previous studies reported differences in the expression of *BOLA2* between human and nonhuman primate iPSCs [42]. In light of our new findings, we aimed to quantify the levels of different *BOLA2* isoforms in human, chimpanzee and bonobo iPSCs. We reanalyzed previously described RNA-seq data [42] generated from human embryonic stem cells (ESCs) HUES6, H1 and H9; human iPSC lines WT-33, ADRC-40, WT-126 and WT9, chimpanzee iPSC lines PR00818 and PR01209, and bonobo iPSC lines AG05253 and PR01086. Moreover, to investigate the expression levels of *BOLA2* during neuronal development, we differentiated human and chimpanzee iPSCs into neural progenitor cells (NPCs) and neurons. We designed a specific bioinformatics analysis to quantify the expression of different *BOLA2* isoforms from RNA-seq.

6.8.2 Cell lines

Human ESCs H1 and H9 are from Wisconsin (WiCell Research Institute, Inc.). According to Thomson *et al*. [58], "Fresh or frozen cleavage stage human embryos, produced by *in vitro* fertilization (IVF) for clinical purposes, were donated by individuals after informed consent and after institutional review board approval. Embryos were cultured to the blastocyst stage, 14 inner cell masses were isolated, and five ES cell lines originating from five separate embryos were derived, essentially as described for nonhuman primate ES cells." Human ESC line HUES6 is from Harvard (Harvard Stem Cell Institute) and described by Cowan *et al*. [59]. This line was obtained from frozen cleavage- and blastocyst-stage human embryos, produced by IVF for clinical purposes, after obtaining written informed consent and approval by a Harvard institutional review board. Human iPSC lines (WT-33, ADRC-40, WT-126 and WT9) and chimpanzee iPS cell lines (PR00818 and PR01209) have been previously described [42, 60].

6.8.3 Cell culture and neuronal differentiation

Established iPSC colonies were kept in feeder-free conditions and passed using mechanical dissociation. To obtain NPCs from human and chimpanzee iPSCs, embryoid bodies (EBs) were formed by mechanical dissociation of iPSC clusters and plated into low-adherence dishes in DMEM/F12 plus N2 and B27 (Invitrogen) medium in the presence of Noggin (R&D) for forebrain induction for approximately 7 days. Then, floating EBs were plated onto poly-ornithine/laminin (Sigma)-coated dishes in DMEM/F12 plus N2 and B27 (Invitrogen) with addition of Noggin. Rosettes were visible to collect after 7 days. Rosettes were then dissociated with accutase (Chemicon) and plated again onto coated dishes in DMEM/F12 plus N2 and B27 and 10ng/ml of FGF2 (R&D). Homogeneous populations of NPCs were achieved after 1-2 passages with accutase in the same conditions. To obtain mature neurons, NPCs were cultured with DMEM/F12 plus N2 and B27 with addition of 1ug/ml of Laminin, BDNF (20 ng/ml), GDNF (20 ng/ml) and cyclic AMP (500 ug/ml) for 8 weeks. The full transcriptomic and functional characterization of primate NPCs and neurons will be described elsewhere (Marchetto *et al*., manuscript in preparation). The use of chimpanzee and bonobo fibroblasts was approved by the US Fish and Wildlife Service, under permit MA206206. Protocols describing the use of iPSCs and human ESCs were previously approved by the University of California, San Diego (UCSD), the Salk Institute Institutional Review Board, and the Embryonic Stem Cell Research Oversight Committee [42].

6.8.4 RNA extraction, RNA libraries, deep sequencing, and data analysis

For RNA library generation and deep sequencing, total cellular RNA was extracted from ~1-5 x10^6 cells using the RNeasy Protect Mini kit or RNeasy Plus kit (Qiagen). RNA-seq datasets derived from iPSCs were previously described [42]. For RNA library generation from human and chimpanzee NPCs and eight-week-old neurons, PolyA^+ RNA was fragmented and prepared into sequencing libraries using the Illumina TruSeq RNA sample preparation kit. NPC-derived sequencing libraries were analyzed on an Illumina HiSeq 2000 sequencer at the UCSD Biomedical Genomics Laboratory (BIOGEM). cDNA libraries were prepared from two human and two chimpanzee NPC lines (two clones each) derived from human WT-33 and ADRC-40 iPSC lines and chimpanzee PR00818 and PR01209 iPSC lines, respectively. Libraries were sequenced using single-end 100 bp reads at a depth of 15–30 million reads per library. Sequencing libraries derived from eight-week-old neurons were analyzed on an Illumina HiSeq 2500 sequencer at the Salk Next Generation Sequencing Core. cDNA libraries were prepared from two human (WT-33 and ADRC-40) and two chimpanzee (PR00818 and PR01209) neurons, one clone each. Libraries were sequenced using paired-end 125 bp reads at a depth of 15–30 million reads per library.

Gene expression was calculated in TPM with Kallisto [61] (version 0.42.1) against a custom catalog of human transcripts, including all human RefSeq transcripts with the three *BOLA2* isoforms (**Table S14** and **Fig. S45**). RefSeq isoforms nearly identical to the canonical *BOLA2* isoform (NM_001039182 and NM_001031827) were not included in the catalog of transcripts. Since the RNA-seq datasets are a mix of PE100 and SE100 reads, we quantified gene expression by using only the first read of PE100 sequencing.

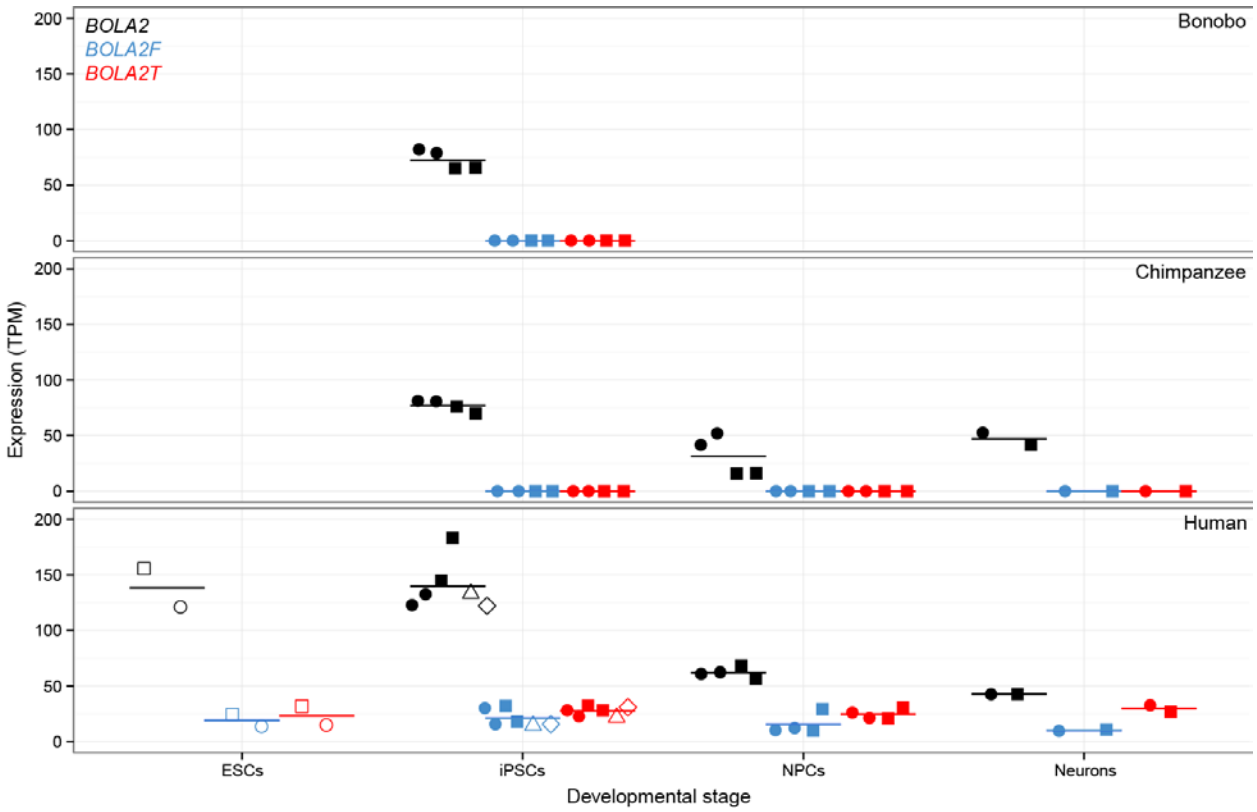


Figure S45. *BOLA2* expression over neuronal differentiation. Quantification of canonical *BOLA2* (black), *BOLA2F* (blue) and *BOLA2T* (red) in human ESCs, and human and chimpanzee iPSCs, NPCs and mature neurons. Expression of *BOLA2* isoforms is shown for each individual clone (points). Samples include human ESCs HUES6, H1 and H9 (one clone each); human iPSC lines WT-33 and ADRC-40 (two clones each); human iPSC lines WT-126 and WT9 (one clone each); chimpanzee iPSC lines PR00818 and PR01209 (two clones each); bonobo iPSC lines PR01086 and AG05253 (two clones each); iPSC-derived human NPCs WT-33 and ADRC-40 (two clones each); iPSC-derived chimpanzee NPCs PR00818 and PR01209 (two clones each); iPSC-derived human neurons WT-33 and ADRC-40 (one clone each); and iPSC-derived chimpanzee neurons PR00818 and PR01209 (one clone each). Different shapes correspond to different cell lines. Horizontal lines show mean values for each species and differentiation stage.

For visualization and quantification of reads spanning *BOLA2* exon junctions in iPSCs, we generated a Sashimi plot [62] showing RNA-seq reads that aligned to the *BOLA2* locus using the Integrative Genomics Viewer (**Fig. S46**). Reads from human iPSC lines WT-33 and ADRC-40 and chimpanzee iPSC lines PR00818 and PR01209 (two clones each) were mapped to the human (GRCh37) or chimpanzee (panTro4, CGSC 2.1.4) reference genomes using STAR with default parameters (version 2.2.0.c) [63]. All reads from human samples were mapped to the telomeric (BP4) *BOLA2-SMG1* duplication in the human reference genome; chimpanzee reads were mapped to chimpanzee reference genome. Chromosome coordinates are shown on top for both genomes.

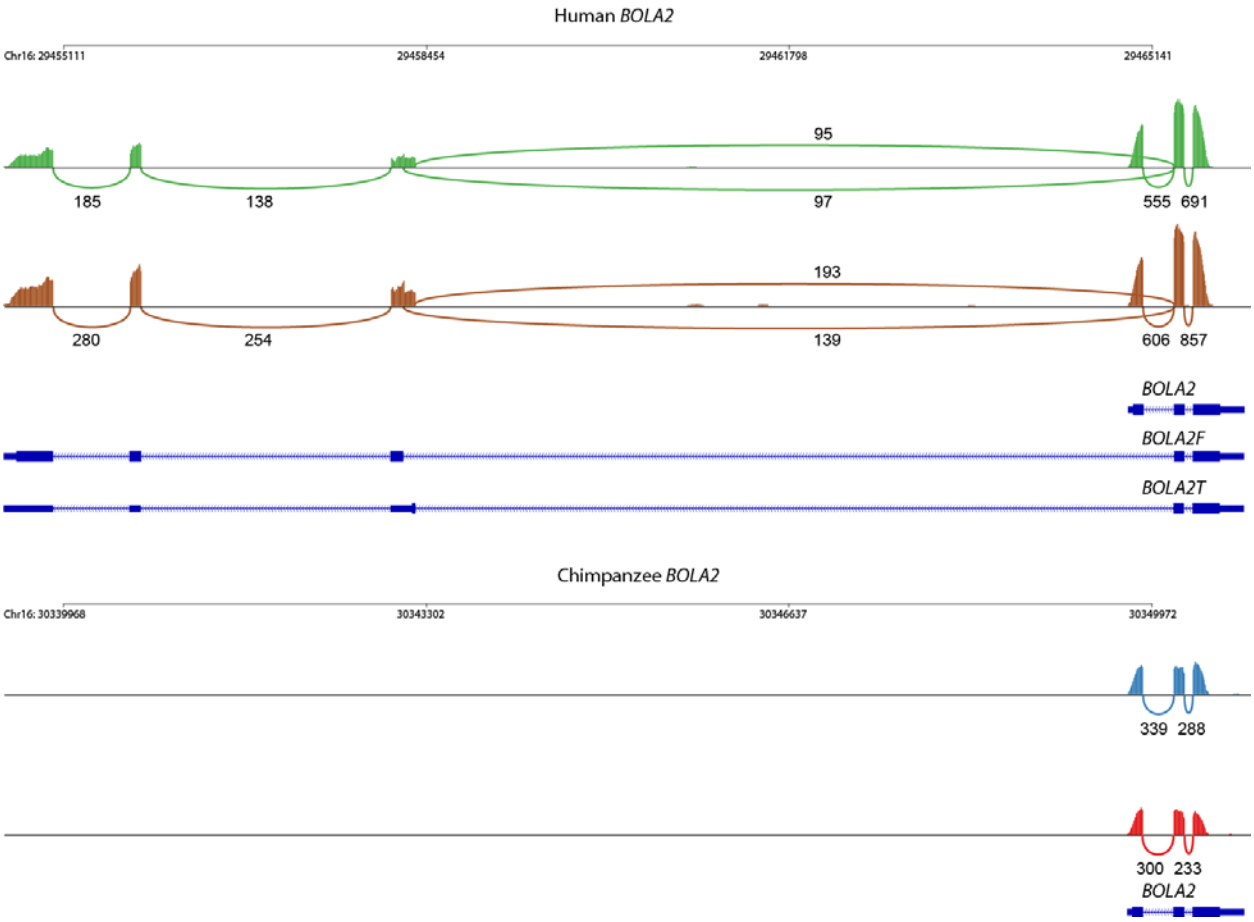


Figure S46. Schematic representation and RNA-seq quantification of *BOLA2* and *BOLA2-SMG1P* fusion transcripts in human and chimpanzee iPSCs. Sashimi plots representing the different splicing variants of *BOLA2* and human-specific *BOLA2-SMG1* in human and chimpanzee iPSC lines WT-33 (green) and ADRC-40 (brown), and chimpanzee iPSC lines PR00818 (blue) and PR01209 (red). The exon junctions (colored lines) and number of reads mapping to each are indicated (black). All reads from human samples were mapped to the telomeric (BP4) *BOLA2-SMG1* duplication in the human reference genome assembly (GRCh37); chimpanzee reads were mapped to the chimpanzee reference genome assembly (panTro4, CGSC 2.1.4). Chromosome coordinates are indicated for both genomes above plots. Three transcript isoforms are represented: *BOLA2*, *BOLA2F* and *BOLA2T*. Mature mRNA transcripts are shown in blue, and thicker lines represent the CDS while thinner lines represent UTRs. *BOLA2* consists of a 1030 nt mRNA coding for an 86 aa protein. *BOLA2F* is 1455 nt long and potentially encodes a 214 aa fusion protein including 54 aa from BOLA2 and 163 aa from SMG1. *BOLA2T* is a 1559 nt mRNA containing a premature stop codon in the first *SMG1P* exon, it contains an ORF of 58 aa, 54 aa from BOLA2 and 4 aa from SMG1P.

7. Susceptibility to recurrent 16p11.2 rearrangements

The organization (section 2.1) and high identity (**Fig. S12**) of *Homo sapiens*-specific duplicated sequences including *BOLA2* implicate them in predisposing chromosome 16p11.2 to recurrent rearrangements in humans associated with disease. We compared all directly oriented segmental duplications flanking the 16p11.2 autism critical region in human, chimpanzee, and orangutan to determine whether this susceptibility is specific to our species.

We identified directly oriented duplicated sequences flanking the autism critical region (**Table S15**) using a modified WGAC pipeline described above (section 2.1). We find that orangutan lacks directly oriented

segmental duplications flanking the autism critical region, while chimpanzee possesses only small blocks of such duplicated sequence, no more than 50 kbp in size, having at most 98.6% average sequence identity (**Fig. S47**). In contrast, human haplotypes have directly oriented duplication blocks flanking the autism critical region that are at least 117 kbp in size and exhibit at least 98.8% average sequence identity. Restricting our analysis to human haplotypes having a duplicate *BOLA2B* paralog, the blocks of interest are at least 147 kbp in size and show at least 99.3% average sequence identity.

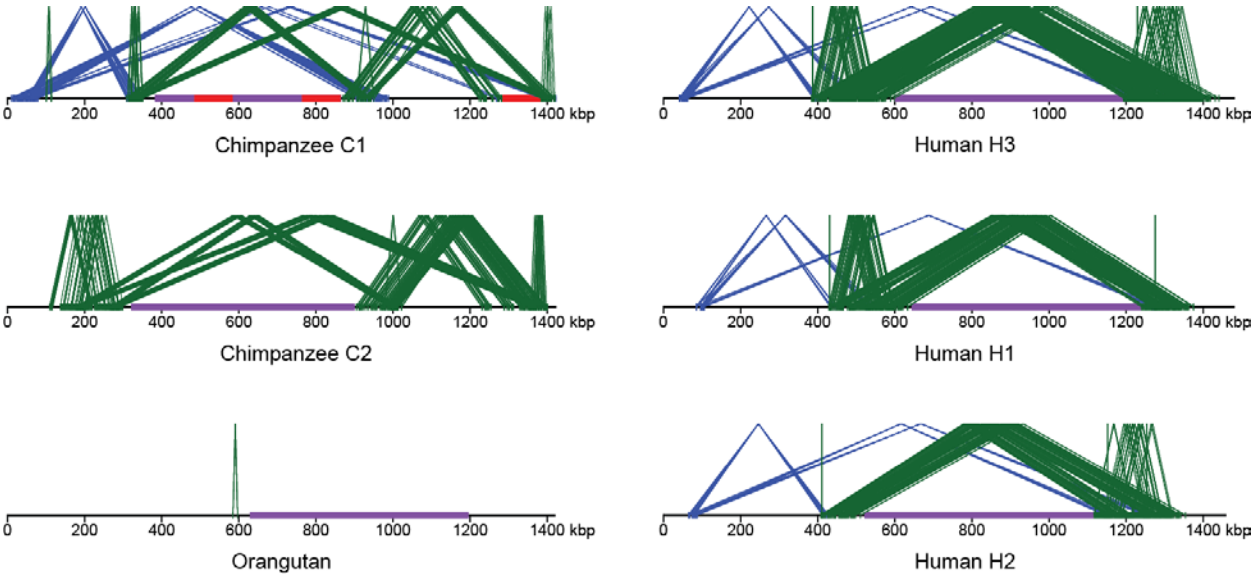


Figure S47. Comparison of duplications around the 16p11.2 autism critical region among apes. Local directly oriented (green) and inversely oriented (blue) intrachromosomal segmental duplications flanking the chromosome 16p11.2 autism critical region (purple) are visualized using Miropeats [12] (s = 1,000). Gaps in the chimpanzee C1 contig are shown in red. The smaller size (<50 kbp) and lower average sequence identity (at most 98.6%) of directly oriented duplications flanking the critical region in chimpanzee compared to human haplotypes including *BOLA2* on both sides of the critical region (at least 147 kbp of directly oriented duplications having at least 99.3% average sequence identity) suggest that susceptibility to NAHR resulting in microdeletions and microduplications at this locus evolved specifically in humans.

Long, identical stretches of sequence shared between duplications promote NAHR [64, 65]. Such regions are abundant at BP4 and BP5 for human haplotypes having *BOLA2* copies at both loci (**Fig. S12**). For each contig sequence, we identified all tracts of perfect sequence identity at least 500 bp in size within the longest contiguous region of homology between directly oriented segmental duplications flanking the autism critical region (**Table S16** and **Fig. S48**). We selected 500 bp as a threshold since it appears to represent a minimal length for efficient processing for mammalian recombination machinery [66]. Neither orangutan nor chimpanzee possess any tracts meeting the criteria above, while such tracts were found in all human haplotypes, with the highest number and longest such tracts occurring in haplotypes including *BOLA2B* (**Table S17**). Long stretches of perfect sequence identity are exclusive to humans and most prevalent on haplotypes containing the *Homo sapiens*-specific *BOLA2* duplication at BP4. These findings corroborate the conclusion that the predisposition to recurrent, disease-associated rearrangements at 16p11.2 is specific to our species.

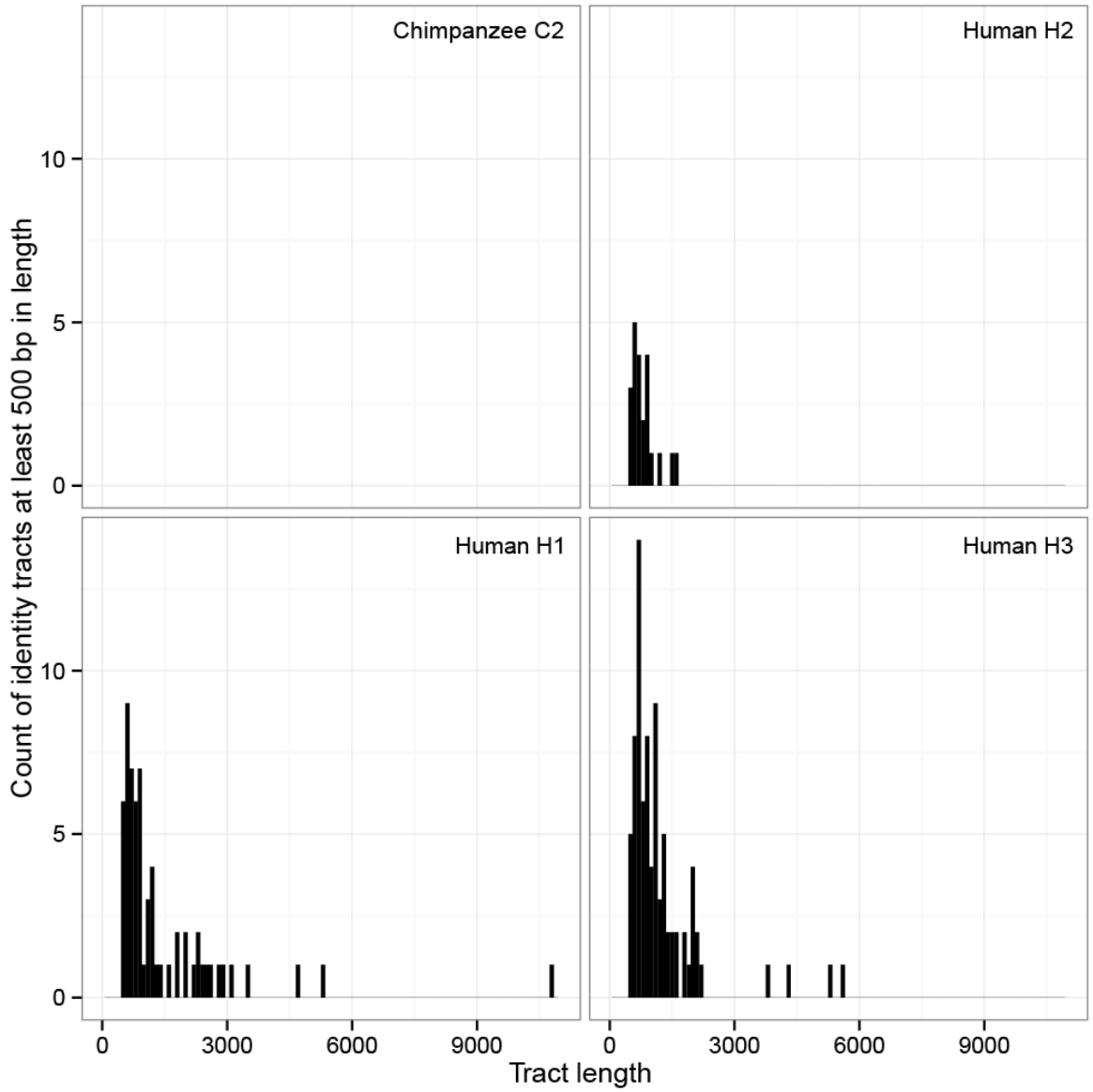


Figure S48. Perfect sequence identity tract lengths (>500 bp) within directly oriented duplications flanking the critical region for human vs. chimpanzee. Histograms show counts of tracts of perfect sequence identity (lacking single-nucleotide variants and indels) between directly oriented segmental duplications of interest within each indicated haplotype and the distribution of these tracts over different size ranges. Human haplotypes having *BOLA2* on both sides of the critical region (bottom panels) contain the highest number of such tracts and the longest such tracts, including one tract spanning 10,774 bp. In contrast, the longest tract of perfect sequence identity between duplications of interest in chimpanzee (considering both the C1 and C2 haplotypes) spans 450 bp.

8. Microdeletion/microduplication breakpoint refinement

8.1 Overview

Chromosome 16p11.2 microdeletions and microduplications associated with autism and developmental delay arise via NAHR between directly oriented segmental duplications at BP4 and BP5. To evaluate the potential role of the *Homo sapiens*-specific duplication, including *BOLA2* in promoting instability at this locus, we localized breakpoints for 151 patients carrying a typical BP4-BP5 16p11.2 rearrangement, corresponding to 72 independent microdeletions and 33 independent microduplications (**Table S18**). We utilized three methods to refine breakpoint locations: i) examination of WGS read depth at unique 30-mer locations in the human reference genome (GRCh37), ii) visualization of marker-specific WGS read count relative frequencies at positions informative for breakpoint mapping, and iii) analysis of marker-specific read count relative frequencies from sequencing data generated using a MIP assay targeting breakpoint-informative sites. We briefly detail each of these approaches below and show that, except for a few cases, they resolve breakpoints as mapping within the ~95 kbp interval corresponding to the *Homo sapiens*-specific duplication from BP5 to BP4.

8.2 Breakpoint refinement using normalized WGS read depth

We generated whole-genome shotgun sequence from three trios and three quads (21 genomes total), each including an initially identified proband having a *de novo* 16p11.2 BP4-BP5 microdeletion. Each genome was sequenced to an average coverage of at least 20-fold using the Illumina HiSeq platform, the Illumina NextSeq platform, or a combination of the two. Each sequence read was decomposed into 30-mers and mapped to the human reference genome GRCh37 using mrsFAST as previously described [5, 31]. We generated copy number variation heat maps showing aggregate and PSCN across the 16p11.2 locus. Additionally, we computed read depth at all positions corresponding to unique 30-mer sequences in the reference genome GRCh37, normalized read-depth values based on overall genome sequence coverage, and visualized the normalized data using custom tracks uploaded to the UCSC Genome Browser (**Fig. S49**) as previously described [5, 31].

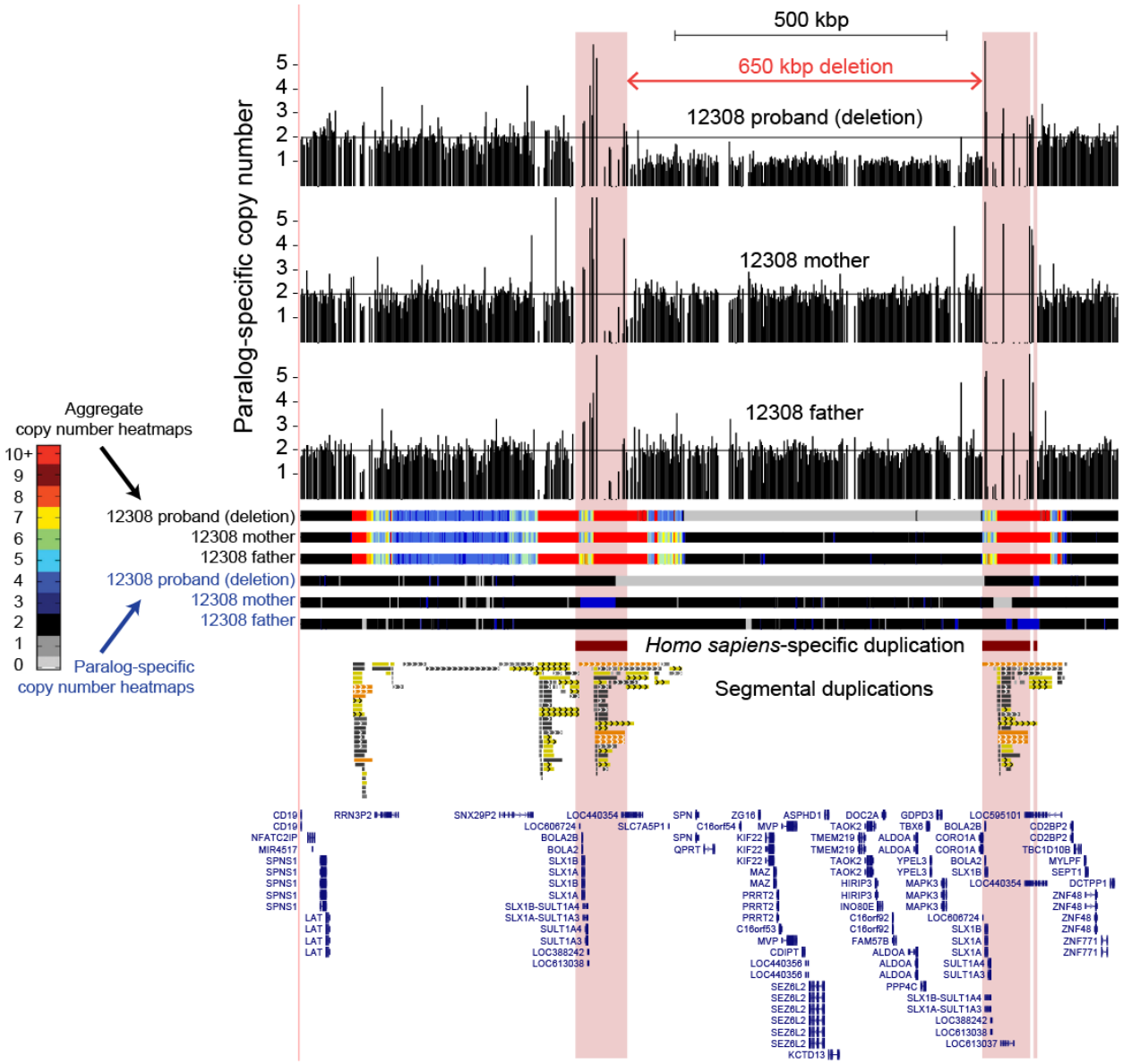


Figure S49. Breakpoint refinement based on familial WGS data from a 16p11.2 microdeletion patient. The top three tracks show normalized read depth at unique 30-mer positions for the proband, her mother, and her father, respectively. Middle tracks (groups of three) indicate aggregate and PSCN heat maps across the locus for these individuals computed in 500 bp intervals (100 bp sliding windows). Observed read-depth signatures reveal a deletion in the proband extending between, but not beyond, the *Homo sapiens*-specific duplicated sequences, a pattern consistent with breakpoints mapping within the *Homo sapiens*-specific duplication (highlighted in pink).

All six families showed the same pattern. Normalized read depth was about half of that observed in parents between the *Homo sapiens*-specific duplicated sequences but equal in probands and parents beyond the *Homo sapiens*-specific duplicated sequences. These observations refine all 16p11.2 microdeletion breakpoints examined using this method to an ~95 kbp interval including *BOLA2* (**Table S18**).

8.3 Breakpoint refinement using marker-specific WGS read count frequencies

In a second approach, we utilized WGS data together with genetic markers identified for PSCN genotyping (section 4.3) to detect breakpoint signatures. Specifically, we expect a pattern of a reciprocal marker-specific copy number shift at the location of unequal crossover [32, 67]. In this scenario, the BP4-BP5 recombinant duplication block formed by NAHR would include markers unique to BP4 before the breakpoint and markers unique to BP5 after the breakpoint (**Fig. S50**). On the other hand, NAHR between sequences at BP4 and BP5 not distinguishable using our PSCN markers, i.e., NAHR between the *Homo sapiens*-specific duplication segments, should not produce a detectable signature. In this case, the BP4-BP5 recombinant would contain the same markers across its entirety as both of the original duplication blocks at BP4 and BP5 from which it derived (**Fig. S50**).

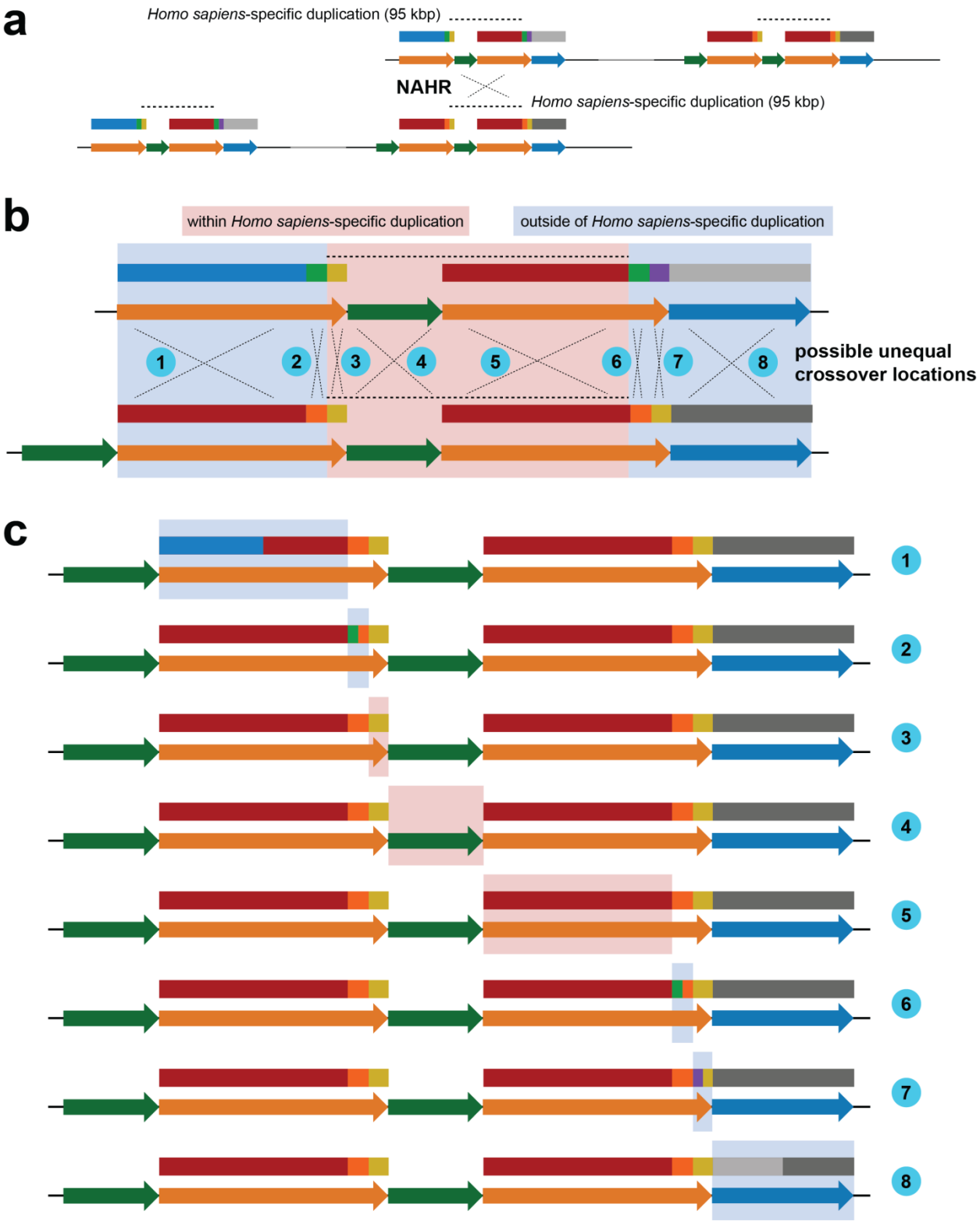


Figure S50. Expected marker-specific copy number patterns for different breakpoint locations. a) Schematic depicts NAHR between directly oriented segmental duplications at BP4 and BP5. This unequal crossover results in 16p11.2 microdeletions and microduplications (**Fig. S3**). Colored arrows and boxes correspond to duplication blocks and sectors within them color-coded as in **Fig. S21**. b) Unequal crossover could occur in eight distinct regions with

regard to duplication block and sector boundaries. Three such regions are located within the ~95 kbp *Homo sapiens*-specific duplication (highlighted in pink). c) Predicted deletion products for NAHR breakpoints within each of the eight regions in panel b. Only unequal crossover events outside the *Homo sapiens*-specific duplication produce recombinants having a sector with non-uniform marker-specific copy number across its extent (highlighted in blue).

For each sequenced genome, we plotted marker-specific read count frequencies at each PSCN marker site (**Fig. S51**), performing the same analysis described in section 4.3. In no cases did we detect a reciprocal marker-specific copy number transition as would be expected if microdeletion breakpoints occurred outside of the *Homo sapiens*-specific duplication. Thus, these data corroborate the above results (section 8.2) that in all seven microdeletion patients (six independent rearrangements), breakpoints map within the ~95 kbp *Homo sapiens*-specific duplication (**Table S18**).

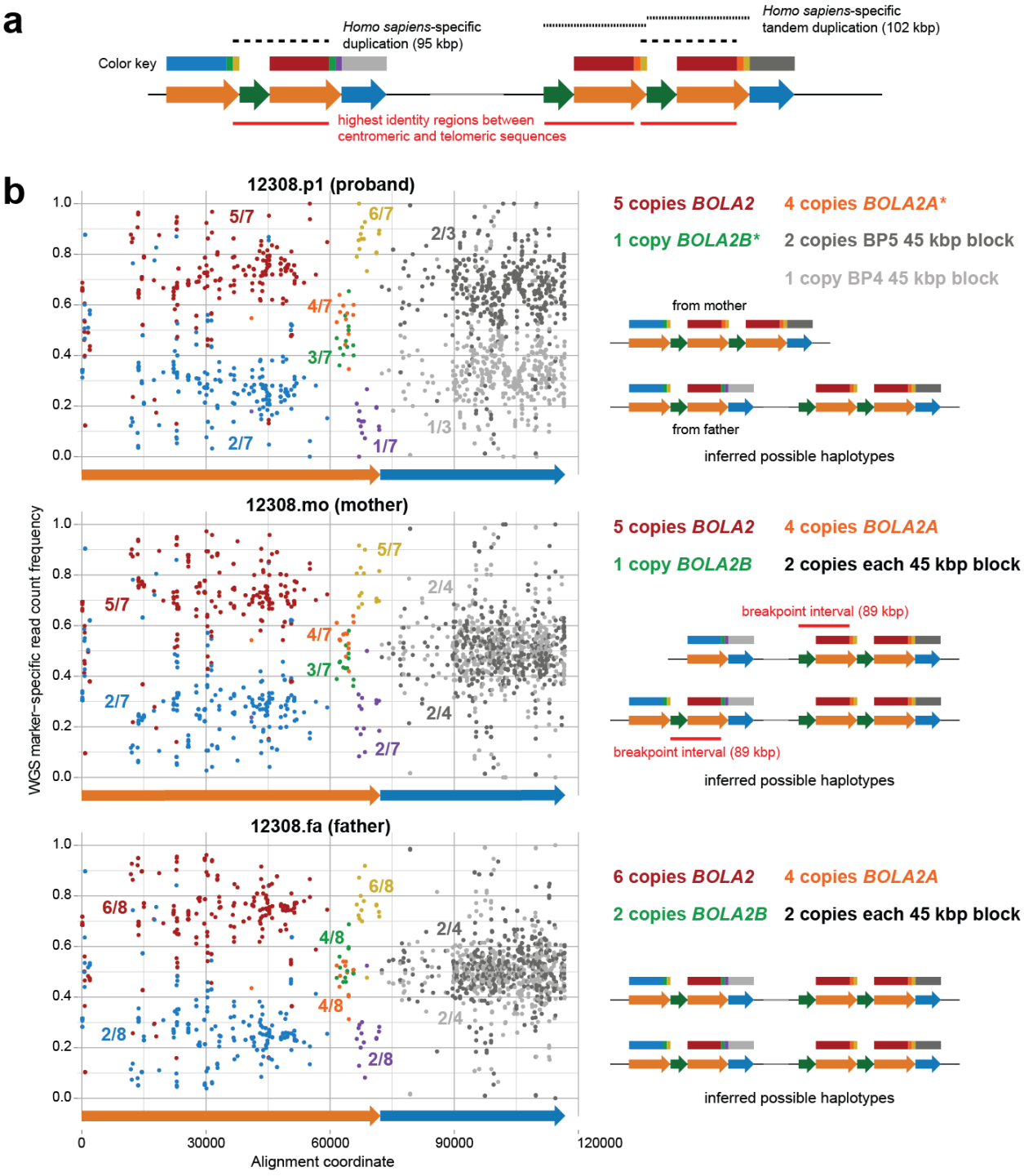


Figure S51. Breakpoint refinement based on WGS PSCN marker data from a 16p11.2 microdeletion patient. a) Schematic depicts the locations of duplication blocks (arrows) and sectors (boxes) color-coded as in **Fig. S21**, as well as the locations of *Homo sapiens*-specific duplication events (dashed lines) and regions of highest sequence identity between BP4 and BP5 (red lines). b) Plots show relative marker-specific read count frequencies (points) determined from WGS analysis for a deletion proband (top), her mother (middle), and her father (bottom). Fractions indicate relative marker-specific copy numbers as in **Fig. S21**, and diagrams adjacent to the plots show inferred haplotype structures for each chromosome 16 homolog for these individuals. Because marker-specific copy number is uniform across each sector in the deletion proband, unequal crossover breakpoints must have occurred within the

Homo sapiens-specific duplication. Red lines in the diagram for the mother indicate the most likely microdeletion breakpoint intervals. We inferred the parent-of-origin for this event using microarray data [68].

8.4 Breakpoint refinement using a MIP assay

We repurposed our paralog-specific *BOLA2* copy number MIP assay to refine microdeletion and microduplication breakpoint locations in a total of 152 individuals corresponding to 105 independent rearrangement events (**Table S18**). Specifically, we used the same MIP pool as in section 4.4 (**Table S9**), including the 47 MIPs described therein targeting markers within the 72 kbp block as well as 54 MIPs targeting markers across the 45 kbp block (**Fig. S21d**). These latter markers enable detection of breakpoints occurring within the 45 kbp block. We performed MIP sequencing and analysis as above (section 4.4), except we mapped sequencing data to that minimal genome augmented with all 45 kbp blocks from our haplotype contigs and blocks of paralogy throughout the human genome (GRCh37). We plotted marker-specific read count frequencies (section 8.3) to determine whether breakpoints for each independent rearrangement map within or outside of the *Homo sapiens*-specific duplication and to define as precisely as possible intervals within which breakpoints occurred.

Breakpoints for at least 101 of 105 rearrangement events localize to the *Homo sapiens*-specific duplication (**Fig. 4.6b** and **Fig. S52**). In two cases, the breakpoints cannot be unambiguously resolved. In the remaining two, microdeletion breakpoints map outside the *Homo sapiens*-specific duplication, instead falling within the 45 kbp block (**Fig. 4.6c** and **Fig. S53**). For these two, we narrow the putative breakpoint intervals to a 1.6 kbp region and a 22 kbp region within this block. Marker-specific read count frequency data over the 45 kbp block for these individuals indicate reciprocal transitions in BP4/BP5 PSCN. The signatures are consistent with these individuals having a total of three 45 kbp blocks at 16p11.2: two from the unaffected haplotype having BP4 or BP5 markers across their entire lengths, and one unequal crossover recombinant from the microdeletion haplotype. This recombinant has BP4 markers at its start and BP5 markers at its end. Interestingly, marker-specific read count frequency data for the sibling of one of these individuals (**Fig. 4.6c**) reveal an ~22 kbp interval within the 45 kbp block showing both an increase in BP5 marker copy number and a decrease in BP4 marker copy number. We conclude that this region likely corresponds to an interlocus gene conversion event between BP5 and BP4, with the former serving as the conversion donor. This event was observed specifically in this family and was inferred to be present in the germline of the father (DNA not available for testing) based on its absence in the mother. Note that the start of the conversion region in the sibling maps to the same location as the PSCN transition in the proband. This interlocus gene conversion, thus, created a high sequence identity interval within the 45 kbp block predisposing this region to unequal crossover between BP4 and BP5 (**Fig. 4.6d**) in this family.

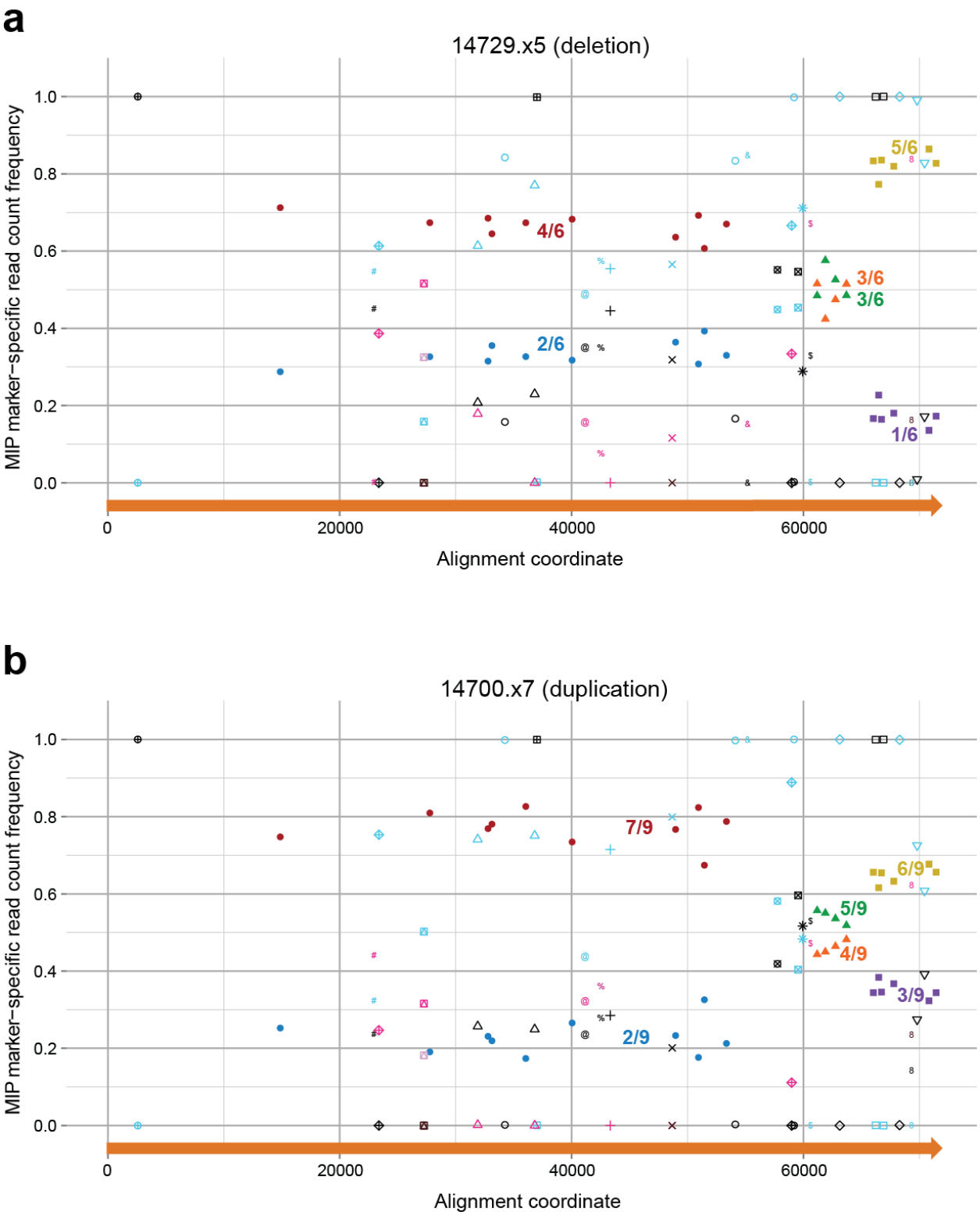


Figure S52. Breakpoint refinement based on MIP PSCN marker data. Plots show relative marker-specific read count frequencies (points) determined using MIPs for a typical microdeletion patient (panel a) and a typical microduplication patient (panel b). Shapes and color code designate different markers, and fractions indicate relative marker-specific copy numbers (as in **Fig. S21**). Because marker-specific copy number is uniform across each sector

for both individuals, in both cases, unequal crossover breakpoints must have occurred within the *Homo sapiens*-specific duplication.

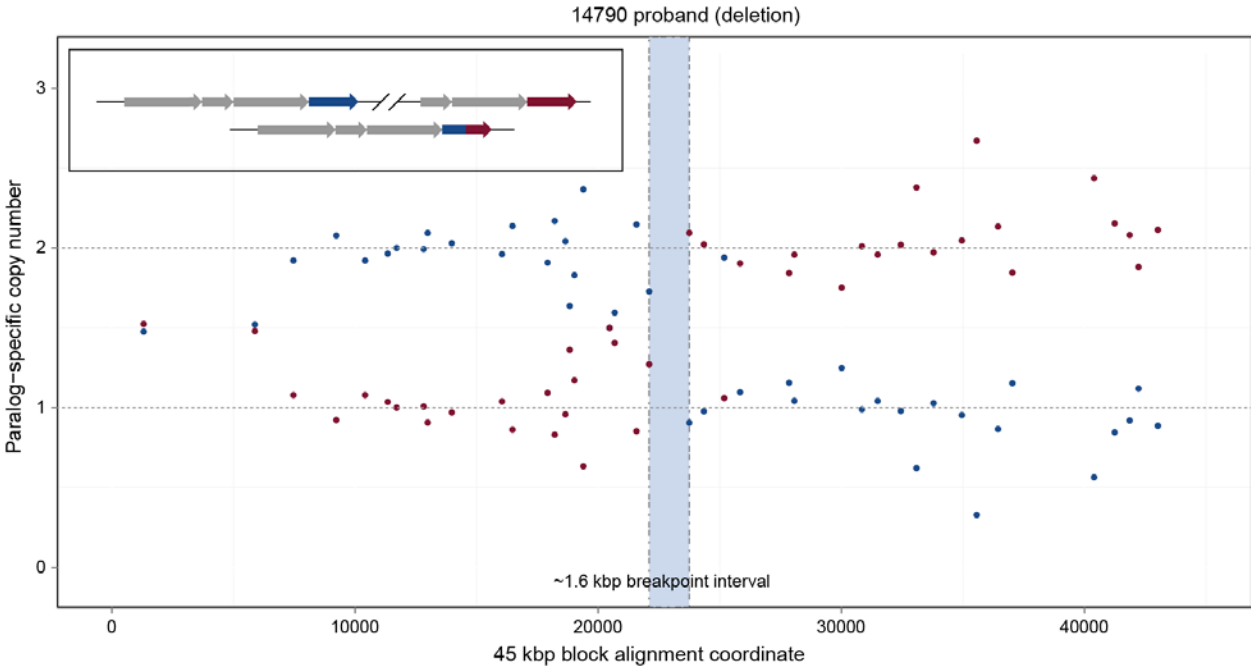


Figure S53. An atypical microdeletion breakpoint within the 45 kbp block. The plot shows PSCN determined using a MIP assay across the 45 kbp duplication block for a microdeletion patient. Points correspond to markers distinguishing BP4 (blue) or BP5 (red) copies. Schematic (upper left) indicates the inferred haplotypes for this individual. The reciprocal copy number transition (highlighted in blue) defines an ~1.6 kbp interval within which the unequal crossover likely occurred. This region is clearly outside of the *Homo sapiens*-specific duplication.

Marker-specific read count frequency data for the remaining two rearrangements showed signatures consistent with breakpoints mapping within the 72 kbp block but outside of the *Homo sapiens*-specific duplication. However, these signatures were the same between patients from the two different families of interest and matched an interlocus gene conversion signature observed in some individuals lacking any 16p11.2 rearrangement (**Fig. S54**). Thus, in these cases, the data are consistent with either atypical breakpoints or breakpoints within the *Homo sapiens*-specific duplication together with 72 kbp blocks affected by interlocus gene conversion.

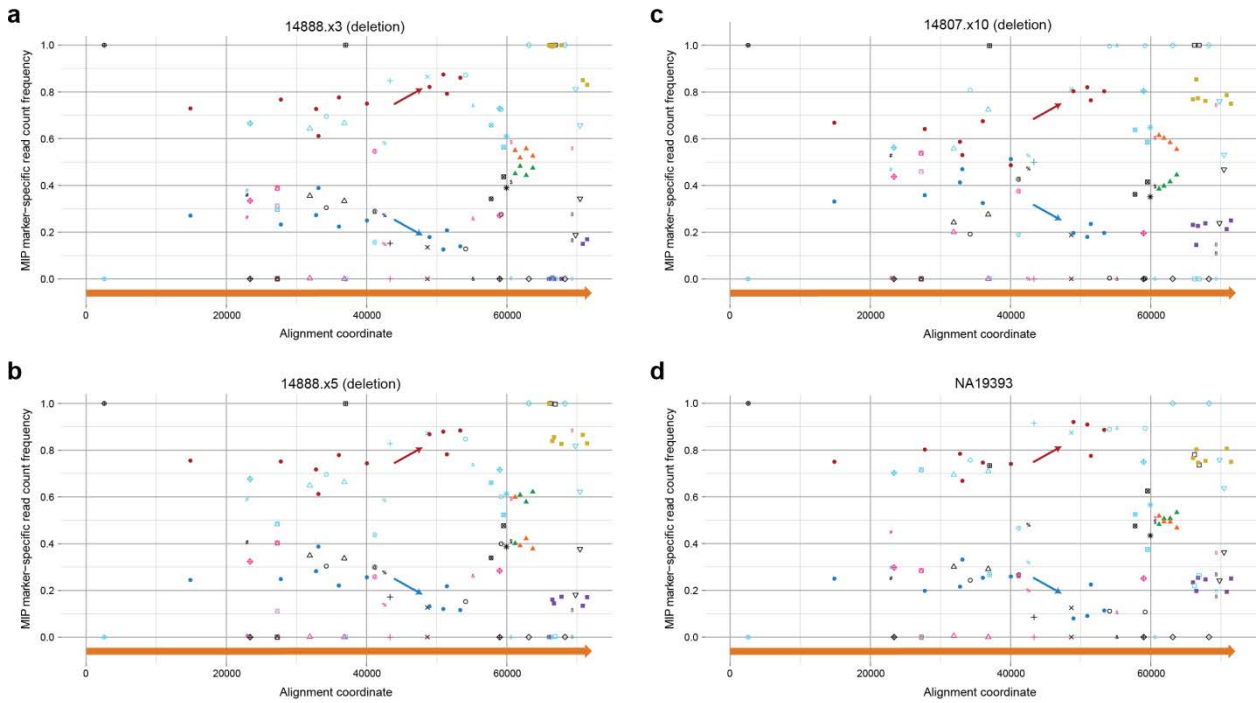


Fig. S54. Ambiguous breakpoint locations. Plots show relative marker-specific read count frequencies (points) determined using MIPs as in **Fig. S52**. Panels a-c show data from microdeletion patients, while panel d corresponds to an individual lacking the microdeletion. Marker-specific copy number is not uniform (highlighted by thin arrows) across the 59 kbp sector for the microdeletion patients, consistent with NAHR breakpoints occurring outside the *Homo sapiens*-specific duplication. However, data from the individual lacking the microdeletion exhibits the same pattern, which is best explained in this case by interlocus gene conversion. Thus, the reciprocal marker-specific copy number shift in the microdeletion patients may reflect atypical breakpoint locations or interlocus gene conversion. In the latter scenario, breakpoints for these individuals would map within the *Homo sapiens*-specific duplication.

9. Additional methods and analyses

9.1 Fluorescence *in situ* hybridization

FISH experiments (**Table S2**) were used to assay aggregate *BOLA2* copy number variation (**Fig. 4.3b** and **Fig. S19**), to show that such variation affects both BP4 and BP5 (**Fig. S25**), to compare 16p11.2 organization between human, chimpanzee, and orangutan (**Figs. S4-S5**), and to assess the duplication from 16q24.2 in human, chimpanzee, gorilla, and orangutan (**Fig. S9**).

Metaphase spreads were obtained from lymphoblast and fibroblast cell lines from human HapMap individuals HG01067, HG02314, NA12275, NA12878, NA19041, NA19091, and NA20127, as well as from chimpanzee (PTR5), gorilla (GGO5), and orangutan (PPY10, PPY13, and PPY16; Coriell Cell Repository, Camden, NJ). FISH experiments were performed using human fosmid clones (**Table S2**) identified based on mapping clone paired-end sequence data [2] to the human reference genome GRCh37. Clones were directly labeled with Cy3-dUTP, Cy5-dUTP (GE Healthcare), or Fluorescein-dUTP (Invitrogen) by nick translation as previously described [5], with minor modifications. Two hundred ng of labeled probe were hybridized on metaphase spreads; hybridization was performed overnight at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, 5 µg COT1 DNA (Roche), and 3 µg sonicated salmon sperm DNA, in a volume of 10 µL. Post-hybridization washing was performed at 60°C in 0.1xSSC (three times, high stringency). Washes for interspecies hybridization experiments were

performed at lower stringency: 37°C in 2xSSC, 50% formamide, followed by washes at 42°C in 2xSSC. Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI, Cy3, Cy5 and fluorescein fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

9.2 RT-PCR

We amplified *BOLA2* transcripts using total RNA (Clontech) from a variety of human tissues (section 6.1). To prepare cDNA libraries, we used the Transcriptor High Fidelity cDNA Synthesis Kit (Roche). For each RNA sample, a template-primer mixture was prepared by combining 9.07 µL PCR-quality water, 1.33 µL total RNA (at a concentration at least 1 µg/µL), and 1 µL oligo dT primer. RNA secondary structures were denatured by incubating this mixture for 10 mins at 65°C, and afterwards the mixture was immediately cooled on ice. An RT-PCR master mix was prepared by combining 72 µL RT reaction buffer, 9 µL RNase inhibitor, 36 µL dNTP mix (10 mM),18 µL DTT, and 19.8 µL reverse transcriptase. 8.6 µL of this master mix was added to each template-primer mixture. Reverse transcription was performed by incubating this 20 µL reaction for 30 min at 50°C, followed by a 5 min incubation at 85°C and cooling on ice.

Using the Expand Long Template PCR System (Roche), we prepared two PCR master mixes, one including primers targeting the canonical *BOLA2* isoform and one including primers targeting fusion isoforms. Each master mix contained 50 µL 10x buffer 1, 17.5 µL dNTP mix (10 mM), 15 µL forward primer (10 µM), 15 µL reverse primer (10 µM), 355 µL PCR-quality water, and 7.5 µL enzyme mix. For each cDNA library, we set up a PCR reaction by combining 46 µL of the master mix with 4 µL of cDNA. Reactions were incubated at 94°C for 2 min, followed by 10 cycles at 94°C for 10 s, 55°C for 30 s, and 68°C for 3 min, followed by 25 cycles at 94°C for 10 s, 55°C for 30 s, and 68°C for 3 min + 20 additional seconds each cycle, followed by 68°C for 7 min and finally 4°C indefinitely.

9.3 Western blotting

Human LCLs were grown in RPMI-1640 medium (Gibco) supplemented with 15% fetal bovine serum and 1% antibiotics (penicillin and streptomycin). Cells were lysed in RIPA buffer (Millipore) supplemented with protease inhibitors. Protein lysates were run on an SDS-PAGE gel and transferred to a PVDF membrane. After blocking the membrane in PBS-T 0.05%, gelatin 1%, the primary antibodies were incubated overnight at 4°C. The membrane was then washed, incubated with the appropriate HRP-conjugated secondary antibody for 1 h at RT, washed and revealed.

9.4 CMV transient expression in HeLa cells

HeLa cells were grown in DMEM (Gibco) supplemented with 10% fetal bovine serum and 1% antibiotics (penicillin and streptomycin). Gateway PLUS shuttle clone for human *BOLA2* corresponding to the annotated 17 kDa protein CDS was obtained from tebu-bio (GC-Z3591), moved to pCS2plus destination vector by LR reaction for CMV transient and constitutive expression in cells, and transfected to HeLa cells using FuGENE reagent (Promega) in medium without antibiotics. After 24 h, cells were lysed in RIPA buffer and whole lysates were analyzed by western blotting. To analyze the conditioned medium, HeLa cells were transfected in absence of serum and after 24 h the conditioned medium was collected and concentrated (10-fold) using 3K centrifugal filter devices (Millipore).

9.5 *BOLA2* 10 kDa Gateway cloning

The BOLA2 10 kDa CDS was amplified using attB sites flanked primers and *Pfu* DNA polymerase (Promega). The gel-purified PCR product was cloned in the pDONR221 vector (Invitrogen) through the BP reaction (Invitrogen). *BOLA2* CDS was moved to the pCS2plus destination vector for CMV constitutive expression through the LR reaction (Invitrogen).

9.6 Immunofluorescence

Human LCLs were grown in suspension and seeded on poly-L-lysinated coverslips for 1 h at RT, then fixed with PBS, PFA 4%. HeLa cells were transfected with the CMV *BOLA2* 10 kDa plasmid for 24 h, then fixed with PBS, PFA 4%. Cells were permeabilized with PBS, Triton-X 0.03% for 10' at RT, blocked with PBS, Triton-X 0.03% and gelatin 1% for 1 h at RT and incubated with the Santa Cruz anti-BOLA2 antibody O/N at 4°C. Cells were washed 3X in PBS, Triton-X 0.03% and incubated with an anti-goat Alexa Fluor 488 conjugated secondary antibody. After washing and mounting with DAPI-vectashield, cells were observed using a Zeiss LSM710 confocal microscope.

9.7 Inversion density analysis

We evaluated the probability of large-scale inversions clustering in genomic space with a permutation test using 27 previously described inversions found in the human or chimpanzee lineages [69]. We first measured the density of observed inversions based on the distance between pairs of inversions on the same chromosome. To minimize error in pairwise comparisons caused by coarse breakpoint resolution, we measured the distance between inversion midpoints as calculated by the minimum start coordinate plus half of the range between start and end coordinates. Inversions that occurred alone on a chromosome were necessarily omitted from the distance calculation leaving 12 of the 27 original inversions. After visual confirmation that the distribution of observed distances was not normal, we selected the median as a summary statistic to compare observed and null distributions. To generate a null distribution for inversion distances, we randomly shuffled the coordinates for the full extent of all 27 published inversions across all GRCh37 genomic space for 1,000,000 iterations and calculated the median distance between midpoints for each iteration. We found that the observed median distance between previously published inversion midpoints (9.1 Mbp) was significantly smaller than expected based on the null distribution ($p = 0.003262$), which had a median of 29 Mbp (**Fig. S55**). Most strikingly, null distribution median distances were never as small as or smaller than the observed 216 kbp median distance between 16p11.2 inversion midpoints ($p < 0.000001$).

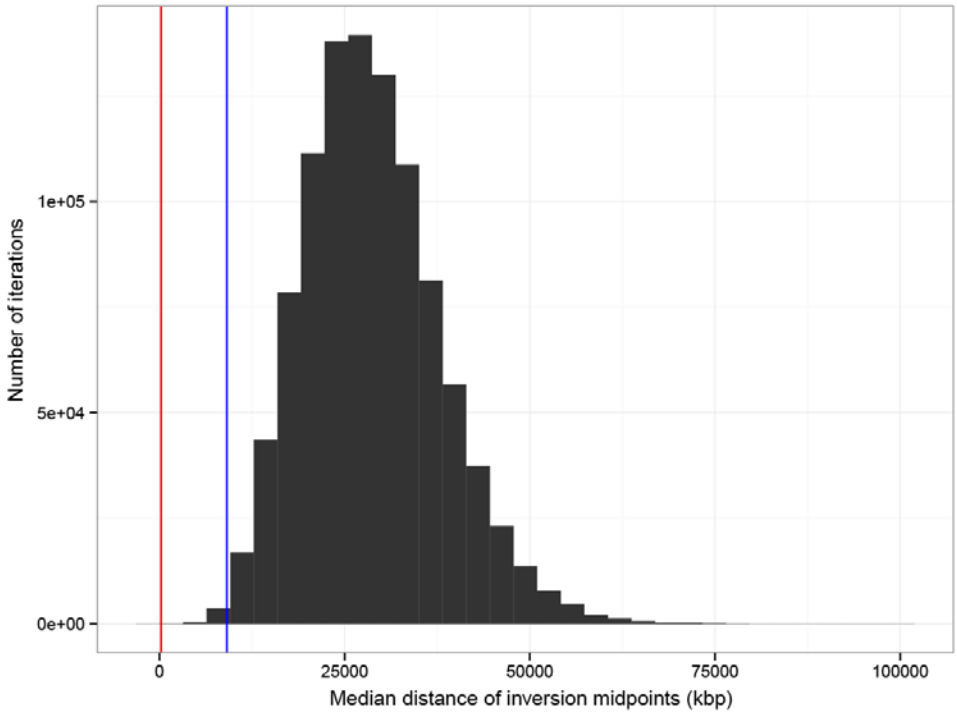


Figure S55. Null distribution for median distance between inversion midpoints for 27 previously published human or chimp inversions randomly placed across the entire genome. The observed median distance between published inversion midpoints (blue line at 9.1 Mbp) is significantly smaller than expected by the null distribution based on 1 million iterations ($p = 0.003262$). The observed median distance between 16p11.2 inversion midpoints (red line at 216 kbp) is also significantly smaller than expected ($p < 0.000001$).

Inversion breakpoints are typically associated with segmental duplications, telomeres, and centromeres. To test whether the density of inversions we observed in human and chimpanzee is significant with respect to the genomic space of these associated sequences, we ran an additional permutation test for 1,000,000 iterations where inversion breakpoints were shuffled across segmental duplications, telomeres (150 kbp from chromosome ends), and centromeres (5 Mbp on either side of annotated centromeres in GRCh37) to create the null distribution. We found that the observed median distance between inversions (9.1 Mbp) was significantly smaller than expected based on the null distribution ($p = 0.005602$) which had a median of ~27 Mbp (**Fig. S56**). Again, null distribution median distances were never as small as or smaller than the observed 216 kbp median distance between 16p11.2 inversion midpoints ($p < 0.000001$).

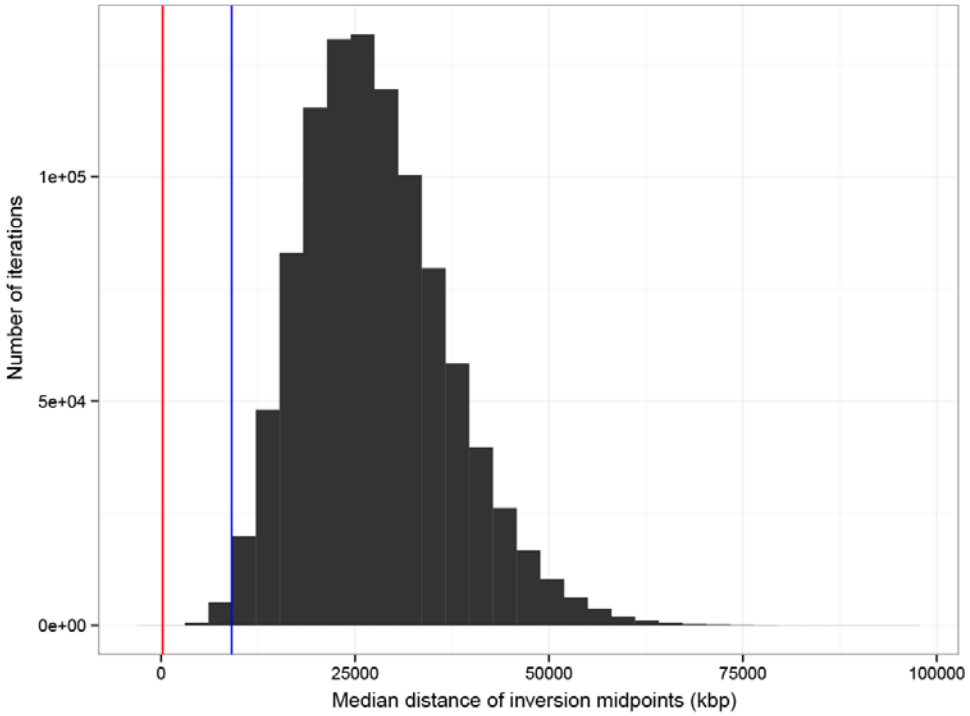


Figure S56. Null distribution for median distance between inversion midpoints for 27 human or chimp inversions whose breakpoints were randomly placed within segmental duplications, telomeres, or centromeres. The observed median distance between published inversion midpoints (blue line at 9.1 Mbp) is significantly smaller than expected by the null distribution based on 1 million iterations ($p = 0.005602$). The observed median distance between 16p11.2 inversion midpoints (red line at 216 kbp) is also significantly smaller than expected ($p < 0.000001$).

Supplementary References

1. Huddleston, J., et al., *Reconstructing complex regions of genomes using long-read sequencing technology.* Genome Res, 2014. **24**(4): p. 688-96.
2. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
3. Dennis, M.Y., et al., *Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication.* Cell, 2012. **149**(4): p. 912-22.
4. Adey, A., et al., *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.* Genome Biol, 2010. **11**(12): p. R119.
5. Antonacci, F., et al., *Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability.* Nat Genet, 2014. **46**(12): p. 1293-302.
6. Zufferey, F., et al., *A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders.* J Med Genet, 2012. **49**(10): p. 660-8.
7. Chin, C.S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.* Nat Methods, 2013. **10**(6): p. 563-9.
8. Chaisson, M.J., et al., *Resolving the complexity of the human genome using single-molecule sequencing.* Nature, 2015. **517**(7536): p. 608-11.
9. Jiang, Z., et al., *DupMasker: a tool for annotating primate segmental duplications.* Genome Res, 2008. **18**(8): p. 1362-8.
10. Bailey, J.A., et al., *Segmental duplications: organization and impact within the current human genome project assembly.* Genome Res, 2001. **11**(6): p. 1005-17.
11. Eichler, E.E. and A.W. Zimmerman, *A hot spot of genetic instability in autism.* N Engl J Med, 2008. **358**(7): p. 737-9.
12. Parsons, J.D., *Miropeats: graphical DNA sequence comparisons.* Comput Appl Biosci, 1995. **11**(6): p. 615-9.
13. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2.3.
14. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics, 2005. **21**(9): p. 1859-75.
15. Johnson, M.E., et al., *Positive selection of a gene family during the emergence of humans and African apes.* Nature, 2001. **413**(6855): p. 514-9.
16. Jiang, Z., et al., *Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.* Nat Genet, 2007. **39**(11): p. 1361-8.
17. Johnson, M.E., et al., *Recurrent duplication-driven transposition of DNA during hominoid evolution.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17626-31.
18. Patterson, N., et al., *Genetic evidence for complex speciation of humans and chimpanzees.* Nature, 2006. **441**(7097): p. 1103-8.
19. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
20. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
21. Gonzalez, J.R., et al., *A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity.* Am J Hum Genet, 2014. **94**(3): p. 361-72.
22. Martin, J., et al., *The sequence and analysis of duplication-rich human chromosome 16.* Nature, 2004. **432**(7020): p. 988-94.
23. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

24. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
25. Sudmant, P.H., et al., *Global diversity, population stratification, and selection of human copy-number variation.* Science, 2015. **349**(6253): p. aab3761.
26. Fu, Q., et al., *Genome sequence of a 45,000-year-old modern human from western Siberia.* Nature, 2014. **514**(7523): p. 445-9.
27. Lazaridis, I., et al., *Ancient human genomes suggest three ancestral populations for present-day Europeans.* Nature, 2014. **513**(7518): p. 409-13.
28. Prufer, K., et al., *The complete genome sequence of a Neanderthal from the Altai Mountains.* Nature, 2014. **505**(7481): p. 43-9.
29. Meyer, M., et al., *A high-coverage genome sequence from an archaic Denisovan individual.* Science, 2012. **338**(6104): p. 222-6.
30. Prado-Martinez, J., et al., *Great ape genetic diversity and population history.* Nature, 2013. **499**(7459): p. 471-5.
31. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes.* Science, 2010. **330**(6004): p. 641-6.
32. Nuttle, X., et al., *Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions.* Nat Methods, 2013. **10**(9): p. 903-9.
33. Illumina, I. *Platinum Genomes*. Available from: http://www.illumina.com/platinumgenomes/.
34. Center, T.N.Y.G., *Unpublished data*.
35. Hiatt, J.B., et al., *Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation.* Genome Res, 2013. **23**(5): p. 843-54.
36. Yang, M.A., K. Harris, and M. Slatkin, *The projection of a test genome onto a reference population and applications to humans and archaic hominins.* Genetics, 2014. **198**(4): p. 1655-70.
37. Hudson, R.R., *Generating samples under a Wright-Fisher neutral model of genetic variation.* Bioinformatics, 2002. **18**(2): p. 337-8.
38. Ewing, G. and J. Hermisson, *MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.* Bioinformatics, 2010. **26**(16): p. 2064-5.
39. McIlwain, D.R., et al., *Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay.* Proc Natl Acad Sci U S A, 2010. **107**(27): p. 12186-91.
40. *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
41. Patro, R., S.M. Mount, and C. Kingsford, *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.* Nat Biotechnol, 2014. **32**(5): p. 462-4.
42. Marchetto, M.C., et al., *Differential L1 regulation in pluripotent stem cells of humans and apes.* Nature, 2013. **503**(7477): p. 525-9.
43. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
44. Computing, R.F.f.S. *R: A Language and Environment for Statistical Computing*. 2014; Available from: http://www.R-project.org/.
45. Migliavacca, E., et al., *A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology.* Am J Hum Genet, 2015. **96**(5): p. 784-96.
46. Zhou, Y.B., et al., *hBolA, novel non-classical secreted proteins, belonging to different BolA family with functional divergence.* Mol Cell Biochem, 2008. **317**(1-2): p. 61-8.
47. Salamov, A.A., T. Nishikawa, and M.B. Swindells, *Assessing protein coding region integrity in cDNA sequencing projects.* Bioinformatics, 1998. **14**(5): p. 384-90.
48. Pedersen, A.G. and H. Nielsen, *Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 226-33.

49. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nat Methods, 2006. **3**(3): p. 211-22.
50. Michel, A.M., et al., *GWIPS-viz: development of a ribo-seq genome browser.* Nucleic Acids Res, 2014. **42**(Database issue): p. D859-64.
51. Desiere, F., et al., *The PeptideAtlas project.* Nucleic Acids Res, 2006. **34**(Database issue): p. D655-8.
52. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
53. Sievers, F. and D.G. Higgins, *Clustal omega.* Curr Protoc Bioinformatics, 2014. **48**: p. 3.13.1-3.13.16.
54. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.
55. Li, H. and C.E. Outten, *Monothiol CGFS glutaredoxins and BolA-like proteins: [2Fe-2S] binding partners in iron homeostasis.* Biochemistry, 2012. **51**(22): p. 4377-89.
56. Li, H., et al., *Human glutaredoxin 3 forms [2Fe-2S]-bridged complexes with human BolA2.* Biochemistry, 2012. **51**(8): p. 1687-96.
57. Witte, S., et al., *Inhibition of the c-Jun N-terminal kinase/AP-1 and NF-kappaB pathways by PICOT, a novel protein kinase C-interacting protein with a thioredoxin homology domain.* J Biol Chem, 2000. **275**(3): p. 1902-9.
58. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
59. Cowan, C.A., et al., *Derivation of embryonic stem-cell lines from human blastocysts.* N Engl J Med, 2004. **350**(13): p. 1353-6.
60. Marchetto, M.C., et al., *A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells.* Cell, 2010. **143**(4): p. 527-39.
61. Bray, N., et al. *Near-optimal RNA-Seq quantification*. arXiv, 2015. 1505.02710.
62. Katz, Y., et al. *Sashimi plots: Quantitative visualization of RNA sequencing read alignments*. 2013. 1306.3466.
63. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
64. Cooper, G.M., et al., *A copy number variation morbidity map of developmental delay.* Nat Genet, 2011. **43**(9): p. 838-46.
65. Dittwald, P., et al., *NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits.* Genome Res, 2013. **23**(9): p. 1395-409.
66. Mundia, M.M., et al., *Nascent DNA synthesis during homologous recombination is synergistically promoted by the rad51 recombinase and DNA homology.* Genetics, 2014. **197**(1): p. 107-19.
67. Nuttle, X., et al., *Resolving genomic disorder-associated breakpoints within segmental DNA duplications using massively parallel sequencing.* Nat Protoc, 2014. **9**(6): p. 1496-513.
68. Duyzend, M., et al., *Maternal modifiers and parent-of-origin bias of the autism 16p11.2 CNV*. 2015.
69. Antonacci, F. and M. Ventura. 2015.
70. *Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders.* Neuron, 2012. **73**(6): p. 1063-7.
71. Fischbach, G.D. and C. Lord, *The Simons Simplex Collection: a resource for identification of autism genetic risk factors.* Neuron, 2010. **68**(2): p. 192-5.

VITA

Xander Nuttle was born in Butler, Pennsylvania in 1988. He grew up in Plum Borough and graduated from Pittsburgh Central Catholic High School in 2006. From 2006-2010 he studied at Duke University in Durham, North Carolina, graduating with a B.S. in Biology, a B.S.E. in Biomedical Engineering, and a Certificate in Genome Sciences and Policy. In 2010, he enrolled in graduate school at the University of Washington where he joined the lab of Dr. Evan E. Eichler in 2011. He greatly enjoyed his time in the lab pursuing his deep curiosity and earned a Doctor of Philosophy in Genome Sciences in 2015. Xander is also passionate about his Catholic faith, playing soccer, scrambling to craggy summits, reading, learning new things, teaching others, and supporting his favorite sports teams.