New methods for haplotyping and *de novo* assembly of genomes and metagenomes

Joshua N. Burton

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

Doctor of Philosophy

University of Washington
2014

Reading Committee:
Jay Shendure, Chair
Maitreya J. Dunham
Phil Green

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

New methods for haplotyping and *de novo* assembly of genomes and metagenomes

Joshua N. Burton

Chair of the Supervisory Committee:
Associate Professor Jay Shendure
Department of Genome Sciences

The study of genomics is made possible by the creation of genome assemblies: strings of sequences that represent the DNA content of a species, or an individual within a species. However, genome assemblies do not spring fully formed from DNA sequencing machines. Sequencing produces small fragments of DNA, and these fragments must be combined into a prediction of an organism's genome by a process called *de novo* assembly. The advent of "next-generation" DNA sequencing technologies over the past decade has vastly increased our capacity to sequence new genomes, but it has exacerbated the difficulty of *de novo* assembly, turning it into one of the foremost challenges in computational biology today. Especially problematic is the short length of many next-generation reads, which deprives genome assemblies of crucial information about sequence contiguity. Here I describe new methods for creating high-contiguity genome assemblies from short next-generation reads. I demonstrate that novel proper library preparation can create short reads that retain long-range contiguity information, and I develop novel algorithms to exploit this information for *de novo* genome assembly. First, I use fosmid clone pools and copy number analysis to perform haplotype resolution on the genome of the famous HeLa cancer cell line. Secondly, introduce the concept of using Hi-C for *de novo* genome assembly. I demonstrate that Hi-C produces signals of genomic contiguity that can be used for chromosome-scale scaffolding of *de novo* genome assemblies. Lastly, I show that Hi-C can also be used for metagenomic deconvolution. These approaches allow us to make productive use of the continual advances in next-generation sequencing and will improve standards for genome assemblies.

ACKNOWLEDGMENTS

During my time in graduate school I have been standing on the shoulders of giants. I have had the great fortune to work with a number of individuals who are far more intelligent and talented than I am. Without them, none of this work would have been conceivable, let alone possible.

First and foremost, I am extremely grateful to my advisor, Jay Shendure, who has been a phenomenal mentor and leader. Jay has enthusiastically supported my ideas and has always provided helpful and incisive advice, enabling me to turn those ideas into new genomic methods. He has patiently shepherded me through the many challenges of a Ph.D., including learning new subfields, conducting my research thoroughly and methodically, submitting papers through the peer review process, presenting my work – in short, becoming a scientist. All the while, he has made the lab an enjoyable place to work hard. Graduate school may be the only opportunity I'll ever have to choose my own boss, and I am confident that I could not possibly have chosen better.

The Shendure lab is very collaborative, and my own intellectual development has benefited greatly from the advice of other lab members. In particular, I would like to acknowledge several lab members who have contributed their own wisdom to my own intellectual development: Andrew Adey, Riza Daza, Martin Kircher, Jacob Kitzman, Akash Kumar, Charlie Lee, Aaron McKenna, Rupali Patwardhan, Ruolan Qiu, Matthew Snyder, and Steve Salipante. I have enjoyed collaborating with the Shendure lab community in scientific as well as non-scientific pursuits.

I would also like to thank many other professors affiliated with the Genome Sciences department. First of all, the professors on my thesis committee – Maitreya Dunham, Evan Eichler, Phil Green, and Brian Browning – have provided invaluable advice, guidance, and encouragement throughout my graduate career. I am thankful to Elhanan Borenstein and Debbie Nickerson, as well as Jay, for instructing and supporting me during my first-year rotations. I have enjoyed the opportunity to serve as a teaching assistant in two Genome Sciences classes, and I am indebted to MK Raghuraman, Anne Paul, Bob Waterston, and David Hawkins for making these classes very enriching experiences for me. Lastly, in

addition to all of the aforementioned professors, I would like to thank Joe Felsenstein, Ben Hall, Bill Noble, and Jim Thomas for giving generous advice at many junctures.

The entire UW Department of Genome Sciences has been very helpful and supportive, and many of my most fruitful collaborations have been with people outside the Shendure lab. I would like to thank Ivan Liachko for his tireless enthusiasm and his faith in the power of the work we have done together. I would also like to thank my other collaborators in the greater UW and FHCRC community: Ferhat Ay, Jason Bielas, David Fredricks, Matt Laurie, Laura Sycuro, Sean Taylor, Nelle Varoquaux, and Adam Waalkes. Other people who have given me helpful advice include Rogan Carr, Keolu Fox, Adam Gordon, Roie Levy, Aaron Miller, Matt Rich, Ian Stanaway, Sanna Sullivan, Jason Underwood, and Benjamin Vernot. Lastly, I would like to thank my fellow members of the incoming graduate student class of 2010 for helping to make my time in graduate school fun and memorable. I hope our friendships endure long after we have all earned our degrees.

On the administrative side, I would like to thank the University of Washington's Genome Training Grant and the National Institutes of Health for their financial generosity. I would also like to thank the administrative coordinators Dawn Counts and Brian Giebel, without whose help there would almost certainly be a clerical error preventing me from graduating.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Opening remarks

This is a very exciting time for the biological sciences. In the 1950's it became clear that the nucleic acids, DNA and RNA, are the source of all genetic information. The 1970's and 1980's witnessed new technologies to sequence these nucleic acids, allowing us to examine this fundamentally digital information for the first time. These sequencing technologies became more and more powerful in the 1990's, as they culminated in the Human Genome Project, the largest biological project in history. The Human Genome Project was a smashing success, and on its heels came new sequencing instruments that provided orders of magnitude more genomic information than ever before. These technologies have made possible an entire new field of research.

Genomics – a word coined in 1986 – is the name of this new field. Genomics is the paradigmatic 21st-century science: a marriage of classical genetics and sequencing technology, with assistance from computer science and Big Data. And we are in the midst of a genomic revolution. Sequencing is so powerful and versatile that it has rapidly become an indispensable tool for anyone engaged in any line of biological inquiry. Not only geneticists but also cell biologists, biochemists, ecologists, evolutionary biologists, bioengineers, and most types of physicians are all coming to rely on sequencing technology for their assays.

Sequencing technology, as powerful as it may be, does nothing biological by itself. The output of a sequencing instrument is not a genome, but merely a long stream of words written in a four-letter alphabet. To a biologist, it is gibberish. The real power of genomics lies in converting this sequencing data into real information about an organism's genome, at which point it can be used to draw further biological inferences. This task, one of the preeminent challenges in genomics today, is called ***genome assembly***, and it is a fundamentally computational process. Genome assembly is the subject of my dissertation.

1.2 Organization of this dissertation

In this dissertation I describe several methods that I have developed for the assembly of genomes at various scales. In **Chapter 2** I give an overview of the history of genome sequencing and assembly methods, as they have gradually developed over the second half of the 20th century and the first decade of the 21st, ending up at the current status quo in the field. I then discuss my specific aims for how I plan to push the status quo forward with the work in my dissertation. In the next three chapters I detail three different computational problems that fall broadly into the category of genome assembly, and I present my solutions to these problems, all of which I have previously published in peer-reviewed journals. In **Chapter 3** I describe the problem of ***scaffolding***, a late step in the genome assembly process. I demonstrate a solution to this problem that utilizes Hi-C, a library preparation technique that was originally developed to study chromatin conformation but which I have repurposed for the task of genome assembly. In **Chapter 4** I extend the principles of Chapter 3 by applying Hi-C to solve another subtype of genome assembly problem, that of ***metagenomic deconvolution***. In **Chapter 5** I describe ***haplotype phasing***, a type of genome assembly problem, in the context of cancer genomes. I develop a novel algorithm and use it to phase the haplotypes in the genome of HeLa, a famous cancer cell line. In **Chapter 6** I attempt to place my work in the broader context of the rapidly moving field of genome sequencing and assembly, and I assess where the field is likely to be headed in the near future and what challenges still await it. Lastly, you may already have noticed that I have placed many genomic terms of art in ***bold italic*** lettering. In the hope of making my writing more accessible, I have defined all of these terms in a glossary at the end of this dissertation.

CHAPTER 2: THE HISTORY OF GENOME SEQUENCING AND ASSEMBLY

Note that terms written in ***bold italic*** have definitions available in the Glossary.

2.1 What is a genome?

The word ***genome*** was coined in 1920 by Hans Winkler, a German botanist, to describe the set of all genetic material in a cell[1]. Winkler's coinage was a portmanteau of ***gene*** and ***chromosome***, reflecting his understanding that both genes and chromosomes were involved in heredity. In 1920, however, neither genes nor chromosomes had a well-established molecular basis.

Many years passed before the molecular basis of genes was understood. In 1944 the Avery-McLeod-McCarty experiment[2] demonstrated, and in 1952 the Hershey-Chase experiments[3] confirmed, that ***DNA*** is the carrier of genetic material. In 1953 James Watson and Francis Crick proposed a structure for the DNA molecule, a double helix joined by hydrogen bonds between the four nucleotides, or ***bases***: adenine, cytosine, guanine, and thymine (A,C,G,T)[4]. The following year, George Gamow was the first to speculate that the sequence of A's, C's, G's, and T's on a strand of DNA might encode information in the form of "a long number written in [base 4]"[5]. It soon became clear that this sequence does indeed encode information – and therefore that genetic information, unlike most types of biological information, is fundamentally digital rather than analog. Fortuitously, this discovery came at a time when powerful computing technologies, capable of handling huge quantities of digital information, were just beginning to appear[6]. Genetics and computer science have been intertwined ever since. The genome is a biological entity, but it is also a computational entity.

2.2 The first genomes are sequenced

Having deduced that a genome is essentially "a long number", the next question scientists naturally sought to ask was, "What number is it?" In other words, how could the information-encoding sequence of bases in a nucleic acid molecule be assayed? This question could not be answered until decades later, when new methods for ***sequencing*** were explored. Walter Fiers developed RNA sequencing techniques and applied them to the RNA genome of the bacteriophage MS2, publishing the first-ever whole sequences of a gene (387 bases) in 1972[7] and of an entire genome (3,569 bases) in 1976[8]. Meanwhile, two different ***DNA sequencing*** techniques were developed independently by Frederick Sanger[9] and by Walter Gilbert and Allan Maxam.[10] In 1977 Sanger produced the first full sequence of a double-stranded DNA genome, that of bacteriophage ΦX174 (5,375 ***base pairs***, or bp)[11]; five years later he sequenced the much larger genome of phage λ (48,502 bp)[12]. Sanger's chain-termination method was adopted much more widely than the Maxam-Gilbert method due to its simplicity and its avoidance of hazardous reagents. So-called ***Sanger sequencing*** also proved to be very amenable to scaling up, and would become the dominant DNA sequencing technology for the next 30 years[13].

According to the central dogma of molecular biology[14], all of the RNA and protein molecules in a cell ultimately derive from information encoded in DNA. Thus, sequencing the genome of an organism is tantamount to creating a comprehensive catalog of the genetic and proteomic capacity of that organism, an incomparably valuable resource for further genetic and biological study. Sequencing the genomes of bacteriophages in the 1970's was an impressive technological accomplishment, but it was clearly only the beginning. Every organism has a genome, and in theory any genome could be sequenced. As Sanger sequencing became more parallelizable and automated, it became more cost-effective, and the idea of sequencing a complete cellular genome became a tantalizing prospect.

In the 1980's, genome sequencing projects were launched for several species. These projects took many years to bear fruit, but they helped to accelerate the pace of improvements in sequencing technology. The first genome to be completed was of the bacterium *Haemophilus influenzae* in 1995 (1.8 megabase pairs (Mbp))[15], followed within a year by the yeast model organism *Saccharomyces cerevisiae* (12.1

Mbp)[16] and the archaeon *Methanococcus jannaschii* (1.7 Mbp)[17]. All three domains of life were now ascertainable at the genomic level.

The ultimate promise of genome sequencing technology was to allow us to sequence the genome of any organism – even ourselves. The concept of a collaborative "human genome project" was first seriously discussed at a conference in 1985[18]. In 1986 it received approval from the United States Congress for public funding[19]. (It was that same year, at one of the early meetings to discuss the project, that Tom Roderick coined the word "genomics" to describe this emerging field[20].) The idea of sequencing a human genome seemed far-fetched at the time; after all, the haploid genome of a human (3 gigabase pairs (Gbp)) is six orders of magnitude larger than that of a small bacteriophage. But the potential benefit to science was incalculable and undeniable. James Watson said of the project: "It's essentially immoral not to get it done as fast as possible"[21]. The ***Human Genome Project*** launched in 1990 as a gargantuan multinational collaboration[22]. It was soon rivaled by a parallel, privately funded project headed by Craig Venter, whose goal was to release a human genome sequence as soon as possible[23]. Spurred by their competition, the two projects raced ahead of schedule and produced draft genomes nearly simultaneously in 2001[24,25].

The Human Genome Project is the largest and most costly biological project ever undertaken, but it has paid for itself many times over in pure economic terms[26], has been emulated as a model by all subsequent large-scale biological projects, and is regarded as a crowning scientific achievement of humanity[27]. In 2014 it is difficult to imagine how human biology or medicine could progress without the human genome sequence[28].

2.3 The genome assembly process

The term ***genome sequencing*** is confusingly imprecise: merely sequencing DNA molecules is not sufficient to produce a genome sequence! No established sequencing technique can be applied directly to entire chromosomes, due to the proteins bound to DNA and the inherent imprecision of sequencing

chemistry. Instead, DNA sequencing produces ***reads***: short contiguous stretches of nucleotide sequence that contain no information regarding their position in the genome. (One method that may someday be used to sequence long molecules of DNA at once is nanopore sequencing[29], but this technology has persistently failed to live up to expectations.) Sanger sequencing produces reads which are rarely more than a kilobase (Kb) in length; yet the human chromosome 2 contains 237 Mbp of euchromatic sequence[30]. How can the former possibly be converted into the latter?

The task of transforming a large number of DNA sequence reads from an organism into a prediction of that organism's genome is called ***genome assembly***. When this is done without any prior knowledge of sequences in the genome, it is called **de novo *assembly***, from the Latin *de novo*, meaning "from the beginning." (These terms refer to the process as well as to its end product.) Genome assembly is analogous to solving a one-dimensional jigsaw puzzle, in which the puzzle pieces may overlap and may be numerous enough to cover the puzzle many times over, but are much smaller than the entire puzzle. If there is no picture indicating what the puzzle should look like when complete, it is a *de novo* assembly.

Sanger and his collaborators performed the first *de novo* assemblies manually and hierarchically, by first assembling gene transcripts and then using read overlaps to merge the transcripts[11,12]. This approach was impractical with larger genomes, but fortunately, contemporary computing technology enabled the development of computer programs to handle the *de novo* assembly process. The first *de novo* assembly software was created by Roger Staden, who in 1979 published a semi-automated FORTRAN program[31] that performed the following three steps: (1) find read-read overlaps; (2) use these overlaps to determine the reads' relative positions; (3) resolve sequence discrepancies within the overlap to merge the reads into a single contiguous sequence, which Staden later termed a ***contig***[32]. As an aid in the consensus process, Staden also devised a quality system to denote the confidence of base calls in a sequence read[31], a system which was later codified in the concept of numerical ***quality scores***[33]. Software tools for *de novo* assembly soon proliferated, mostly following Staden's "overlap-layout-consensus" approach[34]. These tools relied on the principle of high ***coverage***: many copies of the same genome molecule were subjected to the sequencing chemistry to produce enough reads to cover the genome many times over, ensuring that overlaps could be found.

2.4 Draft genome assemblies and the challenge of contiguity

The per-base cost of Sanger sequencing fell 100-fold during the 1990's[24]. This dramatic drop was a consequence of great efforts to refine and automate Sanger sequencing and was a boon to the many large-scale genome projects ongoing at the time. However, it became apparent that simply creating more sequence reads was not sufficient to assemble a genome. Genomes are not random strings of bases: they contain certain repeated sequences, or ***repeats***, as a result of evolutionary history. Repeat elements can take many forms, including but not limited to: ***microsatellites*** (also called short tandem repeats or simple sequence repeats), short sequences (≤6 bp) repeated many times, often in heterochromatic regions[35]; ***mobile genetic elements***, sequences that move and replicate themselves within a genome[36]; and ***segmental duplications***, large blocks (≥1 Kb) of sequence duplicated with high identity (>90%)[37]. Larger genomes tend to contain more repeats; at least 50% of the human genome consists of repeats[24,38]. Repeats pose a serious problem for genome assembly because reads derived from different copies of a repeat will appear to overlap despite not being genomically contiguous.

The first step in a genome project is the creation and sequencing of a sequence ***library*** containing fragments of genomic DNA. The cheapest and most direct type of library is a ***shotgun sequencing*** library, which contains complete genomic DNA and thus represents the entire genome[39]. Shotgun libraries alone are sufficient to create a ***draft assembly***: a set of contigs which contains most of a genome's non-repetitive sequence content but is understood not to be a complete representation of the genome. A draft assembly created from shotgun sequencing typically lacks ***contiguity***, or long-range information regarding the relative position of sequences, due to repeats. Contiguity is destroyed by the fragmentation of the genome into sequence reads and must be reestablished to create a high-quality genome assembly. Any repeat that is longer than a read length and has near-perfect sequence identity will break the assembly's contiguity, because the non-repetitive genomic regions on either side of the different occurrences of the repeat cannot be conclusively linked to one another. Thus the assembly will be fragmented into small contigs[38]. In addition, highly identical copies of a repeat are often collapsed into

a single contig instead of appearing with the proper copy number[40]. More methods are needed to bridge repeats and establish genomic contiguity.

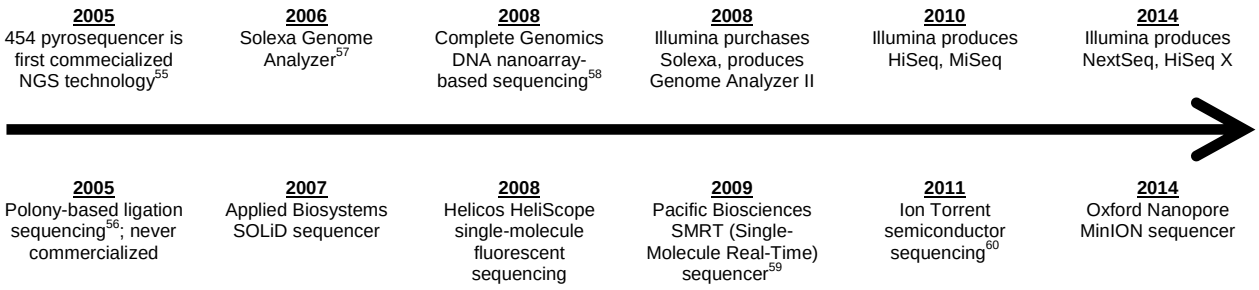
The Human Genome Project pioneered many methods of providing contiguity information, both "bottom-up" and "top-down". "Bottom-up" methods took a set of contigs produced by shotgun sequencing and assembly as input and determined which contigs were close to other contig. One such method was ***mate-pair sequencing*** (or paired-read sequencing), in which long templates of DNA are sequenced at both ends to produce a ***read pair***, such that the genomic distance between the two reads is equal to the length of the DNA template minus the reads' lengths[41]. Paired reads can be aligned onto contigs and used to convert the contigs into ***scaffolds***, which consist of many contigs in a defined order and orientation, separated by ***gaps*** whose sizes can be approximately calculated based on the length of the template library[42]. Most mate-pair libraries have template lengths on the order of 3 kb or less, allowing contiguity to that degree, but ***fosmid*** libraries have template lengths as much as 40 kb or greater[43]. Thus, problematic repeats can be left in gaps and the assembly's contiguity can be built around them. "Top-down" methods used by the Human Genome Project include so-called hierarchical shotgun sequencing, in which genomic DNA was cloned into bacterial artificial chromosomes (BACs)[44]; the genomic regions were assembled separately and then fitted into pre-existing physical and genetic maps of the human genome[45,46]. BAC and fosmid sequencing were effective methods, but they were expensive and laborious, and they did not become cheaper as sequencing costs fell. Craig Venter's team chose to avoid using BACs and instead used mate-pair sequencing to assemble microbial genomes[15,17]; they subsequently decided to apply this approach to the human genome as well in lieu of a "top-down" approach, provoking a lively debate with the public Human Genome Project[47,48].

The process of completing a draft genome assembly as thoroughly as possible is called ***finishing***. It is expensive and time-consuming and is rarely performed to completion. The human genome was not declared "finished" until 2004[49], and finishing the most difficult-to-assemble regions of the human genome is still an ongoing process[50,51].

2.5 The next-generation sequencing revolution

By the end of the Human Genome Project, it was clear that the increased cost-effectiveness of Sanger sequencing had been crucial to the success of the project. It was also clear that continuing improvements in sequencing technology could make possible a great variety of other biological projects[52,53]. Sanger sequencing itself would not deliver these improvements: its throughput had been fully optimized and further increases were only incremental. However, the clear financial and scientific benefits of ever-cheaper DNA sequencing inspired a flood of technical innovation. A large number of new technologies soon appeared, each vying to replace Sanger sequencing and become the new standard. These technologies are collectively referred to by the term ***next-generation sequencing***, or ***NGS***[54].

Figure 2.1: A timeline of advancements in NGS (next-generation sequencing) technology.



The NGS technologies are too numerous to describe in technical detail; see **Figure 2.1** for a timeline. These technologies have been developed by private companies that build and sell branded sequencing instruments and reagent kits. Each type of instrument produces reads with different lengths, error profiles, and limitations at a different price point, and is thus suitable for different biological purposes[61]. The most successful NGS company in recent years has been ***Illumina***, whose sequencing-by-synthesis technology has achieved truly staggering levels of throughput. Illumina's HiSeq X Ten machine, released in 2014, fulfills a long-standing goal of sequencing an entire human genome at 30X coverage (~100 billion bp) for less than $1000, albeit only in bulk[62]; this represents a 10,000-fold reduction in cost over seven years.

Illumina's sequencers have revolutionized genomics and made possible a great number of new biological inquiries[54,63], but they have precipitated new challenges in the area of genome assembly. Like many NGS technologies, Illumina's instruments produce ***short reads***, typically ≤100 bp. Short reads are sufficient for many sequencing applications, but they create two severe problems for assembly tools. First of all, they make the traditional overlap-layout-consensus assembly approach computationally impossible: when shotgun sequencing a genome with short reads, the reads must be incredibly numerous, but the computational power required to detect all read-read overlaps between N reads is on the order of N^2. Hence the advent of NGS technologies spurred the development of novel *de novo* assembly algorithms[64]. Most were based on a mathematical construct called the de Bruijn graph, which converts reads into strings of so-called k-mers. The use of de Bruijn graphs in assembly predates NGS[34] but it has proven to be enormously useful in designing assembly algorithms adapted to short reads[65]. One such algorithm, called ALLPATHS, was developed in part by me and was used to create the first *de novo* assemblies of mammalian-scale genomes solely from short reads[66,67].

An even greater problem with short reads is that assemblies based solely on them lack contiguity. No assembly algorithm could possibly use them to resolve the repeats present in large genomes. Indeed, short-read draft assemblies are observed to lack vast amounts of repeat sequence due to the collapsing of repeats[68]. Thus, although the NGS revolution has enabled a massive increase in the quantity of published genome assemblies, it has also led to a decline in their quality.

2.6 Beyond assembly: alignment, variant detection, and haplotype phasing

The Human Genome Project created a ***reference genome*** for *Homo sapiens*. A species' reference genome is an assembly that is understood to be fairly representative of that species. However, nobody's genome is identical to the human reference genome, and in fact no two individuals have the exact same genome sequence (not even identical twins[69]!) Finding and interpreting inter-individual genetic variation is a central task in the field of population genetics[70], as well as in fields of genomic medicine such as

pharmacogenomics[71]. ***Germline mutations*** give children genetic material that was not present in their parents, and may enter a population's gene pool, at which point they are called ***variants***. Furthermore, our cells acquire ***somatic mutations*** as they divide. Certain somatic mutations may encourage a cell to divide uncontrollably, leading to cancer, which is fundamentally a disease of the genome[72].

Hence, continued advancement in genomics and medicine requires investigations into the genomic differences between individuals and between cells. These investigations are made possible by our knowledge of reference genomes, and they are made economically feasible by the advent of NGS technologies. In a typical variant-detecting workflow, an individual is sampled, DNA is sequenced, and the reads are ***aligned*** to the reference genome. Sequence alignment methods allow for the possibility of mismatches between the read and the reference, and mismatches that appear consistently in the reads in a sample are called as variants[73].

The smallest types of variants are ***single-nucleotide variants*** (SNV) (a single base change, *e.g.*, A→C) and small ***indels*** (insertions or deletions of a small number of bases). SNVs and small indels are easily detected by NGS reads in the form of alignment mismatches, and there now exist many well-established algorithms to perform this detection[74,75]. However, variants also occur in larger forms. ***Structural variants*** (SVs) are a large class of variants encompassing ***copy number variants***, insertions of large sequences such as transposons, ***inversions***, ***translocations***, and chromosomal abnormalities. SVs affect far more bases in the human genome than do SNVs[76], and they are critically important in the etiology of a variety of heritable disorders as well as cancer[77]. However, SVs are difficult to discover with short reads because they are not easily localized. Methods exist to find every type of SV using short reads, but they often require deep sequencing and/or additional assays[78].

An additional challenge in genomic analysis is that most large organisms, including humans, are ***diploid***: they possess two copies of each chromosome. Each chromosome is a distinct molecule with its own ***haplotype*** – that is, its own set of variants that differentiates it from the reference genome. Variants may be ***homozygous*** (appearing on both of an individual's chromosomes) or ***heterozygous*** (appearing on only one). Even if every variant on both chromosomes is detected, additional information is required to determine which heterozygous variants belong on which haplotype. This information is called ***phase***, and

the act of deriving it is referred to as ***phasing*** of variants or haplotypes (also "resolution" of haplotypes, or simply haplotyping.) The phase of variants has important functional consequences, such as in cases of apparent compound heterozygosity, and phased haplotypes contain crucial information about population structure[79,80]. Short reads are insufficient to perform haplotype phasing; some additional source of contiguity information is necessary. In humans, this source may be *a priori* known population structure (statistical phasing or imputation)[81], but the haplotypes are more robust if derived directly from contiguity-preserving library construction methods (molecular phasing)[82].

2.7 Metagenome assemblies

Another challenge in 21st-century genomics is presented by ***microbial communities***. Every ecosystem on the planet, including our own bodies, contains communities of microbial species. However, most microbial species cannot be cultured individually and thus cannot be easily isolated for individual analysis by genome sequencing[83]. DNA can be isolated from a microbial community and sequenced, but all of the organisms' genomes are convoluted together in the sequence library, creating what is called a ***metagenome***[84]. A common approach in studying a metagenome is to subject it to shotgun sequencing and then assemble it, creating a *de novo* ***metagenome assembly***[85]. This is no easy task: if assembling one genome sequence is like solving a jigsaw puzzle, then assembling a metagenome is like taking puzzle pieces mixed together from many different jigsaw puzzles and then solving them all. Sequences that appear in multiple species will be indistinguishable, much like repeats; additionally, because the different species live at different abundances in the community, some genomes will have much greater coverage than others. New metagenome assembly software has recently been developed to address these challenges[86,87], but metagenome assemblies still tend to lack long-range contiguity, especially when short-read sequencing is used.

A metagenome assembly is convoluted: that is, it consists only of a set of contigs, with no *a priori* information about what species they belong to. It contains many genome assemblies, but they cannot be

separated from one another. This may prevent the discovery of novel or unculturable species that play important roles in their ecosystems[85]. Thus, ***metagenomic deconvolution*** – the reconstruction of individual species' genomes from metagenomes – is a pressing need, and has been called "the holy grail" of metagenomics[88]. Many computational methods are used to deconvolute metagenome assemblies, including analysis of base composition[89] and gene content[90], differential coverage binning[91], and comparisons to existing reference genomes[92]. However, many of these methods are powerless to identify species for which reference genomes do not already exist, and even the best of them only allow for modest improvements in metagenome contiguity. A method that could provide complete intra-species contiguity would assemble each species individually, effectively deconvoluting the metagenome.

2.8 My research aims

To provide much-needed contiguity in genome and metagenome assemblies, researchers have broadly taken one of three approaches. The first is to use next-generation long-read sequencing technologies such as Pacific Biosciences, either alone[93] or in conjunction with short reads[94]. The second is to isolate and sequence individual cells[95,96]. Both of these approaches are promising but not yet fully developed; moreover, both of them require great sequencing expenses. The third approach is to develop new types of short-read sequencing libraries that provide long-range contiguity information. **In this dissertation I describe novel computational methods that I have developed to exploit contiguity information in new sequencing libraries and create high-quality *de novo* genome assemblies from short reads.**

My first aim in this dissertation, as described in **Chapter 3** and in **Burton *et al*.**[97], is to design a new method of *de novo* genome scaffolding that exploits the low cost of short-read sequencing to create chromosome-scale scaffolds in a cost-effective fashion. My key insight is to repurpose Hi-C[98], a library construction method that was originally developed in 2009 to study the 3-D structure of genomes. It has been established that Hi-C gives a signal of long-range interactions that allow the creation of a 3-D model for a known genome. I show that Hi-C also provides a signal of short-range interactions that can piece

together a genome that is not yet fully assembled. I develop and release Lachesis, a software tool that applies this principle. I demonstrate the effectiveness of Lachesis by creating low-contiguity *de novo* assemblies of the human, mouse, and fly (*Drosophila*) genomes, then showing that Hi-C can scaffold these assemblies with over 99% accuracy. These assemblies, produced only from publicly available datasets, are the first ever chromosome-scale genome assemblies produced solely from short-read sequencing technologies.

My second aim is described in detail in **Chapter 4** and in **Burton, Liachko *et al.***[99]. Building off of my previous aim, I show that the Hi-C method provides yet another useful signal: intracellular linkage. When applied to a mixed cell population, Hi-C can distinguish which sequences belong together within a single cell, and thus within a single species – enabling metagenomic deconvolution. I develop MetaPhase, a software tool that uses a Hi-C library for this purpose. Using MetaPhase, I demonstrate that a single Hi-C library can deconvolute metagenomic mixtures of as many as 18 species, including representatives of all three domains of life (Bacteria, Archaea, and Eukarya), reconstructing the genomes of individual species with over 99% accuracy. I also show that a genome assembly created with MetaPhase can be further scaffolded using Lachesis, thus creating high-quality *de novo* genome assemblies from nothing but short-read metagenomic sequencing libraries. MetaPhase may make it possible to discover large numbers of novel and unculturable microbial species directly, by assembling their genomes.

My third and final aim, as described in greater detail in **Chapter 5** and in **Adey, Burton, Kitzman *et al.***[100], is to develop methods to phase haplotypes and resolve their copy numbers in a complex cell line. I demonstrate my methods on HeLa, a cancer cell line of singular historical significance[101,102]. My approach builds on the fosmid clone pool sequencing strategy of my colleague Jacob Kitzman[82] to achieve molecular phasing in long haplotype blocks. I adapt the ReFHap method of Duitama *et al*.[103] to create haplotype blocks with an N50 of 550 Kb. I then determine the so-called “haplotype-resolved copy number” of each haplotype in the HeLa genome, and I exploit the unevenness of this copy number to combine the haplotype blocks into haplotype scaffolds with an N50 of 44.8 Mb, many of which encompass entire chromosome arms. These haplotype scaffolds enable the phasing of the HeLa epigenome and the investigation of the oncogenic events in its history.

CHAPTER 3: CHROMOSOME-SCALE SCAFFOLDING OF *DE NOVO* GENOME ASSEMBLIES USING HI-C

This chapter is based on the following peer-reviewed publication[97]:

Joshua N. Burton, Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman and Jay Shendure. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119-1125 (2013).

Jacob Kitzman, Jay Shendure, and I conceived the idea of using Hi-C for *de novo* genome scaffolding. I designed the LACHESIS method (**Figure 3.1**) and wrote the LACHESIS software. Rupali Patwardhan and I performed the *de novo* assemblies. I applied LACHESIS to the human, mouse, and fruit fly genomes (**Figures 3.2, 3.3** and **Table 3.1**) as well as to the simulated genome assemblies (**Table 3.2**). Ruolan Qiu conducted the HeLa Hi-C experiments, and Andrew Adey analyzed the HeLa Hi-C data and called chromosomal rearrangements (**Figure 3.4**). Andrew Adey, Jay Shendure, and I prepared the manuscript, with input from all authors. Jay Shendure supervised the study. The Lachesis software was written in C++ and Perl and is available online at https://github.com/shendurelab/LACHESIS.

3.1 Summary

Genomes assembled *de novo* from short reads are highly fragmented relative to the finished chromosomes of *Homo sapiens* and key model organisms generated by the Human Genome Project. To address this problem, we need scalable, cost-effective methods to obtain assemblies with chromosome-scale contiguity. Here we show that genome-wide chromatin interaction data sets, such as those generated by Hi-C, are a rich source of long-range information for assigning, ordering and orienting genomic sequences to chromosomes, including across centromeres. To exploit this finding, we developed an algorithm that uses Hi-C data for ultra-long-range scaffolding of *de novo* genome assemblies. We demonstrate the approach by combining shotgun fragment and short jump mate-pair sequences with Hi-C data to generate chromosome-scale *de novo* assemblies of the human, mouse and *Drosophila* genomes, achieving—for the human genome—98% accuracy in assigning scaffolds to chromosome groups and 99% accuracy in ordering and orienting scaffolds within chromosome groups. Hi-C data can also be used to validate chromosomal translocations in cancer genomes.

3.2 Introduction

The Human Genome Project defined and achieved high standards for the *de novo* assembly of reference genomes for *H. sapiens* and key model organisms. For example, the public draft human genome, reported in 2001, contained 90% of the euchromatic sequence with an N50 (defined as the length *L* at which 50% of sequence is in contigs of length ≥*L*) of 82 kilobases (Kb)[24,49]. The finished human genome, reported in 2004, contained 99% of the euchromatic sequence with an N50 of 38.5 megabases (Mb) and an error rate of 1 event per 100,000 bases[49]. At both stages, nearly all sequences were assigned, ordered and oriented to chromosomes, although many errors were corrected during finishing[49].

Massively parallel DNA sequencing technologies produce billions of short reads per instrument run at a very low cost per sequenced base, empowering a wide range of experiments[54,63]. However, although

extensive progress has been made in developing algorithms for *de novo* genome assembly from short reads[65], we remain remarkably distant from routinely assembling genomes to the standards set by the Human Genome Project. For example, the human genome was assembled with <40 gigabases (Gb) of Sanger sequencing, but *de novo* assemblies of short reads relying on five- to tenfold more sequence are highly fragmented relative to the finished chromosomes of the *H. sapiens* reference build[67,104].

It is important to recognize that the high quality of the Human Genome Project's genome assemblies is not solely attributable to the length and accuracy of Sanger sequencing reads. Rather, a diversity of approaches was brought to bear to achieve long-range contiguity. For the human genome, this included dense genetic maps, dense physical maps and hierarchical shotgun sequencing of a tiling path of long insert clones[24,49]. Whole-genome shotgun assemblies—typically based on end sequencing of both short and long insert clones—also relied on dense genetic and physical maps to assign, order and orient sequence contigs or scaffolds to chromosomes[105].

Diverse strategies have been developed to boost the contiguity of *de novo* genome assemblies from short reads. These include end sequencing of fosmid clones[67], fosmid clone dilution pool sequencing[82,106], optical mapping[107-110] and genetic mapping with restriction site–associated DNA tags[111]. However, each of these strategies has important limitations. Fosmid libraries and optical mapping are technically challenging and provide only mid-range contiguity. Genetic maps are more powerful but are costly or impractical to generate for many species. Particularly as initiatives such as the 10K Genome Project[112] gain momentum, the genomics field is in need of scalable, broadly accessible methods enabling chromosome-scale *de novo* genome assembly.

Hi-C and related protocols use proximity ligation and massively parallel sequencing to probe the three-dimensional architecture of chromosomes within the nucleus, with interacting regions captured to paired-end reads[98, 113]. In the resulting data sets, the probability of intrachromosomal contacts is on average much higher than that of interchromosomal contacts, as expected if chromosomes occupy distinct territories. Moreover, although the probability of interaction decays rapidly with linear distance, even loci separated by >200 Mb on the same chromosome are more likely to interact than loci on different chromosomes[98].

We speculated that genome-wide chromatin interaction data sets, such as those generated by Hi-C, might provide long-range information about the grouping and linear organization of sequences along entire chromosomes. In exploring this, we developed LACHESIS (ligating adjacent chromatin enables scaffolding *in situ*), a computational method that exploits the signal of genomic proximity in Hi-C data sets for ultra-long-range scaffolding of *de novo* genome assemblies. LACHESIS works in three steps (**Figure 3.1**)—first, clustering contigs or scaffolds to chromosome groups; second, ordering contigs or scaffolds within each chromosome group; and finally, assigning relative orientations to individual contigs or scaffolds. We demonstrate the effectiveness of this approach by combining shotgun fragment and short insert mate-pair (<3 Kb) sequences with Hi-C data to generate reasonably accurate chromosome-scale *de novo* assemblies of the *H. sapiens, Mus musculus* and *Drosophila melanogaster* genomes. We also show that Hi-C data can be used to validate chromosomal rearrangements in cancer genomes.

3.3 Results

3.3.1 Exploiting contact probability maps for de novo *genome assembly*

The input to LACHESIS consists of a set of contigs or scaffolds (the term 'contig' is used in this description of the method to indicate both possibilities), such as are generated by de Bruijn graph–based *de novo* assemblers[65,67], and a genome-wide chromatin interaction data set, such as is generated by Hi-C and related protocols[98,113]. The Hi-C reads are aligned to the contigs, and the number of Hi-C read-pairs linking each pair of contigs is tabulated (**Figure 3.1a**). In a first step, LACHESIS uses hierarchical agglomerative clustering to group contigs that are likely to derive from the same chromosome, exploiting the fact that intrachromosomal contacts are on average more probable than interchromosomal contacts in Hi-C data sets[98] (**Figure 3.1b** and Figure A.3.1). An average-linkage metric[114] is used for this clustering, with linkage defined as the normalized density of Hi-C read-pairs linking any given pair of contigs. The final number of groups is prespecified, ideally set to the expected number of chromosomes.

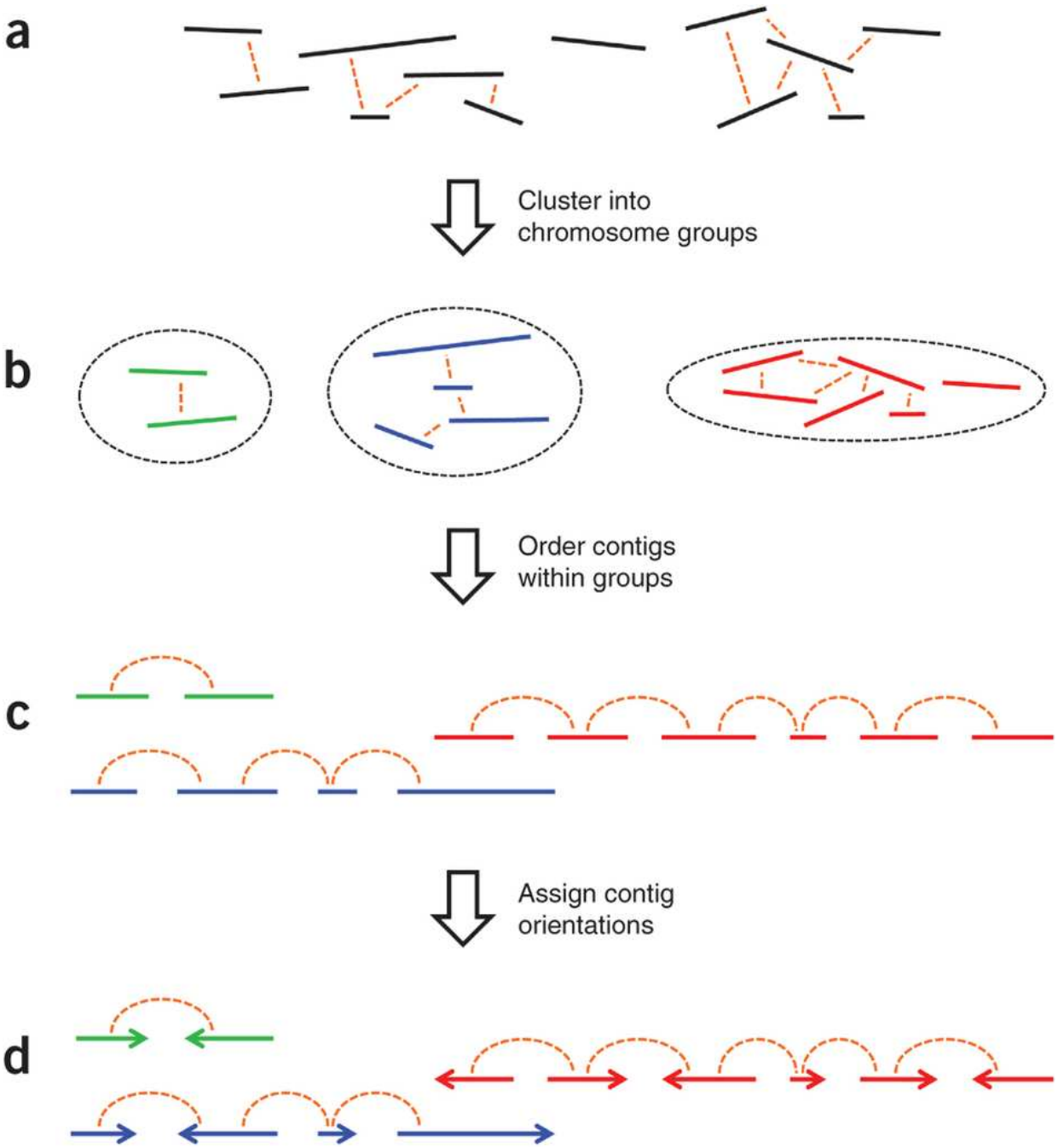


Figure 3.1: The LACHESIS scaffolding method. (**a**) The input consists of a set of contigs (or scaffolds) from a draft assembly and a set of genome-wide chromatin interaction data, for example, Hi-C links. (**b**) Contigs on the same chromosome tend to have more Hi-C links between them, relative to contigs on different chromosomes. LACHESIS exploits this to cluster the contigs into groups that largely correspond to individual chromosomes. (**c**) Within a chromosome, contigs in close proximity tend to have more links than contigs that are distant. LACHESIS exploits this to order the contigs within each chromosome group. (**d**) Lastly, LACHESIS uses the exact position of links between adjacent contigs to predict the relative orientation of each contig.

In a second step, LACHESIS orders contigs linearly within each chromosome group by taking advantage of the higher Hi-C link densities expected between closely located contigs (**Figure 3.1c** and Figure A.3.2). For each chromosome group, a graph is built with vertices representing contigs and edge weights corresponding to the inverse of the normalized Hi-C linkage density between pairs of contigs. A minimum spanning tree is found in this graph, and the longest path in the tree is extracted as the 'trunk', an incomplete but high-confidence ordering of contigs within each chromosome group. To generate a full ordering, contigs excluded from the trunk are reinserted into it at sites that maximize the amount of linkage between adjacent contigs.

In a third step, the ordered contigs are oriented with respect to one another by taking into account precisely where the Hi-C reads map on each contig (**Figure 3.1d** and Figure A.3.3). For each chromosome group, a weighted, directed, acyclic graph is built representing all possible ways to orient the contigs, given the predicted order. The weights are calculated as the log-likelihood of the observed Hi-C links between adjacent contigs in a given combined orientation, assuming[98] that the probability of a link connecting two reads at a genomic distance of x decays as 1/x for $x \geq$ ~100 Kb. The maximum likelihood path through this graph yields a predicted orientation for each contig.

3.3.2 Chromosome-scale assembly of mammalian genomes

We sought to evaluate the effectiveness of this approach for the chromosome-scale *de novo* assembly of mammalian genomes. We focused on human and mouse as test cases because of the availability of the necessary data sets and the high quality of these reference genomes as gold standards for comparison. For human, we used ALLPATHS-LG to assemble previously generated[67] shotgun fragment and short jump (~2.5 Kb) mate-pair sequences to an N50 scaffold length of 437 Kb and a total length of 2.74 Gb. We refer to this below as the 'shotgun assembly'. We intentionally excluded fosmid end sequencing data[67] because libraries of this type require cloning and are laborious to generate. Furthermore, we hoped that the chromatin interaction data would effectively substitute for the ~40 Kb fosmid links while also providing even longer-range contiguity.

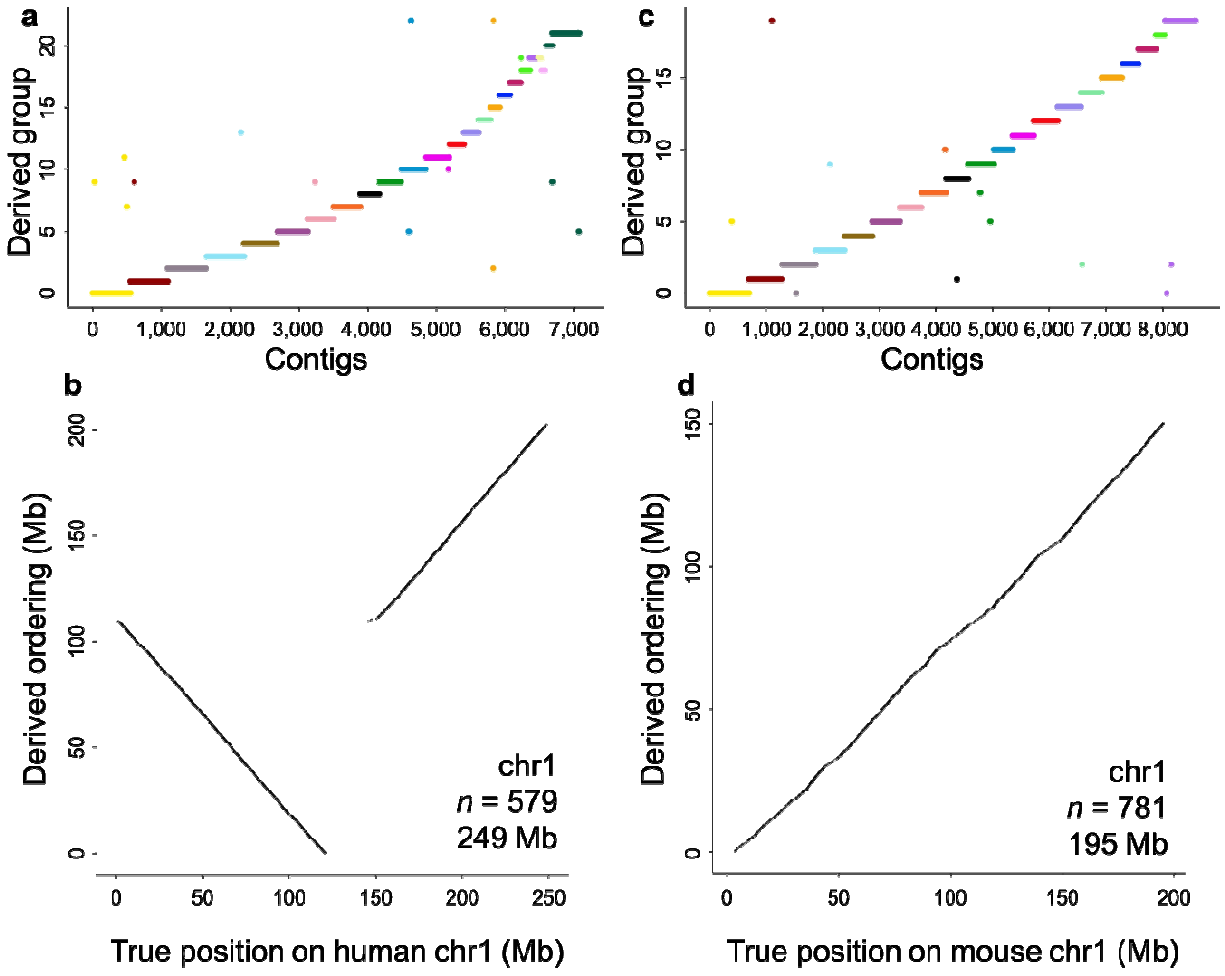
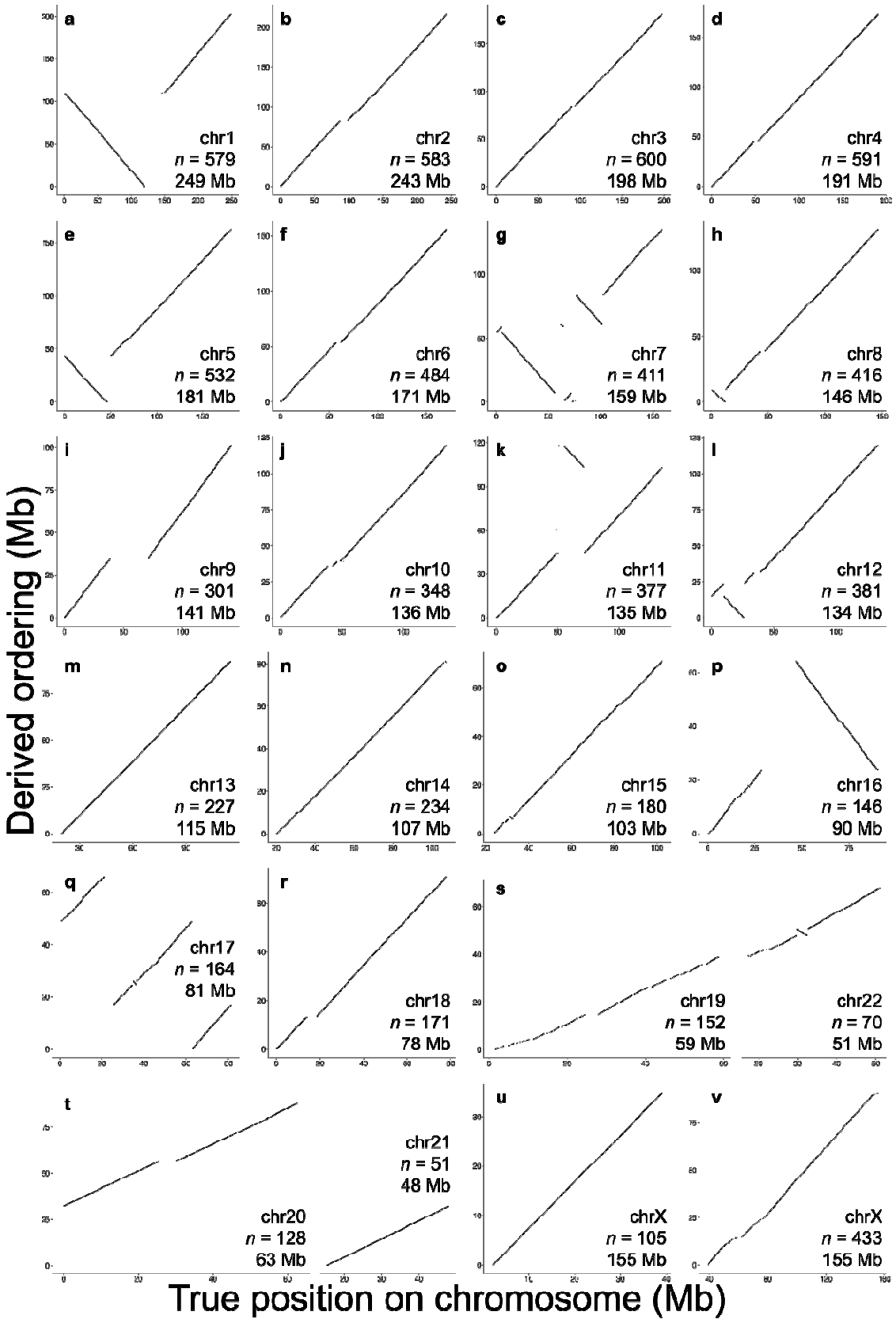


Figure 3.2: Clustering and ordering mammalian sequences with LACHESIS. (**a**) The results of LACHESIS clustering on the *de novo* human assembly. Shown on the *x* axis are the 7,083 scaffolds (total length, 2.49 Gb) that are large (≥25 AAGCTT restriction sites) and not repetitive (Hi-C link density less than 2 times average), which LACHESIS uses as informative for clustering. The *y* axis shows the 23 groups created by LACHESIS, with the order chosen for the purposes of clarity. The color scheme is the standard SKY (spectral karyotyping) color scheme for human. (**b**) The results of LACHESIS ordering and orienting of 579 scaffolds within the group from **a** corresponding to human chromosome 1. On the *x* axis is the true position of these scaffolds along human chromosome 1. On the *y* axis is the order in which LACHESIS has placed these scaffolds. Also listed in the panel are the chromosome name, the number of scaffolds in the derived ordering and the reference length of this chromosome. (**c**) The results of LACHESIS clustering on the *de novo* mouse assembly. Shown on the *x* axis are the 8,594 scaffolds (total length, 1.94 Gb) that are large and not repetitive, which LACHESIS uses as informative for clustering. The *y* axis shows the 20 groups created by LACHESIS, with the order chosen for the purposes of clarity. The color scheme is as in **a**. (**d**) The results of LACHESIS ordering and orienting of 781 scaffolds within the group from **c** corresponding to mouse chromosome 1. The plotting is as in **b**.


a
chr1
n = 579
249 Mb
b
chr2
n = 583
243 Mb
c
chr3
n = 600
198 Mb
d
chr4
n = 591
191 Mb
e
chr5
n = 532
181 Mb
f
chr6
n = 484
171 Mb
g
chr7
n = 411
159 Mb
h
chr8
n = 416
146 Mb
i
chr9
n = 301
141 Mb
j
chr10
n = 348
136 Mb
k
chr11
n = 377
135 Mb
l
chr12
n = 381
134 Mb
m
chr13
n = 227
115 Mb
n
chr14
n = 234
107 Mb
o
chr15
n = 180
103 Mb
p
chr16
n = 146
90 Mb
q
chr17
n = 164
81 Mb
r
chr18
n = 171
78 Mb
s
chr19
n = 152
59 Mb
chr22
n = 70
51 Mb
t
chr20
n = 128
63 Mb
chr21
n = 51
48 Mb
u
chrX
n = 105
155 Mb
v
chrX
n = 433
155 Mb
Derived ordering (Mb)
True position on chromosome (Mb)

(Previous page) Figure 3.3: LACHESIS ordering of scaffolds in a *de novo* human assembly. (**a**–**v**) The results of LACHESIS ordering and orienting on 22 of the 23 chromosome groups in the *de novo* human assembly. For each ordering, only the scaffolds on the dominant chromosome (the chromosome containing the plurality of aligned sequence) are shown. The exceptions are two groups that correspond to fusions of small chromosomes (19 and 22 (**s**); 20 and 21 (**t**)) (Table A.2.2). Within each of these fused groups, the two chromosomes were well separated by ordering (**s**,**t**). The X chromosome clustered into two separate groups (**u**,**v**). Not shown: one very small chimeric group (length, 6.5 Mb; Figure A.3.4w). Also listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering and the reference length of the dominant chromosome.

Metric	*De novo* assemblies		
	Human	Mouse	*Drosophila*
Shotgun assembly metrics			
Total assembly length, including gaps (Mb)	2,739	2,370	127
Number of contigs or scaffolds	18,921	25,964	7,109
N50 contig or ungapped scaffold size (Kb)	437	224	68
Clustering			
% sequence (% contigs) clustered into groups	98.2 (71.5)	98.0 (87.8)	81.2 (64.3)
% clustered sequence (% contigs) mis-clustered	0.14 (1.4)	0.24 (0.5)	3.4 (10.5)
Ordering			
% clustered sequence (% contigs) ordered	94.4 (55.3)	86.7 (42.7)	82.0 (24.5)
% ordered sequence (% contigs) w/ordering errors	0.5 (0.8)	0.5 (1.1)	4.6 (5.2)
% ordered sequence (% contigs) w/orientation errors	1.2 (2.5)	1.9 (4.6)	4.1 (6.1)
High-quality predictions			
% ordered sequence (% contigs) w/high quality	92.8 (79.0)	93.3 (82.9)	94.1 (88.1)
% high-quality sequence (% contigs) w/ordering errors	0.3 (0.4)	0.3 (0.7)	3.3 (3.4)
% high-quality sequence (% contigs) w/orientation errors	0.4 (0.5)	0.5 (1.0)	2.5 (2.7)

Table 3.1: Metrics for LACHESIS-based scaffolding of shotgun assemblies. The human and mouse shotgun assemblies are based on read-pairs from short-insert and ~2.5 Kb jumping libraries, whereas the *Drosophila* shotgun assembly is based solely on read-pairs from short-insert libraries[67]. The human and mouse shotgun assemblies consist of scaffolds, whereas the *Drosophila* shotgun assembly consists of contigs. LACHESIS places scaffolds or contigs into groups and then orders and orients them within each group. An ordering error means that a contig or scaffold's position is out of the expected order with respect to its neighbors. An orientation error means that its orientation is not the orientation implied by its position with respect to its immediate predecessor. 'High-quality predictions' refers to a subset of contigs or scaffolds whose position and orientation in their ordering is deemed more certain; the threshold for high quality is chosen for convenience for each assembly.

After aligning Hi-C read-pairs from a human male embryonic stem cell (ESC) line[115] to this shotgun assembly, we applied LACHESIS to cluster the scaffolds into 23 chromosome groups (the libraries used to generate the shotgun assembly were derived from female DNA[67]), and then to order and orient the scaffolds within each chromosome group (**Figures 3.2 and 3.3**, **Table 3.1**, Figure A.3.4, Table A.2.1, and Table A.2.2). Most scaffolds (*n* = 13,528, comprising 98.2% of the length of the shotgun assembly) were clustered into one of the 23 groups (**Figure 3.2a**). Nearly all of these groups corresponded to individual chromosomes, with the exceptions of the X chromosome, whose two arms were split in separate groups (**Figure 3.3u,v**): one chimeric group containing very little sequence from many chromosomes (6.5 Mb total; Figure A.3.4w); chromosomes 19 and 22, which were 'fused' into a single group (**Figure 3.3s**); and chromosomes 20 and 21, also fused into a single group (**Figure 3.3t**). The fusions are probably due to the greater density of interchromosomal links observed between short chromosomes in Hi-C data[98,116]. Apart from these errors, 98.6% of clustered scaffolds (comprising 99.86% of their sum length) were correctly grouped (**Table 3.1**), suggesting that Hi-C data are highly informative for the clustering of sequences derived from individual chromosomes, including across centromeres.

Within each chromosome group, the vast majority of the length of the clustered scaffolds was successfully ordered and oriented by LACHESIS (94.4% or 2.55 Gb; **Table 3.1**). The predicted orderings are highly concordant with the reference human genome (GRCh37), including across most megabase-scale centromere gaps, except for the occasional rearrangement of large segments within which nearly all scaffolds were well-ordered (**Figure 3.3** and Figure A.3.4). For example, scaffolds corresponding to the long and short arms of chromosome 1 are grouped together and, respectively, very well-ordered, but the reconstructed arms are joined incorrectly (**Figure 3.2b**). To quantify local errors, we defined ordering errors as instances where a contig or scaffold is not in the expected order with respect to its immediate neighbors, and orientation errors as instances where a contig or scaffold is not in the expected orientation implied by its immediate predecessor in the ordering. By these definitions, 99.2% of clustered scaffolds, representing 99.5% of the sum length, were correctly ordered; 97.5% of clustered scaffolds, representing 98.8% of the sum length, were correctly oriented.

Most ordering errors involve the inversion of local segments consisting of one or several contiguous scaffolds (Figure A.3.4). Compared to correctly ordered scaffolds, incorrectly ordered scaffolds are short and are enriched for segmental duplications and simple repeats (Figure A.3.5 and Table A.2.3). This suggests that complexities in the primary sequence are the source of many ordering errors, possibly through inaccuracies in the shotgun assembly or by confounding the mapping of Hi-C read-pairs. Other errors appear to be associated with the nonuniform distribution of biological interactions, for example, chromatin domains at various scales (Figure A.3.6). To address this in part, we calculated a quality score for ordering and orientation, defined as the relative log-likelihood of a contig's predicted orientation to its opposite orientation in the weighted directed acyclic graph. Local accuracy was better for scaffolds with high quality scores (**Table 3.1**). For scaffolds with high quality scores occurring within the assembly trunk, which comprise 2.09 Gb or 76.4% of the overall shotgun assembly, 99.9% of sequence is correctly ordered and 99.7% correctly oriented (Table A.2.1).

We also attempted the chromosome-scale *de novo* assembly of the mouse genome by an identical approach. We first used ALLPATHS-LG to assemble previously generated[67] shotgun fragment and short jump (~2.2 Kb) mate-pair sequences to an N50 scaffold length of 224 Kb and a total length of 2.37 Gb. After aligning Hi-C read-pairs from a mouse ESC line[115] to this shotgun assembly, we applied LACHESIS to cluster the scaffolds into 20 chromosome groups, and then to order and orient the scaffolds within each chromosome group (**Figure 3.2c,d**, **Table 3.1**, Figure A.3.7, Table A.2.1, and Table A.2.4). Most scaffolds (*n* = 22,802, comprising 98.0% of the length of the shotgun assembly) were clustered into one of the 20 groups (**Figure 3.2c**). There was a clear 1-to-1 correspondence between these groups and bona fide chromosomes (GRCm38), although a small part of mouse chromosome 10 (2.6 Mb) was erroneously clustered with chromosome 8 (Table A.2.4). Of the clustered scaffolds, 99.5% (comprising 99.76% of their sum length) were correctly grouped (**Table 3.1**). The majority of the length of the clustered scaffolds was ordered and oriented by LACHESIS (86.7% or 2.02 Gb; **Table 3.1**). Almost all (98.9%) of clustered scaffolds, representing 99.5% of the sum length, were correctly ordered; 95.4% of scaffolds, representing 98.1% of the sum length, were correctly oriented. Overall, the results for chromosome-scale *de novo* assembly of the mouse and human genomes are highly consistent.

3.3.3 Chromosome-scale assembly of the fruit fly genome

To further evaluate the generality of this method, we next applied it to the *de novo* assembly of *Drosophila*, for which a high-quality reference genome is also available as a gold standard for comparison. We first used ALLPATHS-LG[67] to assemble shotgun fragment sequences[117] (without jumping libraries) to an N50 contig length of 68 Kb and a total length of 127 Mb. We then aligned Hi-C read-pairs derived from *Drosophila*[118] to this shotgun assembly and used LACHESIS to cluster the contigs into four chromosome groups. Most contigs (81.2% of the length of the shotgun assembly) were clustered into one of the four groups (Figure A.3.8). This proportion is lower than that for the assemblies described above (≥98% for human and mouse), most likely because of the lower contiguity of the shotgun assembly (N50 contig size of 68 Kb for *Drosophila* versus N50 scaffold size of 437 Kb and 224 Kb for human and mouse, respectively). Nonetheless, the four groups corresponded well to the four *Drosophila* chromosomes (X, 2, 3 and 4), even though chromosome 4 is minuscule compared to the others (1.4 Mb or ~1% of the reference genome). Of the clustered scaffolds, 89.5% (comprising 96.6% of their sum length) were correctly grouped.

We then applied LACHESIS to order and orient the *Drosophila* contigs within each of the four chromosome groups (Table A.2.5 and Figure A.3.9). A lower proportion of the shotgun assembly was ordered (82.0% by length for fly versus 94.4% for human), again likely because the *Drosophila* assembly has shorter contigs than the mammalian shotgun assemblies used above. The predicted order corresponded well with the actual order based on contig alignments to the *Drosophila* reference genome (FB2013_02, euchromatic sequences only), and the right and left arms of chromosomes 2 and 3 were well separated (Figure A.3.9). Once again, a subset of the chromosome groups contained rearrangements of large segments within which nearly all contigs were well ordered. At a local scale, 94.8% of clustered contigs (95.4% of sum length) were correctly ordered, and 93.9% of clustered contigs (95.9% of sum length) were correctly oriented (**Table 3.1**).

Metric	Simulated contig size						
	10 Kb	20 Kb	50 Kb	100 Kb	200 Kb	500 Kb	1 Mb
Number of contigs	309,579	154,794	61,927	30,970	15,489	6,206	3,113
% sequence clustered into groups	30.1	74.2	91.9	92.7	92.9	93.1	93.4
% clustered sequence mis-clustered	1.6	0.47	0.41	0.46	0.66	0.66	0.26
% clustered sequence ordered	48.5	79.9	98.9	99.8	99.97	99.93	99.98
% ordered sequence w/ordering errors	37.2	18.0	4.4	2.2	1.4	0.8	0.8
% ordered sequence w/orientation errors	44.8	28.7	7.7	2.6	1.2	0.8	0.7

Table 3.2: Metrics for LACHESIS-based scaffolding of simulated assemblies. Simulated assemblies were created by breaking up the human reference genome into simulated contigs of varying sizes, and then using LACHESIS to cluster, order and orient the simulated contigs. The simulated contigs' expected order and orientation are derived from their true position in the reference genome. Ordering and orientation errors are defined as in **Table 3.1**.

3.3.4 Robustness to contig size and Hi-C data quantity

Our results for chromosome-scale scaffolding of the human, mouse and fly genomes were based on initial *de novo* assemblies with reasonably high N50s, that is, 437 Kb, 224 Kb and 68 Kb, respectively. To evaluate the power of this approach as a function of the contiguity of this initial assembly, we sought to reassemble simulated contigs of varying size derived from the human reference genome. In each iteration, we split the reference genome into equally sized contigs (10, 20, 50, 100, 200, 500 or 1,000 Kb) and mapped Hi-C read-pairs[115] to these simulated shotgun assemblies. We then used LACHESIS to cluster, order and orient the simulated contigs (results for 100-Kb simulated contigs are shown in Figure A.3.10 and Figure A.3.11). The performance of the method with respect to completeness and local accuracy is robust above an initial N50 of 50 Kb, but degrades rapidly below this point (**Table 3.2**).

In a separate analysis, we down-sampled the sequencing depth of Hi-C data and attempted chromosome-scale scaffolding of the human shotgun assembly (N50 = 437 Kb; Table A.2.6). Although clustering is robust to marked reductions in the amount of Hi-C data, accurate ordering and orienting of scaffolds within chromosome groups requires ~400 million read-pairs. Nonetheless, we note that even the

full amount of Hi-C data used here is <20% of the amount of sequencing data used to generate the initial shotgun assembly (59 Gb versus 303 Gb).

3.3.5 Validating translocations in cancer genomes

We also speculated that the strong intrachromosomal signal observed in Hi-C data might enable the global discovery or validation of interchromosomal rearrangements in cancer genomes, many of which are challenging to detect with methods other than karyotyping because the breakpoints occur in repetitive regions. For example, recent studies combined several mate-pair sequencing strategies to detect rearranged marker chromosomes in the aneuploid HeLa cancer cell line[100,119], but such methods were only successful for a small proportion of rearrangements, and for none of the rearrangements involving centromeric sequences. Of note, the 4C method was previously used to detect chromosomal breakpoints in cancer genomes, but in a targeted rather than global fashion[120].

To test this, we constructed a Hi-C library from HeLa cells and sequenced it to high depth (154 M unique read-pairs). These data were mapped and used to generate a matrix of pairwise link densities between windows of length 1 Mb along the human reference genome. Visual examination of the matrix revealed off-diagonal patches of strong linkage with asymmetric decay, consistent with interchromosomal rearrangements (**Figure 3.4**). Most of these corresponded well to previously described marker chromosomes[121], although we also observed strong evidence for two novel marker chromosomes (der(2;7)(q36;q10), “U1” and der(3;20)(q25;q10), “U2”). We implemented a rearrangement-calling method that successfully identified all of the suspected marker chromosomes, albeit with limited specificity (Figure A.3.12). Using chromatin interaction data in this way may enable the validation of candidate chromosomal rearrangements or the detection of chromosomal rearrangements in heterogeneous cancer cell populations that might not be detected by karyotyping of limited numbers of cells.

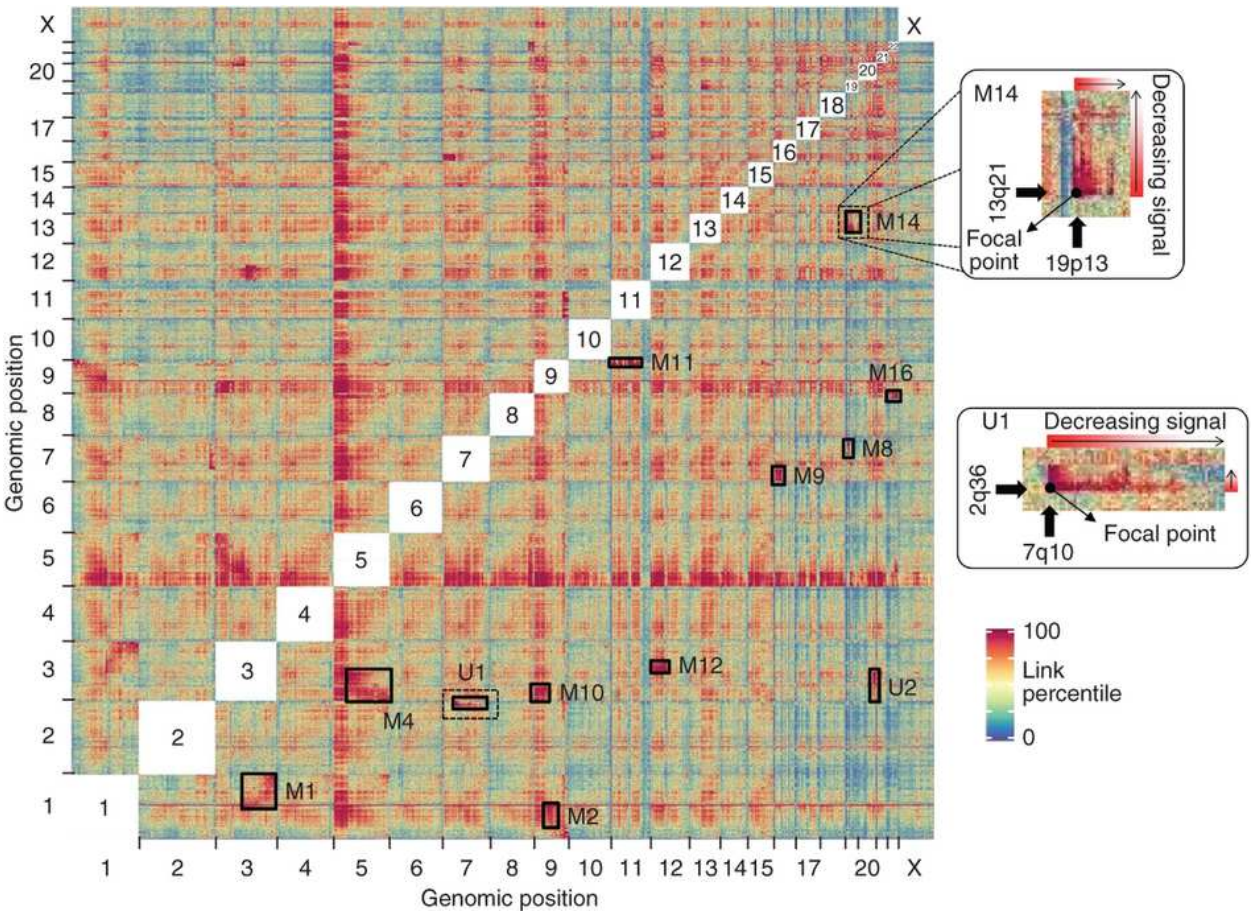


Figure 3.4: Detection of chromosome fusions in HeLa S3 using Hi-C data. Normalized interchromosomal links for a HeLa S3 Hi-C library between megabase windows were derived as described in Online Methods and are represented as an all-by-all heatmap. For visualization purposes, link weights are ranked and converted to a percentile. Previously identified marker chromosomes were identified (M1, M2, M4, M8, M9, M10, M11, M12, M14 and M16) as well as two additional peaks representing previously undescribed marker chromosomes (U1: der(2;7)(q36;q10) and U2: der(3;20)(q25;q10)). Two rearrangements are highlighted (M14 and U1) to demonstrate the signal focal point at the location of the fusion event with asymmetrical signal decay outward in the direction of the sequence contained in the chromosome fusion, thus allowing breakpoint identification as well as orientation.

3.4 Conclusions

Here we demonstrate that genome-wide chromatin interaction data sets, such as those generated by Hi-C, are a rich source of long-range information for assigning, ordering and orienting genomic sequences to chromosomes, including across megabase-scale centromere gaps, as well as for validating chromosomal

translocations in cancer genomes. There are a number of avenues for the potential improvement of this approach, both experimentally and computationally.

Although the experimental methods for Hi-C are straightforward, current protocols require a large amount of material (10^6–10^8 cells). As such, reducing the input requirements is an important technical goal. To date, global chromatin interaction data sets have been generated on organisms including yeast[113], human[98,115,116], mouse[115], fruit fly[118] and *Arabidopsis thaliana*[122]. This is consistent with broad applicability, but demonstrating these protocols on an even more diverse range of organisms is imperative. On a related point, as the success of this approach depends on chromosomes occupying distinct territories in the nucleus, it will be important to further validate LACHESIS in diverse species to confirm that this is ubiquitously the case. We also note that using multiple restriction enzymes (or developing new methods that avoid restriction digestion altogether and/or operate on purified high-molecular-weight genomic DNA) will likely improve performance, particularly for smaller contigs or scaffolds. Along the same lines, even if this approach broadly enables chromosome-scale scaffolding, the contiguity required for the initial *de novo* assembly (~50 Kb) may be challenging to achieve for many organisms. As such, there will remain a strong need for methods delivering 'intermediate' contiguity information in a highly cost-effective and scalable manner.

Computationally, a substantial limitation of our current algorithm is that the clustering step requires the number of chromosomal groups to be specified *a priori*. We assessed whether the scoring metric used during clustering enables reliable inference of chromosome number, but it does not (Figure A.3.13). One potential solution is to order contigs or scaffolds before determining chromosome groups, but this is computationally difficult with large numbers of contigs or scaffolds. Alternatively, statistical methods for predicting the optimal number of clusters may prove useful[123,124].

Ordering and orientation errors were associated with short scaffolds, segmental duplications and simple repeats (Table A.2.3). It is possible that our full exclusion of ambiguously mapping reads may be introducing 'gaps' in contiguity information that increase the probability of errors in such regions. Alternatively, these errors may be secondary to flaws in the initial shotgun assembly. Consistent with the latter, we also ran LACHESIS on a human 'shotgun assembly' that has higher contiguity because it used

fosmid end-pair data[67] (N50 scaffold length 11.5 Mb versus 437 Kb). We achieved chromosome-scale scaffolding of this assembly as well, but with lower accuracy owing to a small fraction of incorrectly joined scaffolds in the input to LACHESIS (Table A.2.1). This suggests that conservative *de novo* assembly before using chromatin interaction mapping for long-range scaffolding may be optimal. Lastly, we note that our use of chromatin interaction data for long-range scaffolding (by LACHESIS) was entirely separate from the initial assembly of contigs/scaffolds (by ALLPATHS-LG). We anticipate that a more integrated approach might improve accuracy.

Starting from shotgun human and mouse genome assemblies, each consisting of tens of thousands of scaffolds, we were able to cluster nearly all scaffolds into groups that overwhelmingly corresponded to individual chromosomes. A high fraction of these assignments were correct (comprising >99% of the sum length of clustered scaffolds). We were further able to order and orient contigs within each chromosome group, including scaffolding across megabase-scale centromere gaps, with surprisingly few errors. As such, we achieved reasonably accurate *de novo* mammalian genome assemblies with chromosome-scale contiguity using just three types of libraries, all generated by *in vitro* methods and sequenced as short read-pairs on a single platform (for human, shotgun fragment (161 Gb); ~2.5 Kb short jump (142 Gb); and Hi-C (59 Gb)). Although its broad applicability beyond the genomes assembled here has still to be demonstrated, our approach may enable a new generation of *de novo* genome assemblies that do not sacrifice the high standards for contiguity set by the Human Genome Project.

3.5 Acknowledgments

We thank Ferhat Ay, Evan Eichler, Joe Felsenstein, Phil Green, LaDeana Hillier, Max van Min, William Noble, Robert Waterston, and members of the Shendure lab for helpful discussions. Some of the sequencing data used in this research were derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta

Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. Our work was supported by grant HG006283 from the National Human Genome Research Institute (NHGRI; to J.S.); a graduate research fellowship DGE-0718124 from the National Science Foundation (to A.A. and J.O.K.); and grant T32HG000035 from the NHGRI (to J.N.B.). J.S. is a member of the scientific advisory board or serves as a consultant for Adaptive Biotechnologies, Ariosa Diagnostics, Stratos Genomics, GenePeeks, Gen9, Good Start Genetics, Ingenuity Systems and Rubicon Genomics.

3.6 Concurrent publications

This manuscript was published simultaneously with another publication in *Nature Biotechnology*[125] that uses the same principle of applying Hi-C to scaffolding genome assemblies. The authors apply a different computational approach for contig ordering, using a multidimensional scaling technique instead of the graph-based technique we used. This publication is also more narrowly focused than ours, as it does not attempt to scaffold any genomes other than the human genome, and it divides the human reference genome into 100-Kb bins rather than creating a human *de novo* assembly. However, it delves deeper into the case of the human reference genome, demonstrating that Hi-C can actually be used to improve this reference by placing 65 previously unplaced contigs.

Similarly, another publication in the same issue of *Nature Biotechnology*[126], which was in fact coordinated for publication with our manuscript, concerns the use of Hi-C for haplotyping genome assemblies. The authors used the principle that a chromosome tends to interact with itself more than with its homologous chromosome, even over long distances. This publication, combined with ours, raises the hope that a single Hi-C library might be used to both scaffold a *de novo* assembly and resolve its haplotypes, creating a high-contiguity diploid genome assembly.

CHAPTER 4: SPECIES-LEVEL DECONVOLUTION OF METAGENOME ASSEMBLIES USING HI-C

This chapter is based on the following peer-reviewed publication[99]:

Joshua N. Burton, **Ivan Liachko**, Maitreya J. Dunham and Jay Shendure. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes* | *Genomes* | *Genetics* **4**(7), 1339-1346 (2014).

Boldface indicates authors who contributed equally to this work.

Ivan Liachko and I conceived the idea of using Hi-C for metagenomic deconvolution, Together we designed the MetaPhase method (**Figure 4.1**). Ivan Liachko and Maitreya Dunham selected and gathered the yeast and bacterial strains for the M-Y and M-3D samples (**Table 4.1**). Ivan Liachko mixed the species together and performed Hi-C and sequencing on the mixtures (**Table 4.2**). I designed and wrote the MetaPhase software and applied both it and Lachesis to the sequencing data (**Figures 4.2, 4.3**). I prepared the manuscript, with input from all authors. Maitreya Dunham and Jay Shendure supervised the study. The software portion of the MetaPhase method was written in C++ and Perl and is available online at https://github.com/shendurelab/MetaPhase.

4.1 Summary

Microbial communities consist of mixed populations of organisms, including unknown species in unknown abundances. These communities are often studied through metagenomic shotgun sequencing, but standard library construction methods remove long-range contiguity information; thus, shotgun sequencing and *de novo* assembly of a metagenome typically yield a collection of contigs that cannot readily be grouped by species. Methods for generating chromatin-level contact probability maps, *e.g.*, as generated by the Hi-C method, provide a signal of contiguity that is completely intracellular and contains both intrachromosomal and interchromosomal information. Here, we demonstrate how this signal can be exploited to reconstruct the individual genomes of microbial species present within a mixed sample. We apply this approach to two synthetic metagenome samples, successfully clustering the genome content of fungal, bacterial, and archaeal species with more than 99% agreement with published reference genomes. We also show that the Hi-C signal can secondarily be used to create scaffolded genome assemblies of individual eukaryotic species present within the microbial community, with higher levels of contiguity than some of the species' published reference genomes.

4.2 Introduction

All ecosystems on this planet include communities of microbial organisms[85,127-130], including our own bodies[131,132]. However, our understanding of microbial communities is limited by our ability to discern which microbial taxa they contain and how these taxa contribute to community-scale phenotypes. Most microbial taxa cannot be cultured independently of their native communities[83] and therefore are not readily isolated for individual analysis, *e.g.*, by genome sequencing. Such unculturable taxa may be difficult to study even if they are abundant[89]. Consequently, many analyses of microbial communities must treat them as a single sample, for example, by shotgun sequencing of a metagenome[85,89,132] or metatranscriptome[133,134].

A central challenge in analyzing a metagenome involves determining which sequence reads and/or sequence contigs originated from the same taxon[90]. Many computational methods have been developed to deconvolute metagenomic assemblies by mapping reads or contigs to assembled microbial genomes[92] or by analyzing base composition[135] or gene abundance[90,129]. However, these strategies are handicapped by the remarkable variety of unculturable species in virtually all microbial communities and the fact that most of these species have not yet been sequenced in isolation[83]. Individual microbial genomes have been deconvoluted from shotgun metagenome reads using methods such as mate-pair libraries[89,136], lineage-specific probes[137], single-cell sequencing[138], neural networks[129,139,140], and differential coverage binning[91,140]. Some *de novo* assembly software has also been adapted to anticipate metagenomic shotgun sequence data[86,87]. These methods have succeeded in isolating whole genomes from abundant organisms in some communities, but they are specific to the communities for which they have been devised and often require prior knowledge of the community's composition. Metagenomic analyses would benefit greatly from a more generalizable methodology that can identify the sequence content belonging to each taxon without any *a priori* knowledge of the genomes of these organisms, especially the genomes of low-abundance taxa. Related to the challenge of determining which contigs belong to the same species are the problems of how to further define and assemble the one or multiple chromosomes that comprise each species' genome, and how to define and assign plasmid content to one or multiple species.

To enable robust reconstruction of individual genomes from within a complex microbial community, additional information beyond standard shotgun sequencing libraries is required. We speculated that contact probability maps generated through chromosome conformation capture methods[141] might inform the species-level deconvolution of metagenome assemblies. One specific method for generating contact probability maps, Hi-C, uses proximity ligation and massively parallel sequencing to generate paired-end sequence reads that capture three-dimensional genomic interactions within a cell[98]. We and others recently exploited the distance dependence of intrachromosomal interactions in Hi-C datasets to facilitate chromosome-scale *de novo* assembly of complex genomes[97,125]. As an additional feature, because crosslinking occurs prior to cell lysis in the Hi-C protocol, each Hi-C interaction involves a pair of reads originating from within the same cell. We speculated that in the context of heterogeneous cell populations (*e.g.*, microbial communities), such pairings might inform the clustering of genome sequences originating

from the same species. Importantly, the efficacy of the Hi-C protocol has recently been demonstrated in bacteria[142,143], implying that this method could be applicable to metagenome samples containing both prokaryotic and eukaryotic cells.

Here, we provide experimental proof-of-concept for this strategy in several contexts while also describing an algorithm for this task, MetaPhase (**Figure 4.1**). We reconstruct the genomes of as many as 18 species from a single synthetic mixture of eukaryotes and/or prokaryotes, including some species with as much as 90% sequence identity to one another, and we generate high-contiguity *de novo* assemblies for individual eukaryotic species present within the synthetic microbial community. In the process, we also present the first demonstration of Hi-C in an archaeal species.

Acronym	Description	Number of Species	Species
M-Y	Mixture of yeasts	13	*S. cerevisiae*, other Saccharomyces; Lachancea, Kluyveromyces, etc.
M-3D	Mixture of 3 domains	18	8 yeasts (Dikarya); 9 bacteria; 1 archaeon

Table 4.1: Contents of the metagenome samples sequenced and analyzed with MetaPhase.

Sample	Library Type	Read Length, bp	Read Pairs, millions
M-Y	Shotgun	101	85.7
	Mate-pair	100	9.2
	Hi-C	100	81.0
M-3D	Hi-C	101	14.3

Table 4.2: Sequencing libraries used in MetaPhase analyses. Hi-C libraries were prepared with the *Hin*dIII restriction enzyme. For description of sample names, see **Table 4.1**.

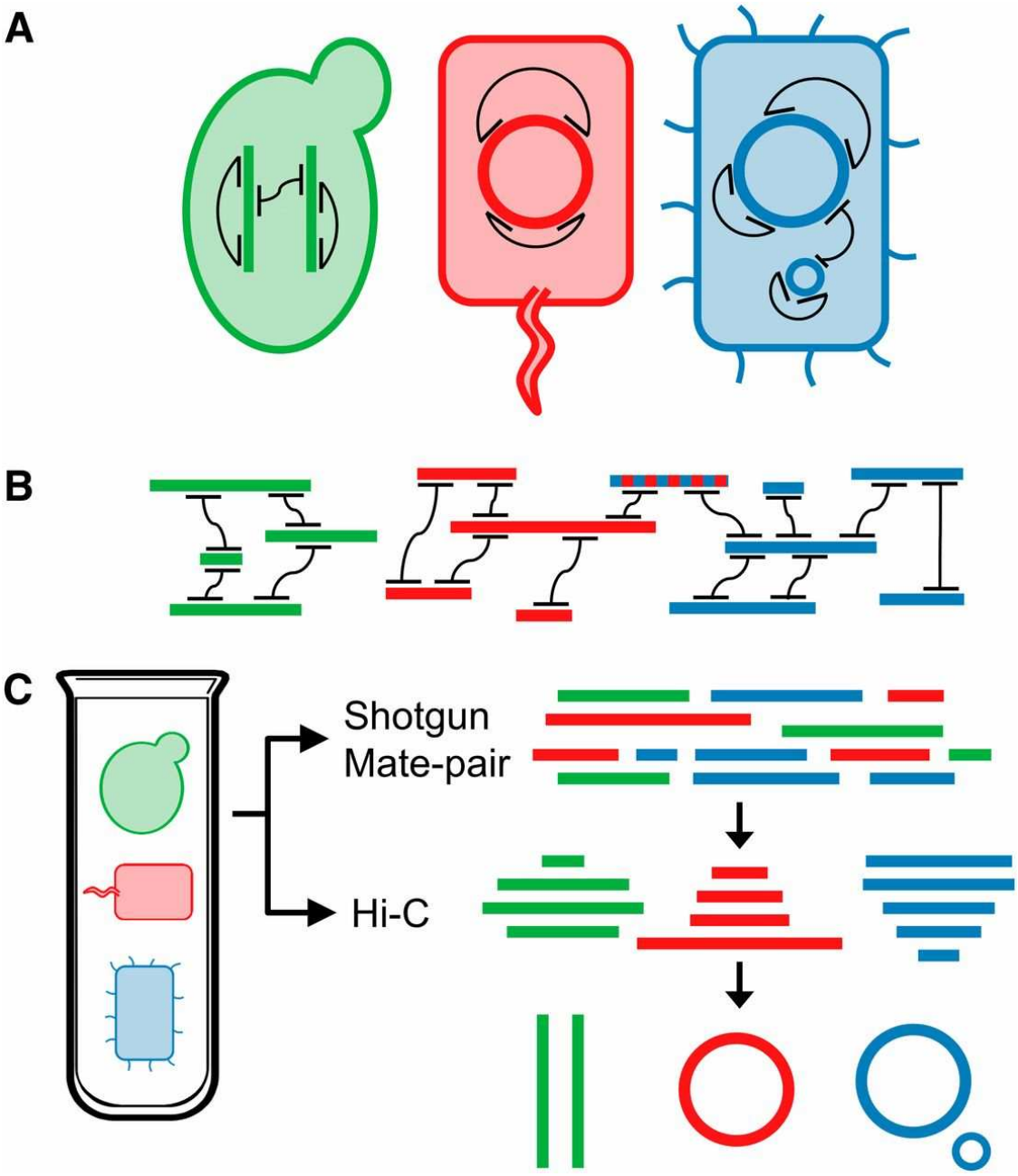


Figure 4.1: Overview of MetaPhase methodology. (**a**) Performing Hi-C on a mixed cell population. Shown are three microbial cells of different species (green, red, blue) with their genomes (thick colored lines or circles), which may or may not include multiple chromosomes or plasmids. A Hi-C library is prepared and sequenced from this sample. The Hi-C read pairs from this library (black lines) represent pairs of sequences that necessarily occur within the same cell. (**b**) Using Hi-C reads to deconvolute individual species' genomes. A shotgun sequencing library from the same sample is used to create a draft *de novo* metagenome assembly, which contains contigs from all species (thick lines). The Hi-C reads are then aligned to this assembly. Because sequences connected by Hi-C links must appear in the same species, the contigs form clusters representing each species. Note that some sequences (*e.g.*, blue/red dotted line) may appear in multiple species, confounding the clustering. (**c**) MetaPhase workflow. A single metagenome sample is used to create shotgun, Hi-C, and (optionally) mate-pair libraries, which are used together to create individual species assemblies.

4.3 Results

4.3.1 Deconvoluting yeast genomes from a synthetic mixture

To evaluate the effectiveness of the proposed strategy, we first applied it to a sample of defined, exclusively eukaryotic composition. Specifically, we created a synthetic metagenome sample consisting of 16 yeast strains ("M-Y") (**Figure 4.2** and **Table 4.1**). The strains include four strains of *Saccharomyces cerevisiae* as well as 12 other species of Ascomycetes at varying genetic distances from *S. cerevisiae*, all of which have publicly available reference genomes (Table B.2.1, Figure B.3.1, and Figure B.3.2). These strains were grown individually to saturation in YPD medium and mixed in approximately similar proportions (with the exceptions of *S. kudriavzevii* and *P. pastoris*, which were mixed in at a much lower proportion to test the sensitivity of this approach). The mixed cell culture was treated with the cross-linking agent, formaldehyde, immediately after mixing the individual strains. Total DNA was isolated from the mixed population culture and prepared for sequencing. This resulted in 92.1 M Illumina read pairs from one shotgun library, 9.2 M Illumina read pairs from one mate-pair library, and 81.0 M read pairs from one Hi-C library (**Table 4.2**).

We used the shotgun and mate-pair (~4 kb) read pairs to generate a draft *de novo* metagenome assembly using IDBA-UD[86] (*see* Supplementary Methods, Appendix B.1). This assembly had 48,511 contigs with a total length of 136 Mb and an N50 contig length of 17.3 kb. Contigs from this assembly covered most of the reference genomes of all 13 yeast species present (average = 96.0%), with the exception of *P. pastoris* (13.7%), which also had a very low fraction of shotgun reads aligning to it (1.2%), confirming its low abundance in the sample (Figure B.3.3).

We next aligned the Hi-C read pairs to the M-Y metagenome assembly, yielding a network of contigs joined by Hi-C links (**Figure 4.2a**). Then, exploiting the fact that sequences connected by Hi-C links are overwhelmingly expected to derive from the same cell, we used the links to cluster these contigs, applying a novel algorithm that combines the steps of Jarvis-Patrick clustering[144] and agglomerative hierarchical clustering[114] (*see* Supplementary Methods, Appendix B.1). Our algorithm suggested the

presence of 12 distinct clusters in the sample based on the metric of intracluster link enrichment (Figure B.3.4). It clustered the majority of the metagenome assembly (111 Mb or 82.2% of total sequence length) into these 12 clusters. Of the remaining 24.1 Mb of sequence not clustered, the majority (99.7%) belonged to contigs that contained no *Hin*dIII sites and thus are not expected to produce a Hi-C signal in this experiment. Bootstrapping tests confirmed the robustness of our clustering method (Table B.3). The 12 clusters match closely with the 12 distinct species present in the draft assembly (excluding *P. pastoris*), and 99.2% of sequence was placed into the cluster representing a species to which it truly belongs (**Figure 4.2b** and Figure B.3.5), allowing for the possibility of a given contig belonging to multiple species.

Further analysis of the clusters demonstrated several strengths and limitations of our method. Some species had greater Hi-C link densities than others after correcting for differences in species abundances (Figure B.3.6). This suggests that some species' cells are more susceptible to lysis during Hi-C than others, and MetaPhase is robust to these differences. However, distantly related species proved easier to separate than closely related species. For example, in the cluster representing *Scheffersomyces stipitis*, 99.88% of the contigs (by length) matched the *S. stipitis* reference genome; however, in the cluster representing *S. cerevisiae*, 3.3% of the contigs (by length) instead aligned uniquely to the genome of closely related *S. mikatae*. We also noted that the sequence content in the *S. cerevisiae* cluster included the contigs that aligned to any of the four *S. cerevisiae* strains' reference genomes. This indicates that although our method is generally successful in merging closely related strains of the same species into a single cluster, genetic variation between strains causes fragmentation of the species' sequence contigs in the metagenome assembly (Figure B.3.3), which in turn hampers our ability to delineate this cluster correctly because smaller contigs produce a weaker and noisier Hi-C signal. Separating this cluster into sub-clusters representing each *S. cerevisiae* strain represents an additional challenge that will require further algorithmic development.

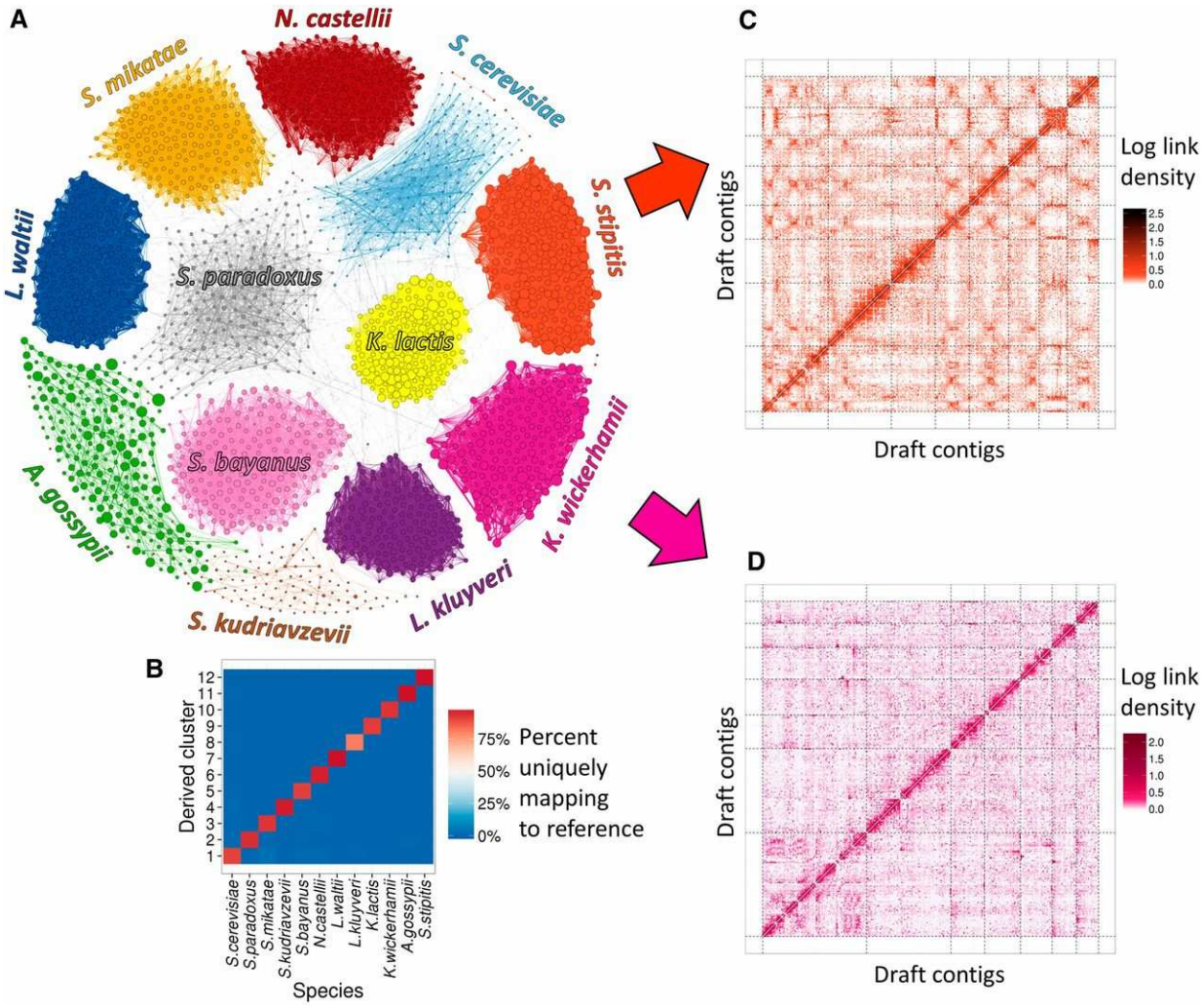


Figure 4.2: MetaPhase clustering results on the M-Y draft metagenome assembly. (**a**) Using Hi-C links to cluster contigs into 12 clusters, one for every species with a substantial presence in the draft assembly. Each contig is shown as a dot, with size indicating contig length, colored by species. Edge widths represent the densities of Hi-C links between the contigs shown. Only 2400 contigs are shown: the 200 largest contigs that map uniquely to each species. (**b**) Validation. This heatmap indicates what fraction of the sequence in each MetaPhase cluster maps uniquely to each of the reference genomes of the 12 present yeast species. Note that not all sequence is expected to map uniquely to one species. *x*-axis: the 12 yeast species. *y*-axis: the MetaPhase clusters. (**c-d**) Lachesis[97] reconstruction of individual species' genomes within the M-Y metagenome assembly. These heatmaps show the Hi-C link density among the contigs in the MetaPhase clusters corresponding to *S. stipitis* (**c**) and *K. wickerhamii* (**d**). The *x*-axis and *y*-axis show the clustering and ordering of contigs by Lachesis. Dotted black lines demarcate chromosomal clusters. Note the expected signals of enrichment within each chromosome and on the main diagonal. The assembly in **c** is similar to the *S. stipitis* reference genome (Figure B.3.7), whereas the assembly in **d** has far higher chromosome-scale contiguity than the best available *K. wickerhamii* reference[145].

4.3.2 Scaffolding individual yeast genomes with metagenomic libraries

We next sought to scaffold the genomic content of individual yeast species from the clusters of contigs representing each species. We ran the contigs in each cluster through our Lachesis software[97] to create chromosome-scale scaffolds. With the *S. stipitis* contig cluster, this approach yielded a scaffold for each of the eight *S. stipitis* chromosomes, with a total scaffolded sequence length of 14.2 Mb (91.7% of the *S. stipitis* reference genome and 95.1% of the portion of the *S. stipitis* genome that appeared in the draft metagenome assembly) (**Figure 4.2c**). These scaffolds matched the reference *S. stipitis* genome assembly fairly well (Figure B.3.7). There were a number of clustering errors, including one chromosomal cluster containing telomeric sequence from four other chromosomes, but the local misassembly rates were quite low: 0.9% and 1.1% for ordering and orientation errors, respectively. We applied this same method to the contig cluster representing *K. wickerhamii*, producing chromosome-scale scaffolds for each of the seven *K. wickerhamii* chromosomes, with a total length of 9.4 Mb (**Figure 4.2d**). These scaffolds, although we emphasize they have not been thoroughly validated, may represent a draft assembly with far higher contiguity than the existing *K. wickerhamii* reference genome[145], which has an N50 contig size of only 36.7 kb. Thus, the MetaPhase approach can be combined with Lachesis to create high-contiguity *de novo* genome assemblies of individual eukaryotic species within metagenome samples.

4.3.3 Concurrently deconvoluting eukaryotic, bacterial, and archaeal genomes

We next asked whether MetaPhase could be applied to deconvolute a metagenome consisting of both eukaryotic and prokaryotic species. Toward a proof of concept, we gathered samples of 18 species including eight yeasts, nine bacteria, and one archaeon, thus representing all three domains of life ("M-3D") (**Table 4.1** and Figure B.3.8). The species were grown in appropriate rich media and mixed together in similar proportions. The proportions were estimated by a combination of spectrophotometric, flow sorting, and counting approaches and were later confirmed by sequence coverage (Table B.2.2).

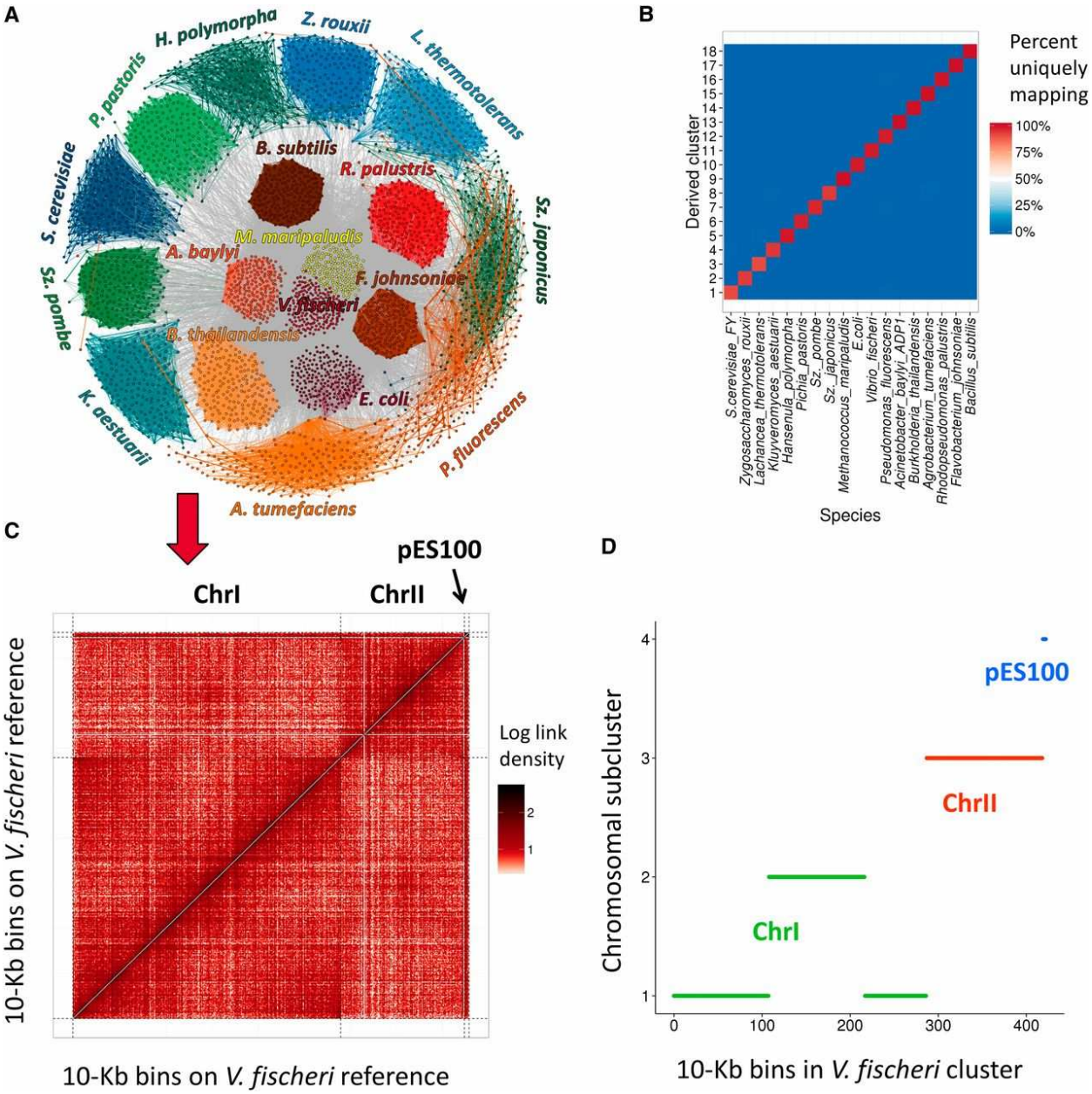


Figure 4.3: MetaPhase clustering results on the M-3D simulated contig assembly. (**a**) The reference genomes of the 18 species from the M-3D sample were split into 10-kb bins. Hi-C links from the metagenome sample were then used to divide the bins into 18 clusters, one for every species. The contigs are illustrated as in **Figure 4.2a**. Blue and green colors are yeast species; yellow is archaea; and red and orange are bacteria. (**b**) Validation. This heatmap has the same key as **Figure 4.2b**. (**c**) Heatmap of the M-3D Hi-C links aligned to the reference genome of *Vibrio fischeri*, one of the bacteria in the sample. The *V. fischeri* genome contains two chromosomes and a 46-kb plasmid, pES100 (demarcated by dotted black lines.) This heatmap has a resolution of 10 kb. (**d**) Applying Lachesis' clustering algorithm to the *V. fischeri* clustered genome to deconvolute the pES100 plasmid from the *V. fischeri*chromosomes. The *x*-axis shows the 424 simulated contigs in the*V. fischeri* cluster derived in (**a**) and (**b**). The *y*-axis shows the four clusters derived by Lachesis. Due to the presence of strong chromatin domains on chromosome I, Lachesis was unable to merge this chromosome into a single cluster and required an input of $N = 4$.

We created a simulated draft *de novo* metagenome assembly for M-3D by splitting the reference genomes of the 18 species into 10-kb contigs. We also experimentally generated a Hi-C sequencing library for the M-3D sample (**Table 4.2**), aligned these reads to the simulated contigs of the draft assembly, and clustered the contigs using Hi-C link frequencies (**Figure 4.3a**). Our algorithm predicted the presence of 18 distinct clusters, consistent with the actual content of the simulated draft assembly and experimental Hi-C data (Figure B.3.4). It clustered 89.1% of the simulated contigs into these 18 clusters; of the unclustered contigs, 85.8% contained no *Hin*dIII restriction sites and thus are not expected to produce a Hi-C signal in this experiment. The 18 clusters clearly matched the 18 species in the sample, with 99.6% of contigs clustered correctly (**Figure 4.3b** and Figure B.3.9). The clusters corresponding to archaeal and bacterial species had a particularly high accuracy rate of 99.87%. Bootstrapping tests confirmed the robustness of our method (Table B.2.3). Thus, our approach can simultaneously deconvolute the genomes of microbes belonging to all three domains of life, making it applicable to real and complex microbial communities.

Finally, we sought to use Hi-C to scaffold the genomic content of prokaryotic species from clustered contigs. Consistent with previous findings[142], we observed in the M-3D sample that both bacterial and archaeal genomes contain a substantially weaker signal of genomic proximity in Hi-C data than do eukaryotic genomes (Figure B.3.10). This suggests that in prokaryotic species, in sharp contrast with eukaryotic species, Hi-C is not very useful for ordering or orienting genomic content within chromosomes. However, Lachesis' clustering algorithm can still be used to deconvolute chromosomes, including plasmids, inside prokaryotic cells. We applied this algorithm to the genome of *Vibrio fischeri* ES114, a bacterial strain present in M-3D that contains two chromosomes and one plasmid, pES100 (**Figure 4.3c**). The chromosomal architecture of *V. fischeri* prevented a complete merging of its chromosome I, but chromosome II and pES100 both formed distinct clusters (**Figure 4.3d**). Thus, MetaPhase and Lachesis are capable of using Hi-C signal not only to deconvolute prokaryotic genomes but also to separate plasmid-derived sequence from chromosomal sequence within clusters corresponding to individual species.

4.4 Conclusions

Here, we demonstrate that contact probability maps such as those generated by Hi-C enable the deconvolution of shotgun metagenome assemblies and the reconstruction of individual genomes from mixed cell populations. Using only a single Hi-C library taken from a metagenome sample, we exploit two different signals inherent to Hi-C read pairing: the intracellularity of each pair, which enables species-level deconvolution, and the correlation of Hi-C linkage with chromosomal distance, which enables scaffolding of the *de novo* assemblies of at least eukaryotic species, as we have previously[97]. All of the sequencing libraries used here were generated by *in vitro* methods and were sequenced on a single cost-effective sequencing platform.

The MetaPhase method is straightforward enough to be applicable to any metagenome sample from which a sufficient number of intact microbial cells can be isolated (10^5–10^8). Furthermore, this approach can be applied to microbial communities containing both prokaryotes and eukaryotes. The application of MetaPhase to diverse microbial communities may permit the discovery and genome assembly of many unculturable and currently unknown microbial species. Additionally, the use of the intracluster enrichment metric (Figure B.3.4) permits a rough estimate of the species diversity within a draft metagenome assembly, a useful piece of information that is not easily measured. However, as with all shotgun metagenomic sequencing, low-abundance species—such as *P. pastoris* in our M-Y sample—will remain challenging to assemble into contigs without very deep sequencing. Additionally, in samples containing species such as dinoflagellates with unusually large genomes[146], even deeper sequencing of both shotgun and Hi-C libraries may be necessary.

We note that as MetaPhase delineates genomic content corresponding to individual microbial species, it also informs the chromosome and plasmid structure of these genomes and, in the case of eukaryotic species, it is capable of facilitating high-contiguity draft genome assemblies. Thus, it makes new species immediately amenable to phylogenetic and functional analysis while concomitantly increasing the power of existing genome databases to classify metagenomic reads via non-*de novo* methods. This method need not be limited to metagenome samples, because any complex cell mixture may be deconvoluted

into individual genomes assuming enough genomic diversity is present that reads can be accurately mapped.

4.5 Acknowledgments

We thank Giang Ong and Charlie Lee for help with sequencing; Riza Daza and Ruolan Qiu for technical assistance; Elhanan Borenstein, Rogan Carr, Roie Levy, and members of the Dunham and Shendure labs for helpful discussions; Adam Gordon for suggesting the name “MetaPhase”; and the following individuals for contributing microbial strains: Houra Merrikh, Carrie Harwood, Peter Greenberg, Amy Schaefer, Tom Lie, John Leigh, Grace Alex Mason, Christine Queitsch, Nick Rhind, Susan Jaspersen, James Cregg, Duncan Greig, Mark Johnston, David Bartel, Bonita Brewer, and the USDA Agricultural Research Service. We thank C. Beitel and colleagues for helpful and ongoing discussions regarding these methods. This work was supported by National Institute of Health (NIH)/National Human Genome Research Institute (NHGRI) grant T32HG000035 (Joshua N. Burton), NIH/NHGRI grant HG006283 (Jay Shendure), NIH/National Institute of General Medical Sciences grant P41 GM103533 (Ivan Liachko and Maitreya J. Dunham), NSF grant 1243710 (I.L. and M.J.D.), and DOE-LBL-JGI grant 7074345/DE-AC02-05CH11231 (J.S.). M.J.D. is a Rita Allen Foundation Scholar and a Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research. While this manuscript was in preparation, a preprint describing a related method appeared in *PeerJ PrePrints*[147]. Note added in proof: this preprint was subsequently published[148].

4.6 Concurrent publications

While preparing this manuscript, we became aware of a competing manuscript that was published as a pre-print in *PeerJ PrePrints*[147]. This manuscript later appeared as a peer-reviewed publication in *PeerJ* [148] at nearly the same time as our publication. The authors apply and demonstrate the same basic idea as

ours: using Hi-C for metagenomic deconvolution. They use simulated rather than real reads for their shotgun library. They apply their method to only one mixture of species, and it is simpler than either of our mixtures, containing only five strains of four bacterial species; furthermore, they do not do any further assembly post-deconvolution, as we do with LACHESIS. However, they do demonstrate chromosome- and plasmid-level deconvolution within their bacterial genomes. Also, unlike us, they demonstrate that Hi-C should be able in theory to enable strain-level deconvolution.

CHAPTER 5: HAPLOTYPE RESOLUTION IN THE ANEUPLOID HELA CANCER CELL LINE

This chapter is based on the following peer-reviewed publication[100]:

Andrew Adey, **Joshua N. Burton**, **Jacob O. Kitzman**, Joseph B. Hiatt, Alexandra P. Lewis, Beth K. Martin, Ruolan Qiu, Choli Lee and Jay Shendure. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).

Boldface indicates authors who contributed equally to this work.

Andrew Adey, Jacob Kitzman, and Jay Shendure conceived the idea of sequencing and haplotype-resolving the HeLa genome. Joe Hiatt, Alexandra Lewis, Beth Martin, Ruolan Qiu, Charlie Lee, and Andrew Adey maintained cell cultures, constructed libraries, and performed DNA sequencing. Andrew Adey called SNVs, indels, and structural variants in HeLa CCL-2 and compared them with the panel of control germline genomes. Andrew Adey constructed the HeLa CCL-2 copy number profile, which I refined to a haplotype-resolved copy number profile (**Figure 5.1**). I called loss-of-heterozygosity regions and regions of genetic ancestry. Jacob Kitzman applied the fosmid clone pool sequencing method and called single-nucleotide variants in the haplotype clone pools. I used these calls to perform haplotype phasing and scaffolding on the variants. Jacob Kitzman and I analyzed the mutation rate in the HeLa genome. Andrew Adey analyzed the HPV integration locus (**Figure 5.2**), mapped the haplotype-resolved variants onto ENCODE data to haplotype-resolve the HeLa epigenome (**Figures 5.3, 5.4**), and performed the HeLa strain comparisons. Andrew Adey, Jacob Kitzman, Jay Shendure, and I wrote the manuscript. Jay Shendure supervised the study.

5.1 Summary

In this chapter we demonstrate new algorithms for haplotype resolution and apply them to the genome of HeLa, a famous cancer cell line derived from the cells of Henrietta Lacks. Jacob Kitzman had previously developed a method[82] to create haplotype-resolved fosmid clone pools, enabling molecular phasing. Here we exploit this method again to enable medium-range haplotype resolution across the HeLa genome. Then we exploit copy number imbalances to extend the haplotype resolution across entire chromosome blocks, and we use this resolution to phase the epigenome of HeLa and explore the etiology of the viral insertion that originally caused the cancer. This work made HeLa the first cancer genome to be haplotype-resolved. Also, by making the HeLa genome publicly available, it led directly to a historic data-sharing agreement between the NIH and the surviving members of Henrietta Lacks' family[102].

5.2 Introduction

The HeLa cell line was established in 1951 from cervical cancer cells taken from a patient, Henrietta Lacks. This was the first successful attempt to immortalize human-derived cells *in vitro*[149]. The robust growth and unrestricted distribution of HeLa cells resulted in its broad adoption—both intentionally and through widespread cross-contamination[150]—and for the past 60 years it has served a role analogous to that of a model organism[101]. The cumulative impact of the HeLa cell line on research is demonstrated by its occurrence in more than 74,000 PubMed abstracts (approximately 0.3%). The genomic architecture of HeLa remains largely unexplored beyond its karyotype[121], partly because like many cancers, its extensive aneuploidy renders such analyses challenging. We carried out haplotype-resolved whole-genome sequencing[82] of the HeLa CCL-2 strain, examined point- and indel-mutation variations, mapped copy-number variations and loss of heterozygosity regions, and phased variants across full chromosome arms. We also investigated variation and copy-number profiles for HeLa S3 and eight additional strains. We find that HeLa is relatively stable in terms of point variation, with few new mutations accumulating after early passaging. Haplotype resolution facilitated reconstruction of an amplified, highly rearranged region of

chromosome 8q24.21 at which integration of the human papilloma virus type 18 (HPV-18) genome occurred and that is likely to be the event that initiated tumorigenesis. We combined these maps with RNA-seq[151] and ENCODE Project[152] data sets to phase the HeLa epigenome. This revealed strong, haplotype-specific activation of the proto-oncogene *MYC* by the integrated HPV-18 genome approximately 500 kilobases upstream, and enabled global analyses of the relationship between gene dosage and expression. These data provide an extensively phased, high-quality reference genome for past and future experiments relying on HeLa, and demonstrate the value of haplotype resolution for characterizing cancer genomes and epigenomes.

5.3 Results

5.3.1 Point variation in HeLa

We generated a haplotype-resolved genome sequence of HeLa CCL-2 using a multifaceted approach that included shotgun, mate-pair and long-read sequencing, as well as sequencing of pools of fosmid clones[82] (Table C.2.1). To catalogue variants, we carried out conventional shotgun sequencing to 88× non-duplicate coverage and reanalyzed 11 control germline genomes in parallel[153] (Tables C.2.2 and C.2.3). Although normal tissue corresponding to HeLa is unavailable, the total number of single-nucleotide variants (SNVs) identified in HeLa CCL-2 ($n = 4.1 \times 10^6$) and the proportion overlapping with the 1000 Genomes Project[154] (90.2%) were similar to controls (mean $n = 4.2 \times 10^6$ and 87.7%, respectively), suggesting that HeLa has not accumulated appreciably large numbers of somatic SNVs relative to inherited variants. Indel variation was unremarkable after accounting for differences in coverage (Figure C.3.1). Short tandem repeat profiles of HeLa also resembled controls, consistent with mismatch repair proficiency (Figure C.3.2).

After removing protein-altering variants that overlapped with the 1000 Genomes Project or the Exome Sequencing Project[155], similar numbers of private protein-altering (PPA) SNVs were found in HeLa (*n* =

269) and controls (mean n = 391). Gene ontology analysis found that all terms enriched for PPA variants in HeLa ($P \leq 0.01$) were also enriched in at least one control (except for 'startle response' in HeLa), suggesting that known cancer-related pathways are not perturbed extensively by point or indel mutations (Figure C.3.3). Although a previous study of the HeLa transcriptome[156] reported an enrichment of putative mutations in cell-cycle- and E2F-related genes, subsequently generated population-scale data sets contain all variants that we observed in these genes, suggesting that they are inherited and benign rather than somatic and pathogenic.

The overlap between PPA variants and the Catalogue of Somatic Mutations in Cancer (COSMIC)[157] was similar for HeLa (n = 1) and control genomes (mean n = 2.6). The gene-level overlap with the Sanger Cancer Gene Census (SCGC)[157] was also similar for HeLa (n = 4) and control genomes (mean n = 8.7). Canonical tumor suppressors and oncogenes were notably absent among the five SCGC genes with PPA variants in HeLa (*BCL11B, EP300, FGFR3, NOTCH1* and *PRDM16*, Tables C.2.3–C.2.5). However, three are associated with HPV-mediated oncogenesis (*FGFR3, EP300, NOTCH1*) and may be ancillary to the dominant role of HPV oncoproteins in HeLa and other HPV^+ cervical carcinomas[158]. Mutations in *FGFR3* have been noted previously in cervical carcinomas, although infrequently and at different residues than observed here[159]. Both *EP300* and *NOTCH1* are recurrently mutated in diverse cancers and are involved in Notch signaling, a pathway that is dysregulated in HeLa[160]. *EP300*, which encodes the transcriptional co-activator p300, interacts directly with viral oncoproteins such as HPV-16 E6 and HPV-16 E7[161]. Although the in-frame deletion of a highly conserved amino acid in *EP300* seems to be somatic (heterozygous within a loss-of-heterozygosity (LOH) region), it is still possible that the others are rare, inherited variants or passenger mutations. Further studies are required to resolve their functional relevance and to assess whether these genes are recurrently altered in HPV^+ cervical carcinomas.

5.3.2 Structural variation in HeLa

Aneuploidy and LOH, which are hallmarks of cancer genomes, were mapped in HeLa by constructing a digital copy-number profile at kilobase resolution (**Figure 5.1**, Figure C.3.4). Read coverage profiles were

segmented by a Hidden Markov Model (HMM) and recalibrated to account for widespread aneuploidy (Figure C.3.5 and C.3.6). 61% of the genome has a baseline copy number of 3, and only a small minority (3%) has a copy number of greater than 4 or less than 2. LOH encompassed 15.7% of the genome, including several entire chromosome arms (5p, 6q, Xp, Xq) or large distal portions (2q, 3q, 6p, 11q, 13q, 19p, 22q) (Figure C.3.7 and Table C.2.6), consistent with previous descriptions of LOH in cervical carcinomas[162]. The overall profile is consistent with published karyotypes of various HeLa strains[121], suggesting that the hypertriploid state arose either during tumorigenesis or early in the establishment of the HeLa cell line.

Structural variants were identified by clustering discordantly mapped reads from 40-kb and 3-kb mate-pair libraries (Figure C.3.8). Twenty interchromosomal links were identified, including links for marker chromosomes M11 (9q33–11p14) and M14 (13q21–19p13). In addition, 209 HeLa-specific deletions and 8 inversions were found (Figure C.3.9 and C.3.11). Only two genes that are impacted by HeLa-specific structural rearrangements intersected with SCGC (*STK11*, *FHIT*), both of which are recurrently deleted in cervical carcinomas[163,164].

5.3.3 Haplotype resolution of the HeLa genome

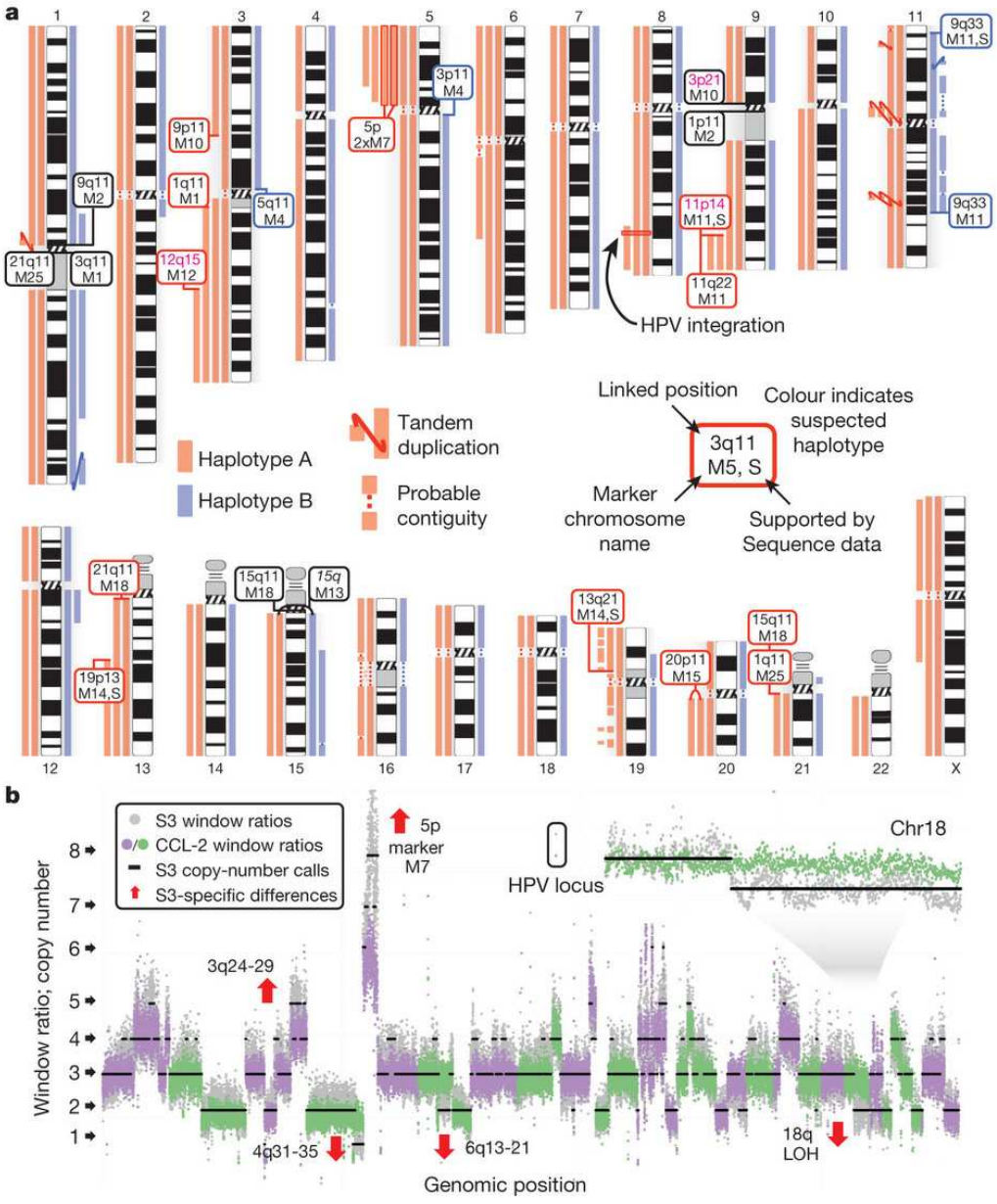
Conventional whole-genome sequencing fails to resolve haplotype phase, an essential aspect of the description and interpretation of non-haploid genomes, including cancer genomes[80]. Recently, several groups have demonstrated genome-wide measurement of local[82] or sparse[165] haplotypes, but these approaches have yet to be applied to aneuploid cancer genomes. To resolve haplotype phase across the HeLa genome, we sequenced pools of fosmid clones[82]. Specifically, we constructed three complex fosmid-clone libraries, and then carried out limiting dilution and shotgun sequencing of 288 fosmid clone pools. In summary, these were estimated to include 518,293 individual non-overlapping clones with a median insert size of 33 kb, for a total physical coverage of 6.3× of the haploid reference genome (Figure C.3.12). The complement of likely inherited heterozygous variants (SNP and indel, $n = 1.97 \times 10^6$) was ascertained by shotgun sequencing and by cross-referencing with calls made by the 1000 Genomes

Project, and then re-genotyped using reads from each clone pool. Alleles that were present at distinct heterozygous sites within a given clone were assigned, or 'phased', to the same inherited haplotype, and the unobserved alleles were implicitly phased to the opposite haplotype. When overlapping clones from distinct pools were merged, this resulted in haplotype blocks with an N50 (the contig size above which 50% of the total length of the haplotype assembly is included) of 550 kb containing 90.6% of heterozygous variants that were probably inherited.

Most of the HeLa genome is present at an uneven haplotype ratio (for example, 2:1 in regions in which copy number = 3). We sought to exploit the resulting allelic imbalance to phase consecutive haplotype blocks (Figure C.3.13). We first calculated the cumulative allelic ratio among shotgun reads for the SNVs residing in each haplotype block, which clustered closely with the underlying haplotype ratio. For example, in non-LOH regions with a copy number of 3 that have ratios of 2:1 or 1:2, allelic ratios calculated for each block had distributions centered on 0.32 or 0.65, close to the expected fractions of 1/3 and 2/3 (Figure C.3.14). Using these ratios, we merged haplotype blocks into scaffolds covering 1.96 Gb or 90.3% of the non-LOH HeLa genome (scaffold N50 of 44.8 Mb). The haplotype-resolved scaffolds were then merged with the copy-number map to produce a global, haplotype-resolved copy-number profile of the aneuploid HeLa genome (**Figure 5.1a**, Figure C.3.15).

Phasing accuracy was independently confirmed by several methods. First, 99.7% of informative read pairs from 3-kb mate-pair sequencing (each read overlapping a phased site) were concordant with the predicted phase. Second, long-insert single-molecule sequencing (Pacific Biosciences RS; mean, 2.97 kb; 90th percentile, 5.1 kb among informative reads) showed that 97.2% of reads were in perfect agreement with the predicted phase, despite the high per-base sequencing error rate of approximately 15% (Figure C.3.16). Third, examination of allelic state across 47.3 Mb of chromosome 18q, which underwent LOH in HeLa S3 but not in CCL-2, showed that out of the 17,761 affected alleles (heterozygous in CCL-2 but at an allele balance of greater than 0.9 among S3 reads), 99.7% corresponded to those phased together on haplotype A in CCL-2 (Figure C.3.17). Finally, windowed analysis of population allele frequencies revealed probable African or European genetic ancestry across

long stretches of the haplotype-resolved genome, consistent with recent admixture and a low switch error rate (Figures C.3.18 and C.3.19).



.

Figure 5.1: Haplotype-resolved copy number of the HeLa genome. (**a**) Copy-number profile of HeLa split by haplotypes. Links denote likely contiguity and tandem duplications. Boxes indicate marker chromosomes identified by copy-number breakpoints (boxes are colored by haplotype; black, unknown; pink text, uncertain locations; S, links confirmed by mate-pair sequencing). (**b**) Windowed copy-number ratios for HeLa CCL-2 (green and purple, alternating chromosomes) and HeLa S3 (grey), with predicted integer copy number for S3 (black). Notable strain differences are indicated by red arrows (for example, reduced copy over chromosome 18q). The window containing the HPV insertion and rearrangement is at elevated copy in both strains.

To measure the frequency of new mutations in the HeLa genome, we examined amplified haplotypes for *de facto* somatic mutations occurring during tumorigenesis or early in the cell line's subsequent passaging. Within LOH regions, these appear as polymorphisms; 2,883 such sites (mean, 1.31 per haploid Mb; Table C.2.7) were confirmed by clone-pool sequencing and allele frequency in shotgun sequencing (Figures C.3.20 and C.3.21). In non-LOH regions, in which one haplotype is amplified but both remain present, the majority of observed heterozygous sites are inherited, as reflected by their substantial overlap with variants from the 1000 Genomes Project (86.7%, n = 2,339,608). Excluding these and sites found in the 11 control genomes, 5,282 sites (mean, 1.32 per haploid Mb) remained at which clones differed in genotype between the two or more amplified copies of the same germline haplotype, with little regional variation in the abundance (Figure C.3.22). In summary, 8,165 somatic mutations were validated with an estimated sensitivity of 61.1%, placing an upper bound on the point-mutational burden sustained by HeLa CCL-2 after aneuploidy. Despite many additional doublings in culture, this point-mutation frequency (2.16 per Mb) is on the lower end of frequencies observed across different cancer genomes[166]. However, without estimates for parameters such as the number of doublings during tumorigenesis, the count of cells explanted, and the number of passages in culture, this estimate of post-aneuploidy mutational burden cannot be rescaled to a rate per base per division.

5.3.4 Comparing HeLa S2 with other strains

Four years after the initial establishment of the HeLa cell line, several additional strains were cloned[167]. One of these, HeLa S3, remains in widespread use today and has been profiled extensively as part of the ENCODE Project. To investigate the divergence between CCL-2 and S3, we carried out shotgun sequencing of S3 to 26× coverage. Outside of S3-specific regions of LOH, 94.5% of rare variants in CCL-2 were shared with S3 (n = 204,841 sites excluding 1000 Genomes Project and segmental duplications, and requiring ≥8× coverage in each genome; Figure C.3.23 and Table C.2.8). Somatic mutations were also shared, though to a lesser degree: 72.4% of clone-confirmed somatic mutations from CCL-2 were

found in S3 (*n* = 8,054 sites with ≥8× coverage in S3), consistent with a low rate of somatic SNV accumulation since the strains diverged in 1955.

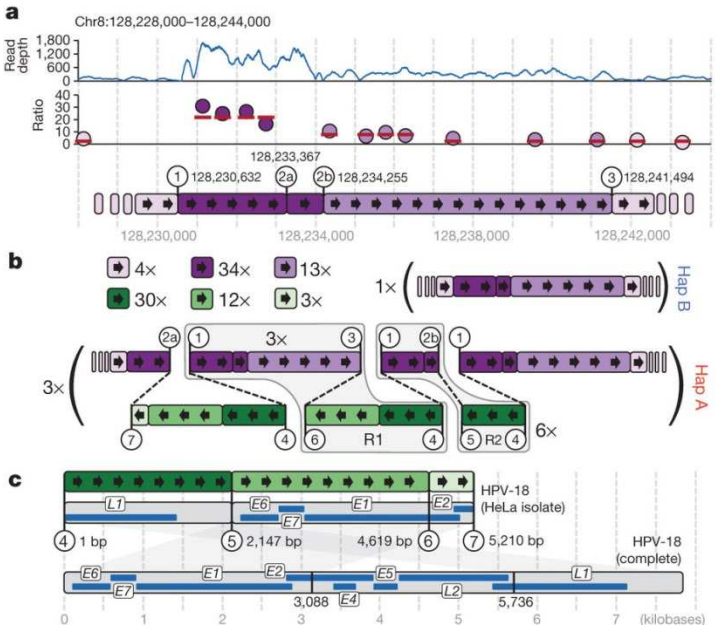


Figure 5.2: The HeLa HPV integration locus. (**a**) Chromosome 8 read depth flanking the HPV integration site (top, blue line), windowed copy-number ratios (purple points, shaded by segment) and integer copy states (red bars, middle), and corresponding segments and breakpoints (circled numbers with genomic coordinates, bottom). (**b**) Proposed HPV integration structure: per-segment copy number (top left), non-rearranged haplotype B (copy number = 1, top right), rearranged haplotype A with HPV insertion (copy number = 3, bottom) carrying approximately 3 and 6 tandem copies of repeats R1 and R2, respectively. Hap, haplotype. (**c**) The partial HPV-18 genome and corresponding genes (grey and blue, top) with breakpoints highlighted by numbered circles. For reference, the entire HPV-18 genome is shown (bottom).

The copy-number profile of HeLa S3 broadly mirrors that of CCL-2 (**Figure 5.1b** and Figure C.3.7 and C.3.24) as well as eight additional HeLa strains that we sequenced lightly (3.5 to 4.3×). We observed some strain-specific differences (Figure C.3.25–C.3.27), consistent with previous reports of karyotypic heterogeneity both among and within strains. Despite some variability, a copy number of three was the dominant state consistently, with a median of 52% of the genome across the eight strains (range 38–60%), similar to its prevalence in CCL-2 (61%). Gains or losses of entire chromosome arms were observed (for example, chr18q, HeLa S3 (**Figure 5.1b**), chr9p, CCL-13 (Figure C.3.28 and C.3.29)), but smaller amplifications and deletions were more common. These may correspond to variability in copy rather than in the content of marker chromosomes present, as suggested by high overall breakpoint concordance between strains (81% of copy-number breakpoints within ±1 Mb were present in ≥2 strains). The additional eight cell lines analyzed here were identified in the 1970s[168] as products of HeLa contamination into other tissue cultures in the preceding two decades. Their shared set of structural

abnormalities reflects their common origin from small founder populations of contaminating cells and reinforces the view that the structural rearrangements resulting in marker chromosomes arose early and are variable in copy number.

5.3.5 HPV integration into the HeLa genome

Nearly all cervical cancer is caused by human papillomavirus (HPV) infection. Within HeLa, a partial copy of the HPV-18 genome is integrated at a known fragile site on chromosome 8q24.21[169,170]. Haplotype and copy-number maps indicate that the flanking regions are present at copy number four, at a haplotype ratio of 3:1. To characterize the structure and copy number of the insertion, we included the HPV-18 genome alongside the human reference during alignment of clone-pool reads. By analyzing patterns of coverage from breakpoint-spanning fosmid clones, read-depth data and breakpoint sequencing, we generated a structural model for the viral integration (**Figure 5.2a,b** and Figure C.3.30 and C.3.31). Two repeat structures (which we designate R1 and R2) consisting of the partial viral genome are interspersed with regions of human chromosome 8q24.21 genomic DNA. The viral genome is present with identical breakpoints on each copy of the amplified haplotype, to the exclusion of the other haplotype, which remains at single copy and lacks integration-associated rearrangements, confirming that integration and rearrangement preceded aneuploidy. The integrated structure contains only two-thirds of the complete HPV-18 genome, including full-length copies of the *E6* and *E7* oncogenes necessary for telomerase activity (amplified to a copy number of approximately 12), but lacking a functional copy of *E2*, an inhibitor of E6 and E7[158] (**Figure 5.2c**). In addition, a distinct portion of the HPV-18 genome, amplified to a copy number of approximately 30 in HeLa, includes an epithelium-specific enhancer that controls *E6* and *E7* transcription[171], possibly contributing to their high expression (Figure C.3.32).

5.3.6 Haplotype and copy number resolution of the HeLa epigenome

Extensive sequencing-based functional genomic data have been generated on HeLa and other cancer cell lines by the ENCODE Project[152], but these have the potential to be misinterpreted if their analysis does not account for aneuploidy and phase. As HeLa CCL-2 and S3 are nearly identical in genotype, we used haplotype and copy-number maps of CCL-2 to assign phase to publicly available[152] functional data generated on S3, including transcription-factor binding, chromatin modification and chromatin-accessibility data sets. We also calculated haplotype-specific gene-expression scores using RNA sequencing (RNA-seq) data generated in this study and by others[151,152] (Figure C.3.33–C.3.35). For each data set, aligned reads were phased by comparison to HeLa CCL-2 haplotype blocks. Corresponding peak scores (chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) and DNase I sequencing (DNase-seq)) or gene-expression values (RNA-seq) called from the full set of reads were divided proportionally based on the abundance of phase-informative mapping to each haplotype, normalized to each haplotype's estimated copy number. Mapping to the human reference genome imposed a slight bias, favoring the reference allele by an average of 1.08-fold. We constructed two HeLa-specific reference sequences by introducing all SNVs from each haplotype onto one or the other; mapping to this reference mitigated most of the bias (to 1.02-fold, or a 75% reduction; Figure C.3.36–C.3.38).

Across the HeLa genome, gene expression is significantly correlated with copy number (P = 0.075; **Figure 5.3a,b**), suggesting a minimal role for gene-dosage buffering. Moreover, on average, each haplotype copy makes a comparable contribution to the transcriptome, despite uneven amplification and, in some cases, rearrangement (**Figure 5.3c,e**). This trend is also observed for histone modifications, DNase hypersensitivity and transcription factor binding (Figures C.3.39 and C.3.40). Transcript allele balances at sites heterozygous in CCL-2 on chromosome 18q closely followed the genomic balance (mean 66% representation of the A allele (two-thirds was expected)), but S3 nearly exclusively matched the A allele (94% of reads), reflecting the S3-specific LOH event (**Figure 5.3d**). However, a small number

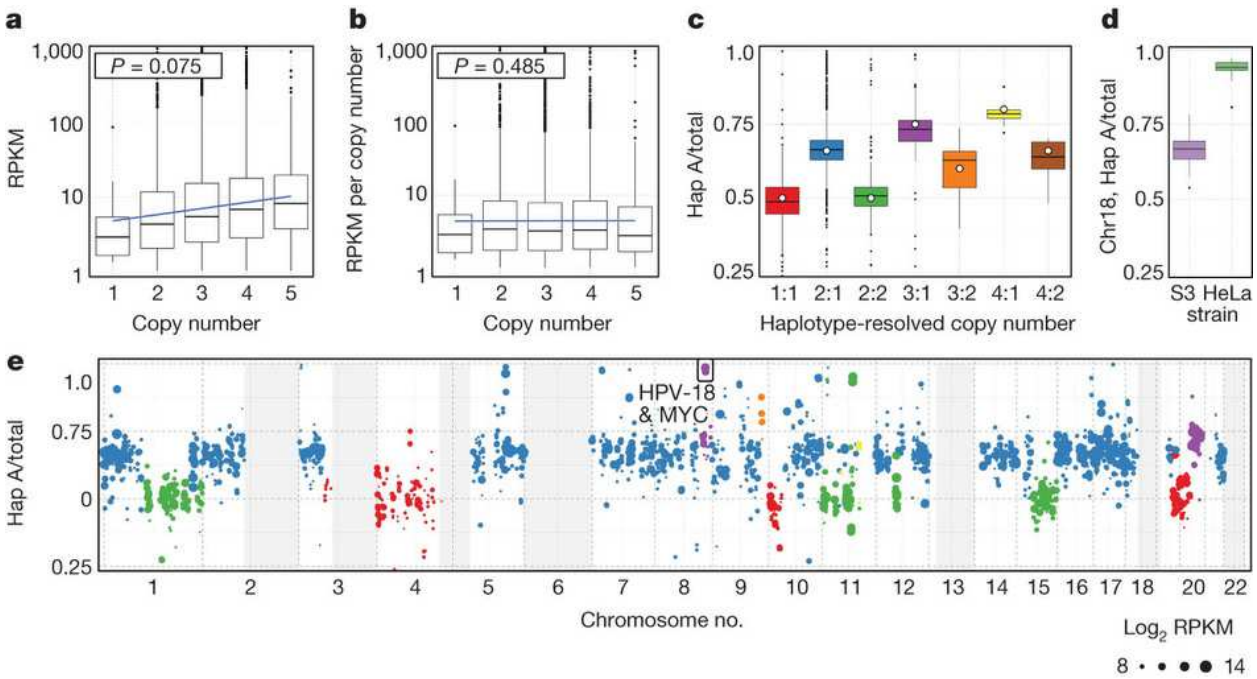


Figure 5.3: Gene expression by haplotype and copy number in HeLa S3. (**a**) Transcript abundance (reads per kb per million (RPKM), for genes with RPKM ≥1) is positively correlated with gene copy. (**b**) Expression per copy (RPKM per gene copy number) does not correlate with copy number. (**c**) Fractional contribution of haplotype A to overall expression (Hap A/total) (RPKM averaged across Mb windows at phased sites) split by haplotype-resolved copy number. Open circles indicate expected fractions. (**d**) Haplotype-A-specific expression in HeLa S3 but not CCL-2 across S3-specific LOH on chr18q. (**e**) Haplotype A fractional contribution to expression across the genome, color-coded by underlying haplotype-resolved copy number as in **c** (point size: $\log_2$ total RPKM, grey boxes: HeLa S3 LOH).

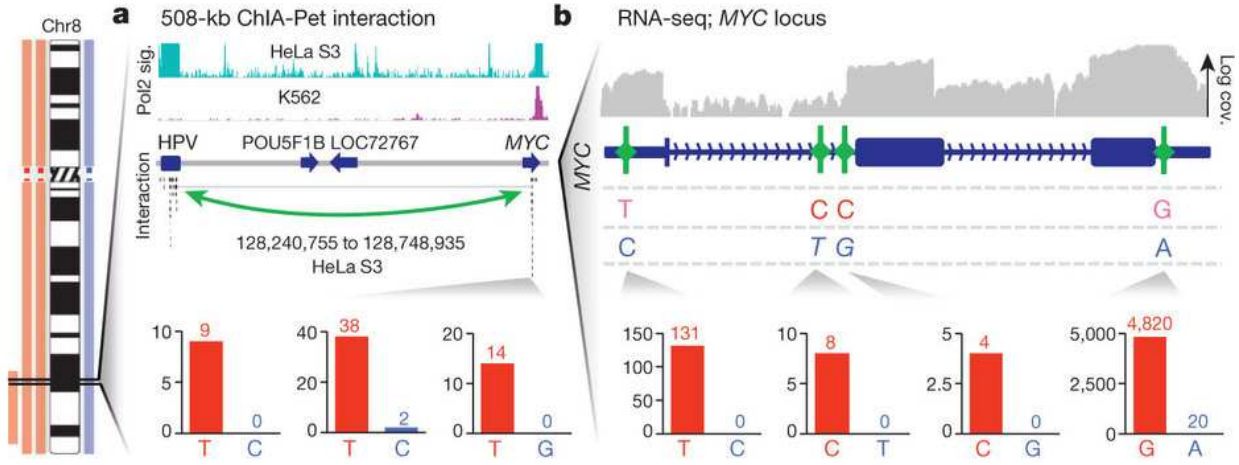


Figure 5.4: Haplotype-specific regulation near the HPV integration site. (**a**) Long-range chromatin interactions between the HPV and *MYC* loci demonstrated by ChIA-PET[171] with the RNA polymerase II signal (top) shown for HeLa S3 and an HPV$^-$ cell line (K562). Chromatin interactions (middle) are indicated by a green arrow. Bar graphs (bottom) show read counts at phased, informative sites in *MYC* (red, haplotype A, blue, haplotype B). (**b**) Transcript abundance in HeLa S3 across the *MYC* locus measured by RNA-seq. Overall coverage is shown in grey (top) with phased, informative sites highlighted by green ticks (pink text, non-reference alleles). Haplotype contributions at each variant are shown in bar graphs (bottom), as in **a**.

of regions showed strong imbalances between each haplotype's contribution to overall patterns of expression, chromatin modification and transcription-factor binding (2.4% of ENCODE peaks, excluding those in LOH regions; Figure C.3.41–C.3.44). Interestingly, the HPV-18 insertion locus and proto-oncogene *MYC* (separated by approximately 500 kb) were among the regions with the most highly haplotype-imbalanced regulation in the genome (Figure C.3.45). Phased RNA-seq data indicate that *MYC* is highly expressed, but almost exclusively from the HPV-18-integrated haplotype (mean ratio, 95:1; **Figure 5.4b** and Figure C.3.46). Phased ENCODE tracks and long-range chromatin interaction data (ChIA-PET (chromatin interaction analysis with paired-end tag sequencing)[172]; **Figure 5.4a** and Figure C.3.47) across the region indicate that transcription-factor occupancy, active chromatin marks and long-distance physical contacts are also nearly exclusive to the HPV-integrated, transcriptionally active haplotype. Taken together, these data implicate viral integration as a strong activator of *MYC* expression[173], acting in *cis* rather than in *trans* and possibly mediated by the epithelium-specific viral enhancer amplified to a copy number of approximately 30 within the R1 repeat structure (**Figure 5.2b**)[171]. This strong *cis* interaction—between the amplified, integrated genome of a DNA tumor virus and a canonical proto-oncogene—may underlie the robust growth characteristics of the HeLa cell line, and provides indirect support for the hypothesis that inherited risk loci for cancer at chromosome 8q24 operate through activation of *MYC*[174].

5.4 Conclusions

In summary, we present a haplotype-resolved genome and a haplotype-resolved epigenome of a human cancer. Our study not only provides an overdue genomic analysis of the human cell line that is possibly the most commonly used in biomedical research but also represents a unique view into a cancer genome and epigenome enabled by the acquisition of haplotype information.

5.5 Acknowledgments

The genome sequence described in this paper was derived from a HeLa cell line. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. We also thank Martin Kircher, Matthew Snyder, Akash Kumar and Rupali Patwardhan as well as other members of the Shendure laboratory for advice and suggestions. We thank the Stamatoyannopoulos and Malik laboratories for cell aliquots. Our work was supported by a gift from the Washington Research Foundation; grant HG006283 from the National Genome Research Institute (NHGRI, to J.S.); grant CA160080 from the National Cancer Institute (to J.S.); a graduate research fellowship DGE-0718124 from the National Science Foundation (to A.A. and J.K.); grant T32HG000035 from the NHGRI (to J.N.B.); and grant AG039173 from the National Institute of Aging (to J.B.H.). J.S. is the Lowell Milken Prostate Cancer Foundation Young Investigator. J.S. is a member of the scientific advisory board or serves as a consultant for Ariosa Diagnostics, Stratos Genomics, Good Start Genetics, and Adaptive Biotechnologies.

5.6 Concurrent publication

While this publication was in press, another publication describing the HeLa genome appeared in *G3*[119]. This publication described the genome and transcriptome of the HeLa cell line in detail, including a list of SNV and structural variant calls, and it speculated on the etiology of the structural variation. However, it did not involve any haplotype resolution of the HeLa genome or epigenome, nor did it analyze HPV integration locus to determine the possible etiologies of the cervical cancer. The *G3* publication incited criticism from people who claimed that its variant list constituted an invasion of the Lacks family's genetic privacy. Francis Collins of the NIH became involved, and our own *Nature* publication was held in press while Collins and the Lacks family worked together to hammer out a historic data-sharing deal. This deal was made public concurrently with our publication[102].

CHAPTER 6: CHALLENGES AND FUTURE DIRECTIONS

6.1 Where do we go from here?

In this chapter I will attempt to place the work I have described in this dissertation into the greater context of the rapidly evolving field of genomics. Acknowledging the very small but not quite infinitesimal role that my own effort has played in advancing the progress of genomics, I will discuss the implications of my projects, the difficulties which have hindered them and which still remain challenges in the field, and the tools that are currently being developed to overcome these difficulties. I will give my predictions for where the field is likely to be headed in the future.

6.2 More reads

The story of genomics over the past decade can be encapsulated by the story of improvements in DNA sequencing technologies. The cost per base of sequencing has trended relentlessly downward over the past 25 years, at a rate substantially faster than Moore's Law[175], and therefore the range of genomic and biological questions that can be feasibly asked on a limited budget has continued to grow. Thus we have the developers of sequencing technologies such as Illumina, 454, Pacific Biosciences, and Ion Torrent (**Figure 2.1**) to thank for the ongoing revolution in genomic science.

Although next-generation sequencing technologies were becoming commercially available as early as 2005[55], the period of swiftest decline in sequencing costs occurred between 2007 and 2011, when Illumina bought and developed Solexa's sequencing-by-synthesis technology and greatly increased its throughput[175,176]. By 2011, Illumina had established market dominance, it had left its competitors far behind technologically, and its rate of innovation began to slacken. The promise of continuing to outpace Moore's Law looked questionable for a few years. However, new competitors were waiting in the wings,

and the year 2014 has happily witnessed a return to the speedy improvements in sequencing technology. In this calendar year alone, we have witnessed the first high-coverage dataset of long reads sequenced from the human genome, from Pacific Biosciences[177]; the first public release of nanopore sequencing data, from Oxford Nanopore[178]; and the announcement of Illumina's HiSeq X, which promises to resequence human genomes for less than $1,000 each[62]. It does not require great courage to predict that the per-base cost of sequencing will continue to drop rapidly over the next several years and that this will enable yet more studies across the field of genomics.

6.3 Longer reads

Not all bases are created equal: the length of the reads on which the bases are located also matters greatly. A billion short reads may provide less long-range contiguity than a single 10-Kb read that spans a long, near-perfect repeat element. Fortunately, many of the sequencing technologies now challenging Illumina produce reads that are long as well as numerous.

In the long term, perhaps the greatest promise comes from nanopore sequencing technologies, in which a nucleic acid molecule is passed through a pore protein embedded in a membrane and the voltage across the membrane is used as a readout. After decades of hope[29,179], nanopore sequencing appeared to be on the verge of becoming a reality in 2012 with Oxford Nanopore's announcement of its minION sequencer, based on an α-hemolysin nanopore[180]. The sequencer's specifications seemed too good to be true, and they were, at least in 2012. Two years later, amid much hype, Oxford was pressured to release some data from its MinION instrument; the reads were low in number and high in systematic errors[178]. It is clear that Oxford's α-hemolysin nanopore sequencing chemistry is far from perfected, as is its informatic analysis of the electrical signal. But Oxford is expending great effort in improving their workflow, and their error rate will certainly decrease. Meanwhile, other groups are exploring the possibility of sequencing DNA using other protein nanopores such as MspA[181].

With the advent of high-throughput nanopore sequencing likely still years away, the strongest player in long-read technologies is currently Pacific Biosciences, or “PacBio”, which uses a unique single-molecule real-time (SMRT) technology to produce long reads with trackable signals for methylation and other chemical modifications. PacBio’s instruments were first made commercially available since 2009 but produced too few reads to be cost-effective for most applications at that time. However, PacBio increased the throughput of its sequencers fourfold in 2012 and 2013 and may do so again in 2014. They have also succeeded in producing ever-greater read lengths, with the longest verified report at over 54 kilobases[182], although the number of extremely long reads is quite low because PacBio reads have a geometrically decaying length distribution.

Read length is more criticial for some genomic applications than for others, and the market is already becoming segmented into users who truly need long reads and users who do not. For resequencing studies in which SNPs are called, no long-range contiguity is necessary, and Illumina short reads suffice. However, the additional medium-range contiguity provided by long reads is useful for calling structural variants[94]. More relevant to the subject at hand, long reads are extremely helpful for any kind of *de novo* genome or metagenome assembly.

6.4 More complete genome assemblies

In **Chapter 3** of this dissertation I demonstrated a novel method to create high-contiguity *de novo* genome assemblies using Hi-C. In this approach, the Hi-C library preparation is difficult, but the sequencing itself can be carried out inexpensively on Illumina instruments, so the approach is quite cheap. However, the wide adoption of this approach may be challenged by the advance of PacBio, which is well-positioned to capture a large share of the *de novo* single-genome assembly market. PacBio offers a number of improvements over Illumina for accurately determining primary sequence content in *de novo* genome assemblies. PacBio’s single-molecule technology is less susceptible to GC bias than Illumina’s PCR-based method, leading to smoother genome-wide coverage[183]. Furthermore, although PacBio reads have

notoriously high error rates, the errors are distributed nearly randomly, rather than systematically biased as in Illumina reads[184]. Thus, if PacBio reads are sequenced to sufficiently high coverage, they can be used to error correct one another, and the resulting error-free reads can be used to generate a high-quality and high-contiguity *de novo* assembly[93].

In practice, this approach remains prohibitively expensive for general applicability to *de novo* assemblies of large genomes; a few more iterations of technological improvement are necessary. But PacBio reads are already being used to assemble small bacterial genomes, for which the total quantity of sequencing required is low[185], and the resulting assemblies are of high quality[186]. PacBio sequencing can also be used to improve the contiguity and completeness of existing *de novo* assemblies[187]. The ability of long reads to reach into difficult-to-assemble genomic regions, such as telomeres, centromeres, and complex rearrangements, promises to produce more complete assemblies of large genomes than ever before over the next few years[94].

It is worth emphasizing that the use of longer reads in *de novo* genome assembly is by no means redundant with the method I have developed to apply Hi-C sequencing to genome scaffolding (see **Chapter 3**). These approaches work at different contiguity ranges: long reads provide high-resolution sequence contiguity in the range of 100-10,000 bp, including difficult-to-assemble repeat regions, while Hi-C provides low-resolution but genome-wide contiguity. Either of these approaches, long shotgun reads or Hi-C scaffolding, would improve the quality of any assembly. However, using both approaches – sequencing long shotgun reads to create a *de novo* assembly, then creating a Hi-C library and using it to scaffold the assembly – would produce a very high-quality assembly with both long contigs and chromosome-scale contiguity! Note, however, that only the shotgun library will benefit from long-read sequencing, not the Hi-C library. Hi-C reads only need to be long enough to align unambiguously to the *de novo* assembly, and Illumina reads are suitable for this purpose, except in the rare case of large and repetitive genomic regions that cannot be assembled into contigs containing unique sequence.

It is quite possible that shotgun sequencing with PacBio or another long-read sequencing technology, possibly abetted by a crosslinking-based approach such as Hi-C, will become the dominant approach to *de novo* genome assembly in a few years. I am hopeful that this will lead to a restoration of the high

quality standard, originally set by the Human Genome Project, for publication-quality genome assemblies. I also anticipate that it will enable yet more post-genomic biological analysis.

6.5 Longer haplotypes

Another straightforward application of longer reads is molecular phasing of haplotypes. There is great demand for a molecular phasing methodology that could be applied to human genomes cheaply and reliably. Illumina's short reads cannot phase haplotypes on their own, for the simple reason that heterozygous SNVs are rarely close enough on the genome that a single short read will cover two or more of them. However, a long read such as produced by PacBio can easily cover many SNVs, and a sufficiently high-depth sequencing dataset could be applied directly into a standard haplotype phasing algorithm, such as the one I used in **Chapter 5** of this dissertation[103]. Here as in the *de novo* assembly problem[93], the high error rates of PacBio reads must be overcome with deeper sequencing depth in order to produce reliable base calls, but it can be done. Work from Jason Chin at Pacific Biosciences (in press) demonstrates that a sufficient depth of PacBio shotgun reads can be used to assemble both haplotypes of an *Arabidopsis thaliana* genome into a single string-based assembly, effectively performing *de novo* assembly and haplotype phasing simultaneously. If the cost of PacBio sequencing continues to decrease over the next few years, this could easily become a standard approach for molecular haplotype phasing.

Long reads are not the only way to achieve haplotype phasing; short reads can do the trick as well, if accompanied by proper contiguity information. In **Chapter 5** I illustrated the power of combining fosmid clone pools with short-read sequencing for haplotype phasing in the HeLa genome. Fosmids are expensive and time-consuming to create, but other library preparation technologies are showing promise. An alternative method, developed in part by my colleague Andrew Adey, achieves haplotype resolution in human genomes via Tn5 transposition and combinatorial indexing[188]. Additionally, in a paper[126] published concurrently with my publication of the Lachesis method, it was demonstrated that Hi-C can provide long-range haplotype resolution. I foresee a near future in which molecular phasing methods become

sufficiently cost-effective and reliable so as to overtake statistical phasing methods for many applications in human genomics.

6.6 More complete gene pools

At the 2014 AGBT conference in Marco Island, Paul Blainey announced a prediction: in the future of sequencing, sample number would become a more important driver of discovery than sequencing depth. In other words, sequencing reads are now so inexpensive that any number of samples can be sequenced and analyzed at a reasonable price point, so long as the samples can be acquired and loaded into sequencing machines in a massively parallel fashion.

Blainey's prediction may sound bold, but it seems entirely reasonable in light of the current state of human genomics research. Individuals' genomes are being sequenced at a faster rate than ever before. The pages of *Nature Genetics* and the *American Journal of Human Genetics* are filled with genome-wide association studies (GWAS) and exome sequencing studies. As sequencing costs drop, whole-exome sequencing is gaining popularity over GWAS, and even whole-exome sequencing is starting to be abandoned in favor of whole-genome sequencing, which captures a fuller range of non-coding variation. It is easy to imagine a future in which whole-genome sequencing is so cheap that it is performed as a matter of course on all individuals in developed countries, and its results are readily available to researchers performing association studies. These studies gain increases in statistical power not by sequencing people to greater depth – their genotype calls are already reasonably high-quality – but by sequencing as many people as possible[189]. There is thus enormous pressure on sequencing and bioinformatics pipelines to handle ever-larger sample numbers, and per Blainey's prediction, that pressure will only increase. This will lead to new developments in high-throughput sample preparation, such as improvements in microfluidics and barcode multiplexing.

This massive increase in sample number has influenced cancer research as well. Thanks in part to the efforts of The Cancer Genome Atlas (TCGA), the number of cancer genomes sequenced has increased

prodigiously in recent years[190], and this trend is certain to continue. The numbers for cancer genomes are not as high as in association studies: while a GWAS might require 10,000 individuals or more in order to establish statistically significant associations, a cancer genome study can implicate driver genes and pathways with only tens or hundreds of genomes[191]. TCGA and other consortia have performed genome sequencing studies for every major cancer type and have identified large numbers of likely driver genes and mutations. The next great challenge in cancer genomics will be to develop high-throughput methods for annotating these genes and predicting their likely functional effects[192]. It is worth noting that these methods will be hindered by the lack of haplotype phasing analyses in these sequencing studies, which makes it impossible to determine the functional effects of compound heterozygotes. Notwithstanding our efforts on the HeLa genome, there is still no cost-effective or scalable method of genome-wide haplotype phasing in cancer genomes. A method of this type is urgently needed to enable more complete functional analyses of cancer genomes.

6.7 More complete pan-genomes

There is vast genetic variation within the human race, but it is dwarfed by the massive degree of genetic variation within microbial species, particularly bacteria. Bacteria are capable of gain or loss of individual genes rapidly, due to their short generation times and their capacity for horizontal gene transfer. This is why, within a single bacterial species, there is great inter-strain variation not merely in the form of SNVs or CNVs, but in fundamental gene content. One standard indicator of bacterial species divergence is an average nucleotide identity (ANI) of 95% or less[193]. (This allows very different organisms to be lumped together into the same species: if the 95% threshold were applied to humans, it would place us in the same species as all other catarrhine primates[194]!) Characterizing bacterial genome variation is imperative to understanding the growing problem of bacterial antibiotic resistance, as well as to efforts to discover new natural products such as antibiotics and engineered communities for biofuels and bioremediation applications.

The set of all genes that may be present in strains of a given species is referred to as that species' "pan-genome"[195]. In bacterial species with wide geographic distributions, the pan-genome may be quite large. For example, *E. coli*, the workhorse of bacterial genetics, is known to have a pan-genome several times larger than the genome of any one strain[196].

The only way to characterize the pan-genome of a bacterial species is to sequence as many isolates as possible, assemble their genomes *de novo*, and compare their genomes' gene content. Thanks to falling sequencing costs and a renewed appreciation of the importance of bacterial genome variation, the number of pan-genome sequencing studies is on the rise. My colleagues Steve Salipante and David Roach have recently led an effort[197] to sequence 312 pathogenic *E. coli* strains isolated from nosocomial infections in an ICU. They characterize the *E. coli* pan-genome and demonstrate that it is even larger than previously appreciated. I am currently collaborating with them on another ongoing project to sequence ICU isolates of a wider variety of pathogenic species (*manuscript in preparation*). We are finding that certain bacterial species such as *Rothia mucilaginosa* are like *E. coli* in that they have much more inter-species variation waiting to be discovered, while other species such as *Staphylococcus aureus* seem to have relatively small pan-genomes. These findings have lessons for treating and containing bacterial infections.

6.8 New microbial species

If the amount of variation within a bacterial species seems vast, it is still minuscule compared to the amount of variation between bacterial species. The bacterial kingdom harbors far vaster genetic variation than any other kingdom of life. New bacterial phyla are still being discovered regularly[138], and there are no doubt many others still waiting to be discovered. However, the challenge of unculturable microbial species hinders our understanding of the full scope of bacterial diversity. It is estimated[198] that fewer than 1% of bacterial species can be cultivated in isolation from the environments in which they naturally occur. There is ongoing debate about whether we could increase this number somewhat by expanding our

range of cell culturing techniques, but it is clear that most microbial species cannot be studied by culturing – and in any event, this process unavoidably changes their genomes via selective pressures. A more straightforward approach for studying these bacteria's genomes would be to sequence the cells directly. It is possible to sequence the genomes of bacteria without culturing them through the use of single-cell sequencing techniques[138], but these methods are expensive and do not provide complete genome coverage. An ideal approach would be to sequence the genomes of the entire community simultaneously and then deconvolute the individual species from the metagenome assembly.

Various computational methods have been developed to achieve metagenomic deconvolution, but the method I have developed with my colleague Ivan Liachko (see **Chapter 4**) represents the first development of a library preparation method for the purpose of metagenomic deconvolution, and it has greater power to resolve individual genomes. I believe deeply that this is an exciting next step in metagenomics and that any researcher who studies microbial communities of any type should consider applying it to their samples. To that end, Ivan and I are currently engaged in several collaborations to sequence metagenomic samples, and we are preparing manuscripts for two of these collaborations that will demonstrate the value of our method.

Whether or not I am personally responsible, I anticipate that molecular techniques for metagenomic deconvolution will soon become widely accepted and will lead to the discovery of vast new troves of microbial diversity. Untold numbers of unknown species are living all around us, as well as inside us, and are waiting to become visible to science. It is my earnest hope that these methods will give us a greater understanding of these microbes and a fuller picture of the tree of life on this planet.

6.9 Conclusion

In this dissertation I have described new methods for resolving the haplotypes of human genomes and for the *de novo* assembly of genomes of all sizes. I have introduced the first haplotype-resolved cancer genome, the first mammalian genome assembly with chromosome-scale contiguity assembled from short

reads, and the first high-contiguity microbial genomes assembled from a metagenomic mixture. None of these methods is complete or conclusive; more work needs to be done to improve their effectiveness as well as to apply them to new cases. However, I hope that these methods will enable future researchers to achieve new understanding of the endless diversity of genomic life.

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

The material in this Appendix, like that of **Chapter 3**, is based on the following peer -reviewed publication[97]:

Joshua N. Burton, Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman and Jay Shendure. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119-1125 (2013).

Supplementary methods are in section A.1. Supplementary tables are in section A.2. Supplementary figures are in section A.3.

A.1 Supplementary methods for Chapter 3

Input data sets. In the Hi-C procedure[98], DNA in a nucleus is cross-linked, then cut with a restriction enzyme, leaving pairs of distally located but physically associated DNA molecules attached to one another. The sticky ends of these fragments are biotinylated and then ligated to each other to form chimeric circles. Biotinylated circles are enriched for, sheared again, and then processed to sequencing libraries in which individual templates are chimeras of the physically associated DNA molecules from the original cross-linking.

Four Hi-C data sets were used, corresponding to human cells, mouse cells, *Drosophila* tissue and HeLa cells. The human data set was produced from human ESCs (hESCs)[115]. The hESC replicates 1 and 2 were used (NCBI SRA accessions: GSM862723, GSM892306) for a total of 734 M read-pairs. The mouse data set was produced from mouse ESCs (mESCs)[115]. The mESC replicates 1 and 2 were used (NCBI SRA accessions: GSM862720, GSM862721) for a total of 806 M read-pairs. The *Drosophila* data set was produced from embryos[118] and includes 363 M read-pairs (NCBI SRA accession: SRX111555). The HeLa data set was produced as part of this study (see "Chromosome Fusion Detection in HeLa", below) and includes 305 M read-pairs.

Two types of shotgun assemblies were created as inputs to LACHESIS. First, we created shotgun assemblies for human, mouse and *Drosophila* by downloading the appropriate sequence libraries from SRA and assembling them with ALLPATHS-LG. **Table 3.1** shows statistics for these three assemblies. Second, simulated shotgun assemblies were made by breaking up the human reference genome into contigs of varying sizes, ranging from 10 Kb to 1 Mb. **Table 3.2** shows statistics for these assemblies.

Shotgun assemblies. To create the human shotgun assembly, we downloaded the sequence files[67] corresponding to the fragment library and two short jumping libraries for individual NA12878 from the NCBI Short Read Archive (NCBI SRA accession SRA024407). The files were converted from sra to fastq format, and formatted as required by the ALLPATHS-LG assembler using the PrepareAllPathsInputs.pl script included with the ALLPATHS-LG distribution. The reads were assembled using the ALLPATHS-LG

assembler[67] (version r41985) with the following parameters (the rest being default): HAPLOIDIFY = TRUE, MAX_MEMORY_GB = 400, THREADS = 32, EVALUATION = STANDARD. Insert size estimates (mean and s.d.) for each library were specified based on the values provided previously[67]. Scaffolds in this assembly were treated as contigs by LACHESIS. Because we intentionally excluded fosmid end sequencing data, this assembly had far less mid-range contiguity than the full *de novo* assembly produced previously[67] (N50 scaffold length 437 Kb versus 11.5 Mb), and thus it more closely represents a typical *de novo* assembly created exclusively from *in vitro* libraries.

To create the mouse shotgun assembly, we downloaded the sequence files[67] corresponding to the fragment and three short jumping libraries from the NCBI Short Read Archive (NCBI SRA accession SRA009956). The libraries were assembled using the ALLPATHS-LG assembler[67] (version r41985) with the following parameters (the rest being default): HAPLOIDIFY = TRUE, MAX_MEMORY_GB = 500, THREADS = 32. Insert size estimates (mean and s.d.) for each library were specified based on the values provided previously[67].

To create the *Drosophila* shotgun assembly, we downloaded the sequence files for *Drosophila*(*Drosophila* Genomic Reference Panel[117] corresponding to sequencing runs SRR516038 (Sample DGRP-348) and SRR516001 (Sample DGRP-821) from the NCBI Short Read Archive. SRR516038 served as the "fragment" library as per ALLPATHS-LG terminology. The ALLPATHS-LG assembler also requires a "jumping" library. We were unable to find a previously sequenced jumping library for *Drosophila*. As a work-around, we used a standard shotgun library with a slightly higher insert size (SRR516001) and artificially converted it into a jumping library by flipping the orientation of reads. All files were first converted from sra to fastq format, then formatted as required by the ALLPATHS-LG assembler using the PrepareAllPathsInputs.pl script included with the ALLPATHS-LG distribution. Insert size distributions for these libraries (mean = 205 bp, s.d. = 25 bp for fragment library; mean = 320 bp, s.d. = 52 bp for jumping library) were obtained by aligning a subset of reads to the *Drosophila* reference genome using BWA[199]. The reads were assembled using the ALLPATHS-LG assembler[67] (version r41985) with the following parameters (the rest being default): HAPLOIDIFY = TRUE, MAX_MEMORY_GB = 300, THREADS = 16, VAPI_WARN_ONLY = True.

Aligning Hi-C reads. Hi-C reads were aligned to shotgun assemblies or reference genomes using BWA[199] with default parameters. Reads were considered artifactual if they did not align within 500 bp of a restriction site, as recommended[116]. Non-uniquely aligning reads were assigned a mapping quality of 0 by BWA and were excluded from subsequent analysis. Additionally, read-pairs were considered for downstream analysis only if both reads in the pair aligned to contigs from the assembly.

Clustering contigs or scaffolds into chromosome groups. Contigs or scaffolds (the term 'contig' is used in this description of the method to indicate both possibilities) were placed into groups using hierarchical clustering (Figure A.3.1). A graph was built, with each node initially representing one contig, and each edge between nodes having a weight equal to the number of Hi-C read-pairs linking the two contigs. The contigs were merged together using hierarchical agglomerative clustering with an average-linkage metric[114], which was applied until the number of groups was reduced to the expected number of distinct chromosomes (counting only groups with more than one contig). Repetitive contigs (contigs whose average link density with other contigs, normalized by number of restriction fragment sites, was greater than two times the average link density) and contigs with too few restriction fragment sites (<5 for the simulated human assembly; <25 for the human and mouse *de novo* assemblies; <250 for the *Drosophila* assembly) were not clustered. However, after clustering, each of these contigs was assigned to a group if its average link density with that group was greater than four times its average link densities with any other group.

Ordering contigs or scaffolds within chromosome groups. Each group of contigs or scaffolds (the term 'contig' is used in this description of the method to indicate both possibilities) was ordered using the following algorithm (Figure A.3.2). First, a graph was built as in the clustering step, but with the edge weights between nodes equal to the inverse of the number of Hi-C links between the contigs, normalized by the number of restriction fragment sites per contig. Short contigs (<5 restriction fragment sites for the simulated human assemblies; <20 sites for the human and mouse *de novo* assemblies; <20 Kb for the *Drosophila de novo* assembly) were excluded from this graph. A minimum spanning tree was calculated for this graph. The longest path in this tree, the "trunk", was found. The spanning tree was then modified

so as to lengthen the trunk by adding to it contigs adjacent to the trunk, in ways that kept the total edge weight heuristically low.

After a lengthened trunk was found for each group, it was converted into a full ordering as follows. The trunk was removed from the spanning tree, leaving a set of "branches" containing all contigs not in the trunk. These branches were reinserted into the trunk, the longest branches first, with the insertion sites chosen so as to maximize the number of links between adjacent contigs in the ordering. Short fragments (<5 restriction fragment sites for the simulated human assemblies; <20 sites for the human and mouse *de novo* assemblies; <40 Kb for the *Drosophila de novo* assembly) were not reinserted; as a result, many small contigs that were clustered were left out of the final LACHESIS assembly.

Orienting contigs or scaffolds. The orientation of each contig or scaffold (the term 'contig' is used in this description of the method to indicate both possibilities) within its ordering was determined by taking into account the exact position of the Hi-C link alignments on each contig (Figure A.3.3). It was assumed that, as demonstrated in previous Hi-C studies[98], the likelihood of a Hi-C link connecting two reads at a genomic distance of x is roughly $1/x$ for $x \geq$ ~100 Kb. A weighted, directed, acyclic graph (WDAG) was built representing all possible ways to orient the contigs in the given order. Each edge in the WDAG corresponded to a pair of adjacent contigs in one of their four possible combined orientations, and the edge weight was set to the log-likelihood of observing the set of Hi-C link distances between the two contigs, assuming they were immediately adjacent with the given orientation.

For each contig, a quality score for its orientation was calculated as follows. The log-likelihood of the observed set of Hi-C links between this contig, in its current orientation, and its neighbors, was found. Then the contig was flipped and the log-likelihood was calculated again. The first log-likelihood was guaranteed to be higher because of how the orientations were calculated. The difference between the log-likelihoods was taken as a quality score.

Validation. To determine the true position of the contigs or scaffolds in the shotgun assemblies, we aligned them to the human, mouse or *Drosophila* reference genome using BLASTn[200] with parameters `-perc_identity 99 -evalue 100 -word_size 50`. For each contig, a "truth placement" on reference was

derived as follows. First, the chromosome was chosen containing the plurality of aligned sequence from the contig. Second, the single best alignment to this chromosome (measured by *E*-value) was used to "seed" a chromosomal region. Third, the other alignments to this chromosome were considered by descending *E*-value, and the region was extended to include as many of them as possible without exceeding the total length of the assembly contig.

Chromosome fusion detection in HeLa. A single, complex Hi-C library was constructed for the HeLa S3 cancer cell line (ATCC CCL2.2; grown in DMEM with 10% FBS and 1× Pen. Strep.) according to a published[201] protocol. This library was sequenced on two lanes of Illumina HiSeq 2000, followed by read trimming to 50 bp to eliminate ligation-spanning reads that confound alignment. Reads were aligned to the human reference genome using BWA[199] with default parameters, followed by removal of PCR duplicates. Reads were then assigned to genomic windows containing approximately one megabase of sequence (mean = 955,176 bp) that were determined by bases of unique mappability to the genome. Links between windows were normalized first to the number of HindIII restriction sites present in the window to account for biases inherent to restriction-based library preparation, then to the total count of short pairs within the window (defined as pairs with an insert size ≤1 Kb) to account for the underlying copy number of the window.

Rearrangements were called by first identifying stretches of ≥10 consecutive windows within a row where ≥80% of windows have a link score ≥1 s.d. above the mean of the entire row. Stretches of windows present in columns were called using the same parameters. Windows present in outlier stretches for both rows and columns were defined as outlier windows. These windows were then clustered with all proximal windows ≤2 windows away and the outlier window count and density within the outer borders of the cluster determined. Outlier spans and clusters are shown in Figure A.2.12.

A.2 Supplementary tables for Chapter 3

Metric		*De novo* assemblies			
		Human	Human hi-contiguity, from ref. 65	Mouse	*Drosophila*
Assembly metrics	Total assembly length, gapped (Mb)	2,739	2,773	2,370	127.2
	Length / reference length	93.0%	94.1%	87.0%	75.4%
	N. contigs or scaffolds	18,921	3,811	25,964	7,109
	N50 contig/scaffold, ungapped (Kb)	437	11,547	224	68
% sequence (% contigs) clustered into groups		98.2% (71.5%)	99.0% (53.3%)	98.0% (87.8%)	81.2% (64.3%)
% clustered sequence (% contigs) mis-clustered		0.14% (1.4%)	4.7% (7.9%)	0.24% (0.5%)	3.4% (10.5%)
Full orders	% clustered sequence (% contigs) ordered	94.4% (55.3%)	99.4% (28.6%)	86.7% (42.7%)	82.0% (24.5%)
	% ordered sequence (% contigs) w/ ordering errors	0.5% (0.8%)	8.4% (9.5%)	0.5% (1.1%)	4.6% (5.2%)
	% ordered sequence (% contigs) w/ orientation errors	1.2% (2.5%)	6.4% (10.3%)	1.9% (4.6%)	4.1% (6.1%)
	% ordered sequence (% contigs) w/ high quality	92.8% (79.0%)	88.4% (51.6%)	93.3% (82.9%)	94.1% (88.1%)
	% high-quality sequence (% contigs) w/ ordering errors	0.3% (0.4%)	4.7% (3.7%)	0.3% (0.7%)	3.3% (3.4%)
	% high-quality sequence (% contigs) w/ orientation errors	0.4% (0.5%)	3.4% (3.3%)	0.5% (1.0%)	2.5% (2.7%)
Trunks	% ordered sequence (% contigs) in trunks	88.4% (88.5%)	82.4% (76.2%)	90.4% (88.4%)	70.7% (70.6%)
	% sequence in trunks (% contigs) w/ ordering errors	0.2% (0.4%)	5.3% (7.5%)	0.2% (0.4%)	3.0% (4.0%)
	% sequence in trunks (% contigs) w/ orientation errors	1.1% (2.3%)	2.8% (7.5%)	1.7% (4.2%)	1.9% (3.5%)
	% sequence in trunks (% contigs) w/ high quality	93.0% (79.4%)	92.4% (56.8%)	93.6% (83.5%)	94.7% (89.6%)
	% high-quality sequence in trunks (% contigs) w/ ordering errors	0.1% (0.2%)	3.0% (2.0%)	0.1% (0.2%)	2.1% (2.5%)
	% high-quality sequence in trunks (% contigs) w/ orientation errors	0.3% (0.3%)	1.0% (0.8%)	0.4% (0.8%)	1.1% (1.6%)

Table A.2.1 | Metrics for the *LACHESIS* scaffolding results. This is a more detailed version of **Table 3.1**. Results for the human *de novo* assembly exclude the chimeric group not shown in **Figure 3.3**.

Figure	Dominant chrom(s)	Sequence length in grouped scaffolds				Sequence length in ordered scaffolds			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom(s)	Other chroms	None		Dominant chrom(s)	Other chroms	None
3a	chr1	210.9	99.9%	0.01%	0.07%	202.6	100%	-	-
3b	chr2	224.4	99.9%	0.02%	0.05%	216.8	100%	-	-
3c	chr3	190.6	99.3%	0.6%	0.02%	182.8	99.3%	0.7%	-
3d	chr4	181.0	99.98%	0.01%	0.01%	173.6	100%	-	-
3e	chr5	170.5	99.9%	0.01%	0.09%	162.1	100%	-	-
3f	chr6	164.9	99.2%	0.8%	0.02%	156.8	99.2%	0.8%	-
3g	chr7	143.9	99.8%	0.03%	0.18%	134.8	100%	-	-
3h	chr8	136.7	99.8%	0.15%	0.01%	131.3	99.9%	0.1%	-
3i	chr9	106.7	99.9%	0.03%	0.09%	101.0	100%	-	-
3j	chr10	125.9	99.6%	0.3%	0.09%	119.8	99.8%	0.2%	-
3k	chr11	125.7	99.9%	0.01%	0.10%	118.3	99.99%	0.01%	-
3l	chr12	126.0	99.9%	0.1%	0.04%	119.8	99.9%	0.1%	-
3m	chr13	93.9	99.96%	0.007%	0.03%	92.2	100%	-	-
3n	chr14	84.8	99.7%	0.2%	0.05%	81.4	99.8%	0.2%	-
3o	chr15	75.5	99.8%	0.01%	0.2%	71.0	100%	-	-
3p	chr16	68.3	99.6%	0.06%	0.3%	64.3	100%	-	-
3q	chr17	73.4	99.7%	0.1%	0.2%	65.9	100%	-	-
3r	chr18	72.4	99.95%	0.02%	0.04%	70.8	100%	-	-
3s	chr19, chr22	82.8	99.9%	0.1%	0.03%	67.9	57.6%, 42.4%	-	-
3t	chr20, chr21	91.2	99.8%	0.2%	0.01%	88.0	63.2%, 36.6%	0.2%	-
3u	chrX	36.7	99.9%	0.03%	0.05%	34.8	100%	-	-
3v	chrX	104.5	99.5%	0.01%	0.4%	90.9	100%	-	-
Supp. Figure 4w	chr16	6.5	23.5%	47.6%	29.0%	2.3	42.8%	54.2%	3.0%

Table A.2.2 | Contents of *LACHESIS*' orderings in the human *de novo* assembly (Figure 3.3). For each of the 23 groups, there is a "dominant chromosome" in the reference genome to which the plurality of alignable sequence aligns. This chart shows what fraction of the scaffold length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no chromosomes. The last row corresponds to the small, chimeric chromosome group described in the main text.

Repeat type	UCSC Genome Browser track name	Enrichment near the edges of mis-ordered scaffolds
Segmental duplications (>1 Kb length, >90% similarity)	Segmental Dups	**6.38**
Microsatellite repeats (dinucleotide, trinucleotide)	Microsatellite	**1.24**
Simple tandem repeats (4 or more nucleotides)	Simple Repeats	**2.87**
RepeatMasked regions	RepeatMasker	**0.93**
Interrupted repeats called by RepeatMasker	Interrupted Rpts	**0.94**

Table A.2.3 | Enrichment of repetitive sequences in error-prone regions. Human genomic regions corresponding to several different types of repetitive sequence elements were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). For each scaffold in the human *de novo* assembly created by *LACHESIS*, a 5 Kb region was extracted around each of its ends. These edge regions were then overlapped with the repeat elements. The enrichment shown for each type of repeat element is the ratio of the frequency with which that element co-occurs with the ends of one of the 61 scaffolds marked with ordering errors, divided by the frequency with which it co-occurs with the ends of one of the 7,604 scaffolds not marked with ordering errors.

Supp. Figure	Dominant chrom(s)	Sequence length in grouped scaffolds				Sequence length in ordered scaffolds			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom(s)	Other chroms	None		Dominant chrom(s)	Other chroms	None
7a	chr1	176.5	99.7%	0.3%	0.02%	150.8	99.7%	0.3%	-
7b	chr2	167.3	99.3%	0.6%	0.1%	149.4	99.4%	0.6%	-
7c	chr3	142.0	99.8%	0.2%	0.04%	119.8	99.9%	0.1%	-
7d	chr4	136.2	99.9%	0.008%	0.1%	118.0	100%	-	-
7e	chr5	136.2	99.93%	0.02%	0.05%	119.1	100%	-	-
7f	chr6	134.6	99.7%	0.2%	0.1%	114.1	99.8%	0.2%	-
7g	chr7	120.2	99.8%	0.01%	0.2%	102.4	100%	-	-
7h	chr8	119.6	97.8%	2.2%	0.04%	106.5	97.6%	2.4%	-
7i	chr9	113.9	99.93%	0.01%	0.06%	103.1	100%	-	-
7j	chr10	116.5	99.8%	0.1%	0.1%	101.0	99.9%	0.1%	-
7k	chr11	113.8	99.5%	0.4%	0.06%	106.8	99.5%	0.5%	-
7l	chr12	104.7	99.9%	0.02%	0.1%	89.9	100%	-	-
7m	chr13	106.1	99.8%	0.02%	0.2%	91.9	100%	-	-
7n	chr14	106.2	99.8%	0.002%	0.2%	92.1	100%	-	-
7o	chr15	95.2	99.96%	0.03%	0.02%	83.7	100%	-	-
7p	chr16	89.6	99.99%	0.003%	0.004%	79.0	100%	-	-
7q	chr17	84.3	99.7%	0.06%	0.2%	73.1	100%	-	-
7r	chr18	82.3	99.93%	0.06%	0.003%	73.1	100%	-	-
7s	chr19	55.6	99.94%	0.01%	0.04%	50.3	100%	-	-
7t	chrX	122.4	99.7%	0.1%	0.2%	90.9	99.9%	0.1%	-

Table A.2.4 | Contents of *LACHESIS*' orderings in the mouse *de novo* assembly (Figure A.3.7). For each of the 20 groups, there is a "dominant chromosome" in the reference genome to which the plurality of alignable sequence aligns. This chart shows what fraction of the scaffold length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no chromosomes.

Supp. Figure	Dominant chrom	Sequence length in grouped contigs				Sequence length in ordered contigs			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom	Other chroms	No euchromatic sequence		Dominant chrom	Other chroms	No euchromatic sequence
9b	X	18.4	75.8%	2.3%	21.9%	12.9	85.0%	2.7%	12.3%
9c	4	2.5	46.5%	21.9%	31.7%	0.74	93.2%	4.0%	2.8%
9d	2	41.1	58.4%	1.9%	39.7%	32.6	71.1%	2.2%	26.7%
9e	3	40.4	82.6%	2.0%	15.4%	37.5	85.2%	1.9%	12.8%

Table A.2.5 | Contents of *LACHESIS'* orderings in the *D. melanogaster de novo* assembly (Figure A.3.9). For each of the four groups, there is a "dominant chromosome" in the reference genome to which the majority of euchromatic sequence aligns. This chart shows what fraction of the contig length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no euchromatic chromosome. Note that a substantial fraction of the *D. melanogaster* assembly may consist of heterochromatic sequences as it does not align to the four euchromatic chromosomes of the reference assembly.

Number of Hi-C pairs, before filtering	Percent of total Hi-C coverage	% (by length) of sequence clustered	Clustering error rate, excluding fusions	% (by length) of sequence ordered	Ordering error rate	Orienting error rate
51,493,359	7.0%	95.97%	0.36%	92.53%	14.8%	12.5%
113,961,921	15.5%	96.97%	0.25%	92.72%	6.3%	6.4%
175,873,230	24.0%	97.08%	0.16%	92.81%	4.6%	5.0%
237,662,270	32.4%	97.13%	0.16%	92.79%	4.0%	4.6%
404,341,129	55.1%	97.92%	0.14%	92.95%	1.0%	1.7%
568,435,079	77.4%	98.04%	0.13%	92.96%	0.8%	1.4%
734,185,216	100%	98.22%	0.15%	93.02%	0.5%	1.2%

Table A.2.6 | The effect of Hi-C down-sampling on *LACHESIS* assembly quality. *LACHESIS* was provided with varying quantities of Hi-C read coverage with which to scaffold the shotgun human assembly. As read coverage increased, the total amount of sequence placed by *LACHESIS* increased slightly, while error rates decreased significantly. The bottom row describes the same assembly as in **Figure 3.2a, 3.2b** and **Table A.2.2**.

A.3 Supplementary figures for Chapter 3

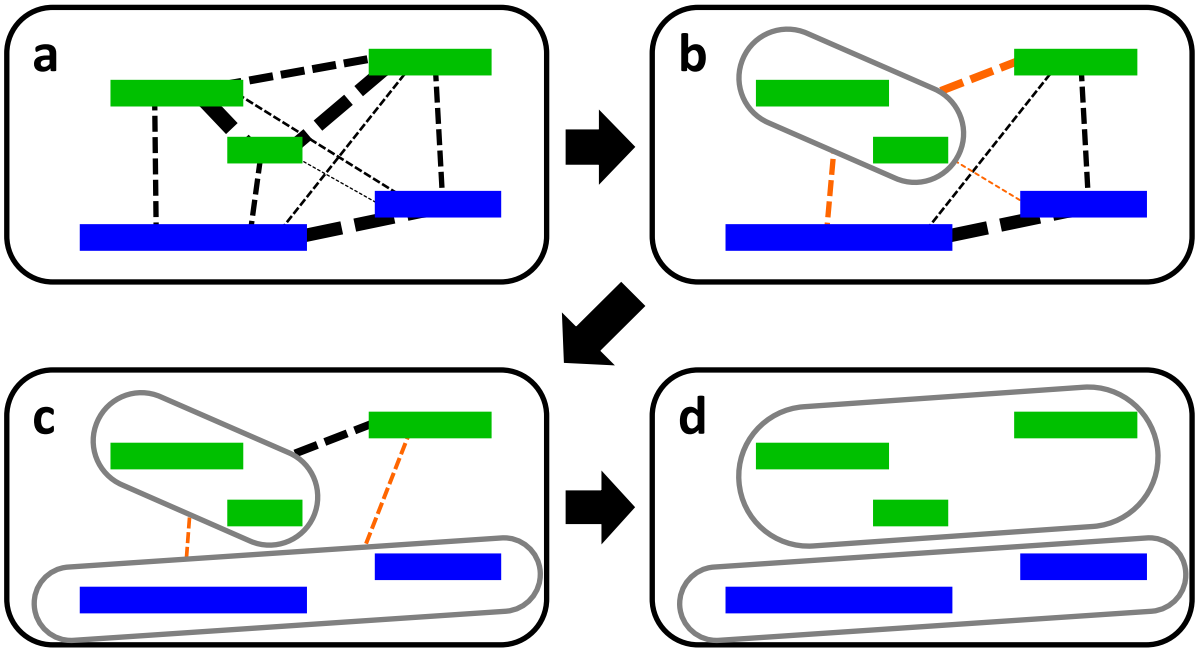


Figure A.3.1 | An illustrated overview of the *LACHESIS* clustering algorithm. a. An assembly consisting of five contigs, which in truth belong to two chromosomes (green and blue). Hi-C links between the contigs are shown as black dotted lines, with thicker lines indicating higher normalized link density. **b.** The agglomerative hierarchical clustering algorithm begins. The two contigs sharing the highest normalized link density are merged together to create a cluster (gray oval). The new link densities between this cluster and each other contig (orange dotted lines) are calculated as the average (normalized) linkage between the two contigs in this cluster and the other contig. **c.** Again, the two contigs sharing the highest normalized link density are merged to create a cluster. New average link densities are calculated (orange dotted lines); note that the link density between the two multi-contig clusters is the average of four original link densities. **d.** Another merge. The user-specified limit of two clusters has been reached, so the algorithm is complete. It has correctly found groups for each chromosome.

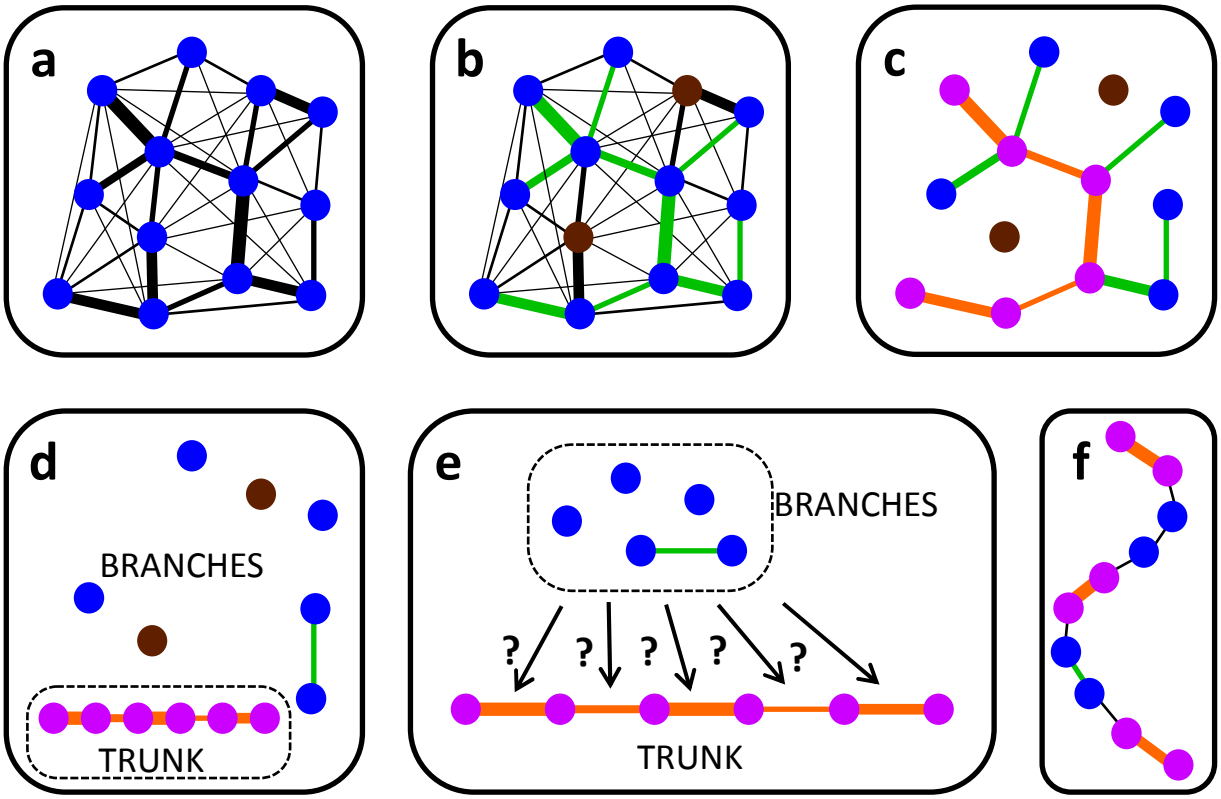


Figure A.3.2 | An illustrated overview of the *LACHESIS* ordering algorithm. a. A group of contigs depicted as a graph. Each blue vertex indicates a contig, and the edges between vertices indicate normalized Hi-C link densities (for clarity, edges are not shown between all pairs of contigs). **b.** A spanning tree (a set of edges that connects all vertices with no loops) is found (green edges). The edges of the spanning tree are chosen to have the maximum possible link densities. Short contigs (dark brown dots) are not included in the spanning tree. **c.** The longest path in the spanning tree (magenta dots, orange edges) is found. This path constitutes the "trunk", an initial contig ordering with high accuracy but low completeness. **d.** The trunk is removed from the spanning tree, leaving a set of vertices and edges called "branches", many of which consist of a single isolated vertex. **e.** Lastly, the branches are considered for reinsertion into the trunk at all possible positions and orientations. Each possible reinsertion site is given a "score" equal to the sum of the reciprocals of all link distances. Very short branches are not reinserted. **f.** The final contig ordering.

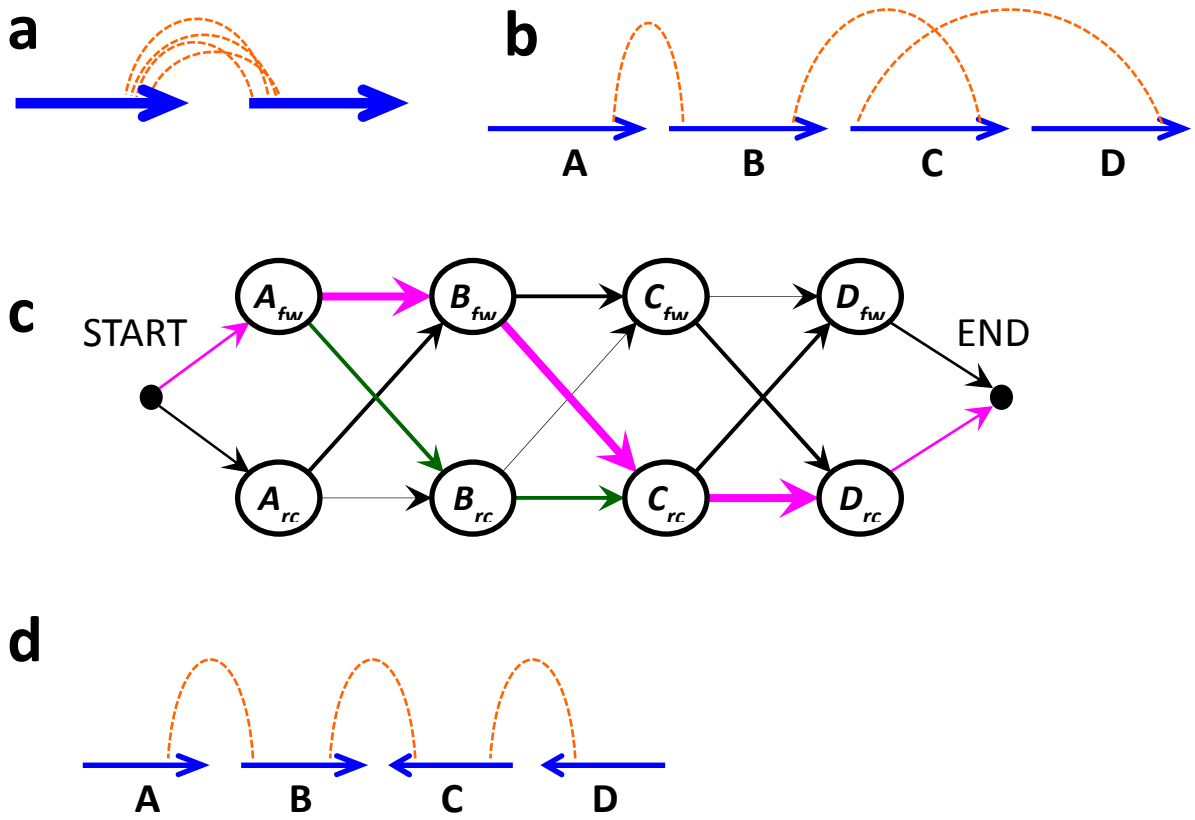


Figure A.3.3 | An illustrated overview of the *LACHESIS* orienting algorithm. a. A pair of contigs connected by several Hi-C links, with the exact location of the aligned Hi-C reads shown (orange dotted lines). All of the reads in these links are localized to one end of each contig, which suggests that the contigs should be placed in the orientation shown; any other orientation would increase the perceived length of the links. Note that this is the only time *LACHESIS* uses the exact location of the reads in a Hi-C link, as opposed to the mere fact of a link between two contigs. **b.** An ordering of four contigs A,B,C,D, with arbitrary initial orientations. The exact locations of the Hi-C read alignments between adjacent contigs are shown (for clarity, only one link per adjacency is shown). **c.** A weighted directed acyclic graph (WDAG) describing all possible ways in which these four contigs could be ordered. The edges exiting the start node and entering the end note all have the same weight. The edge weights between each pair of contigs (arrows) are set to the log-likelihoods of observing the Hi-C links between those two contigs in the two orientations, given that longer links are less likely; larger numbers (thicker arrows) indicate more likely orientations. The likeliest path through the WDAG (magenta arrows) is shown. The orientation quality score is calculated as the differential to the log-likelihood caused by choosing a particular orientation; for example, for contig B, the log-likelihood is the difference between the weights of the magenta arrows entering and leaving node $\boldsymbol{B}_{fw}$ and the weights of the alternative nodes entering and leaving $\boldsymbol{B}_{rc}$ (dark green arrows). **d.** The contig orientations corresponding to the likeliest path found in **c**.

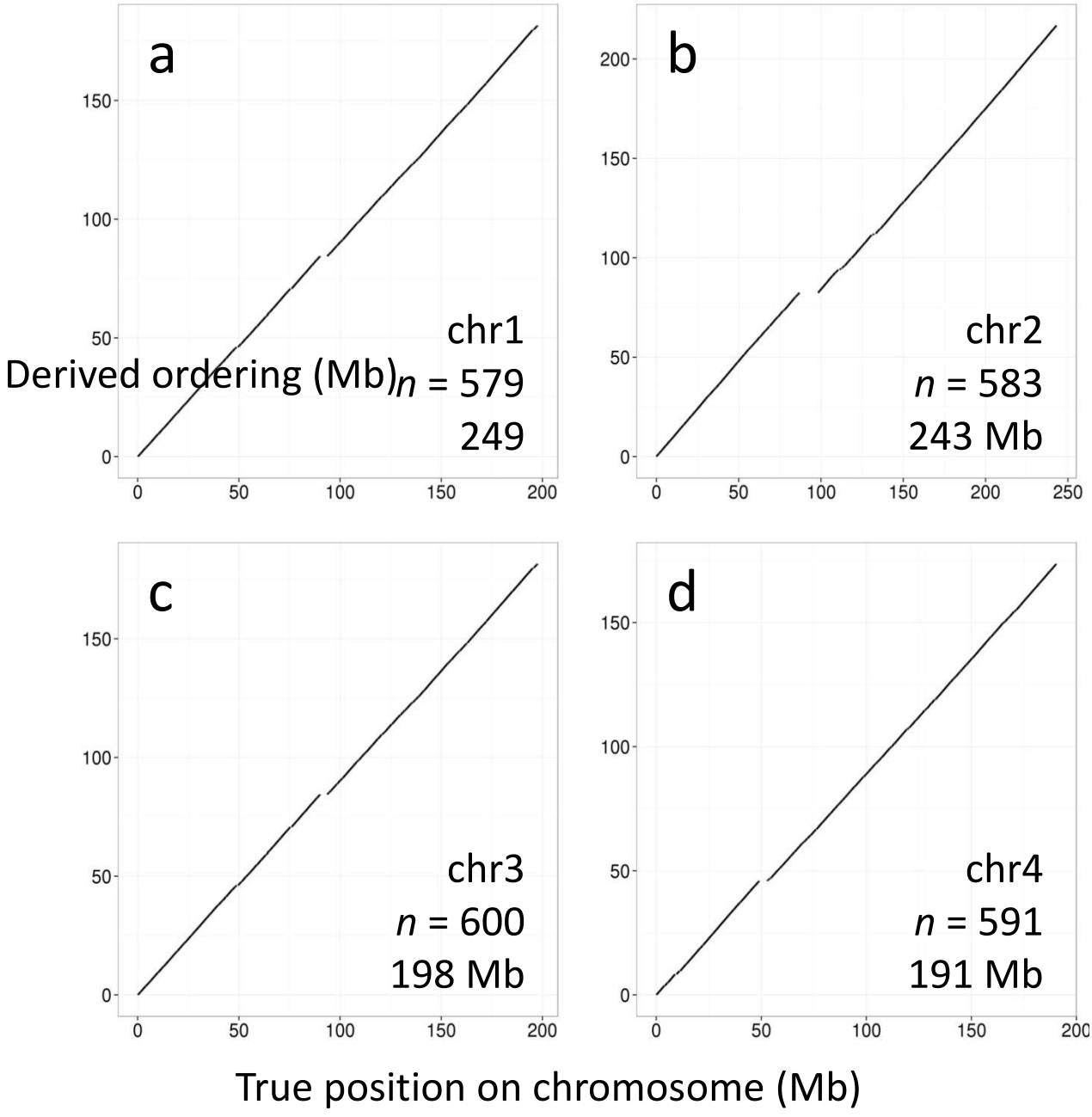


Figure A.3.4 (page 1 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

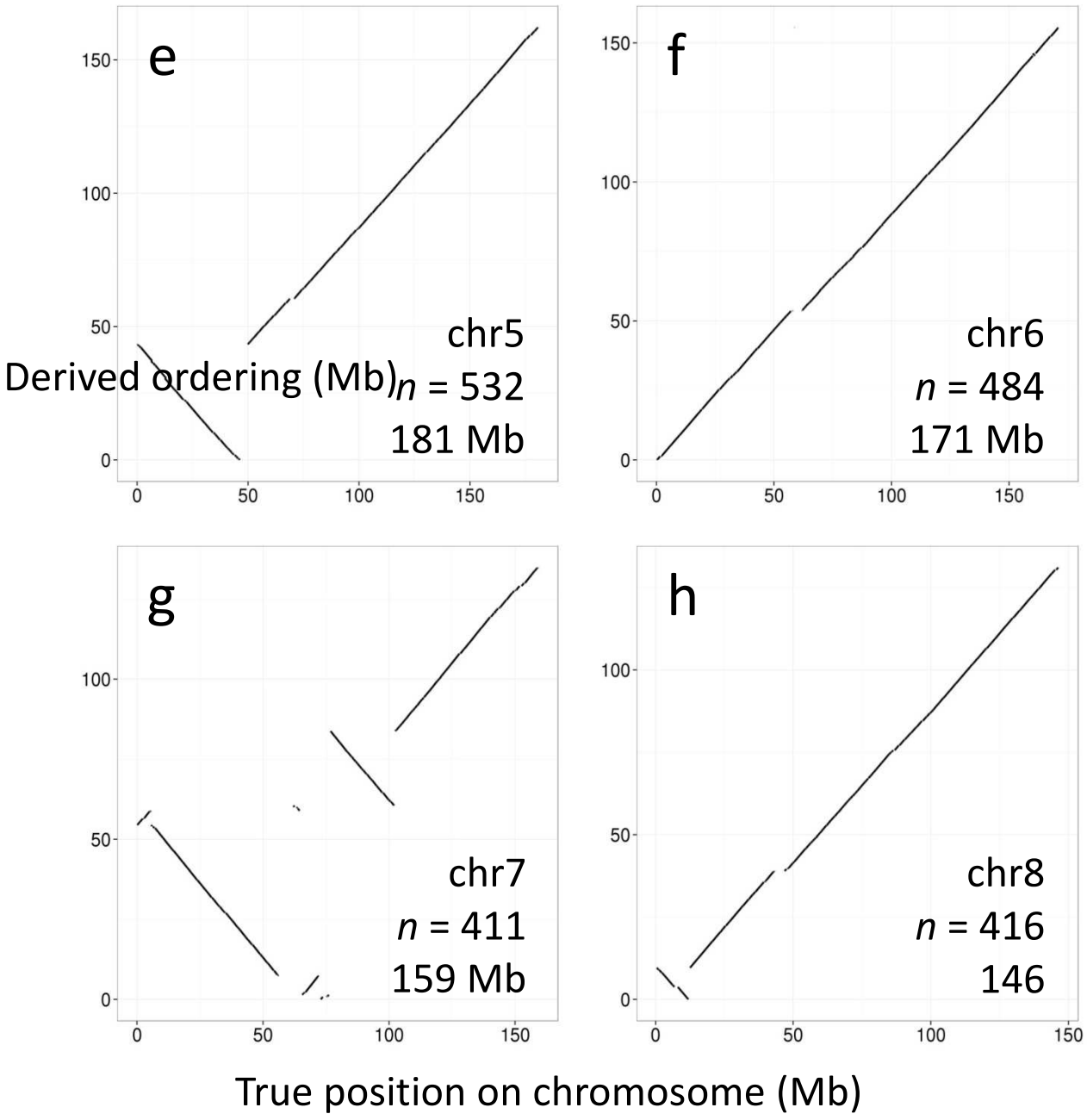


Figure A.3.4 (page 2 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

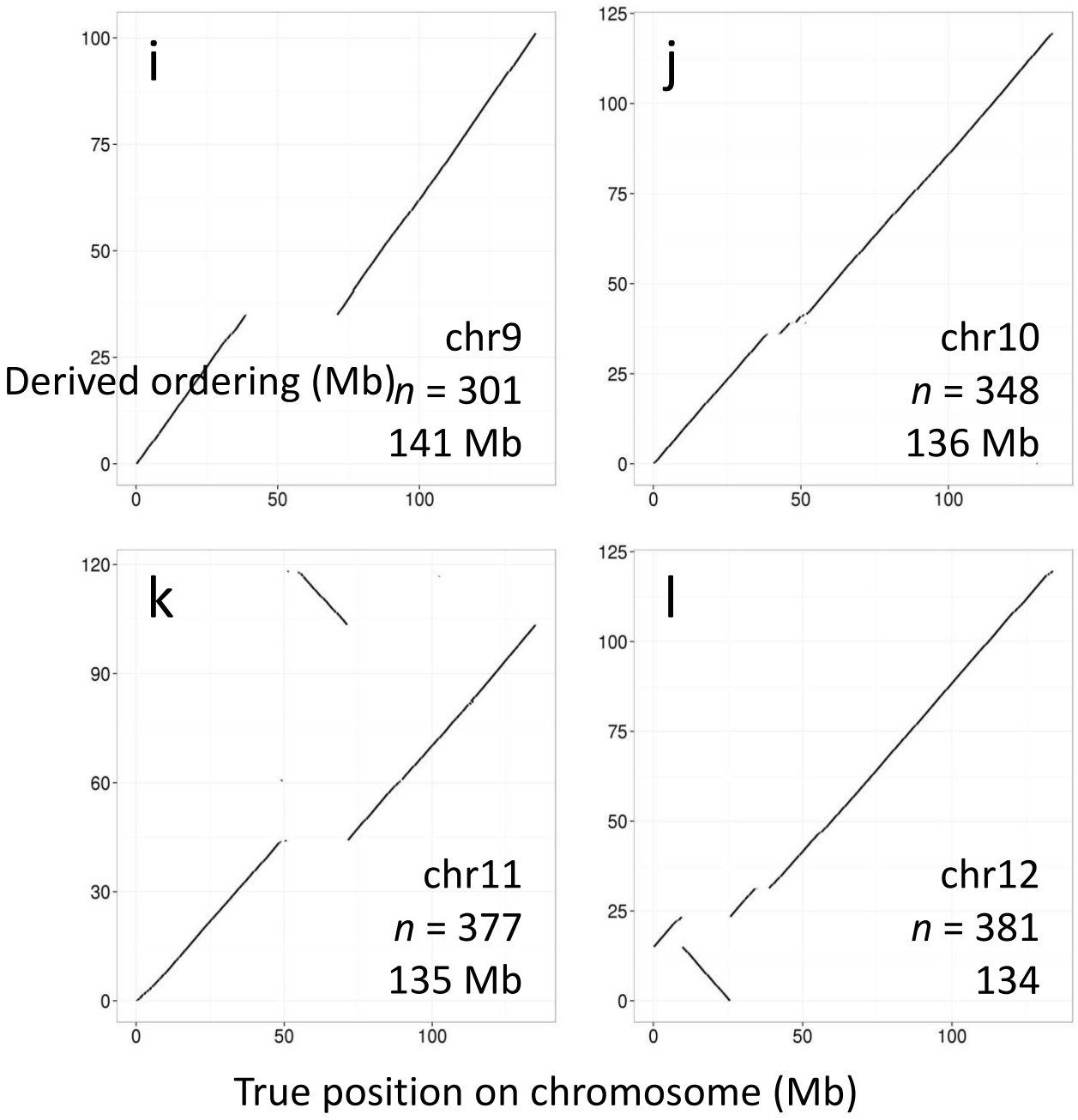


Figure A.3.4 (page 3 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

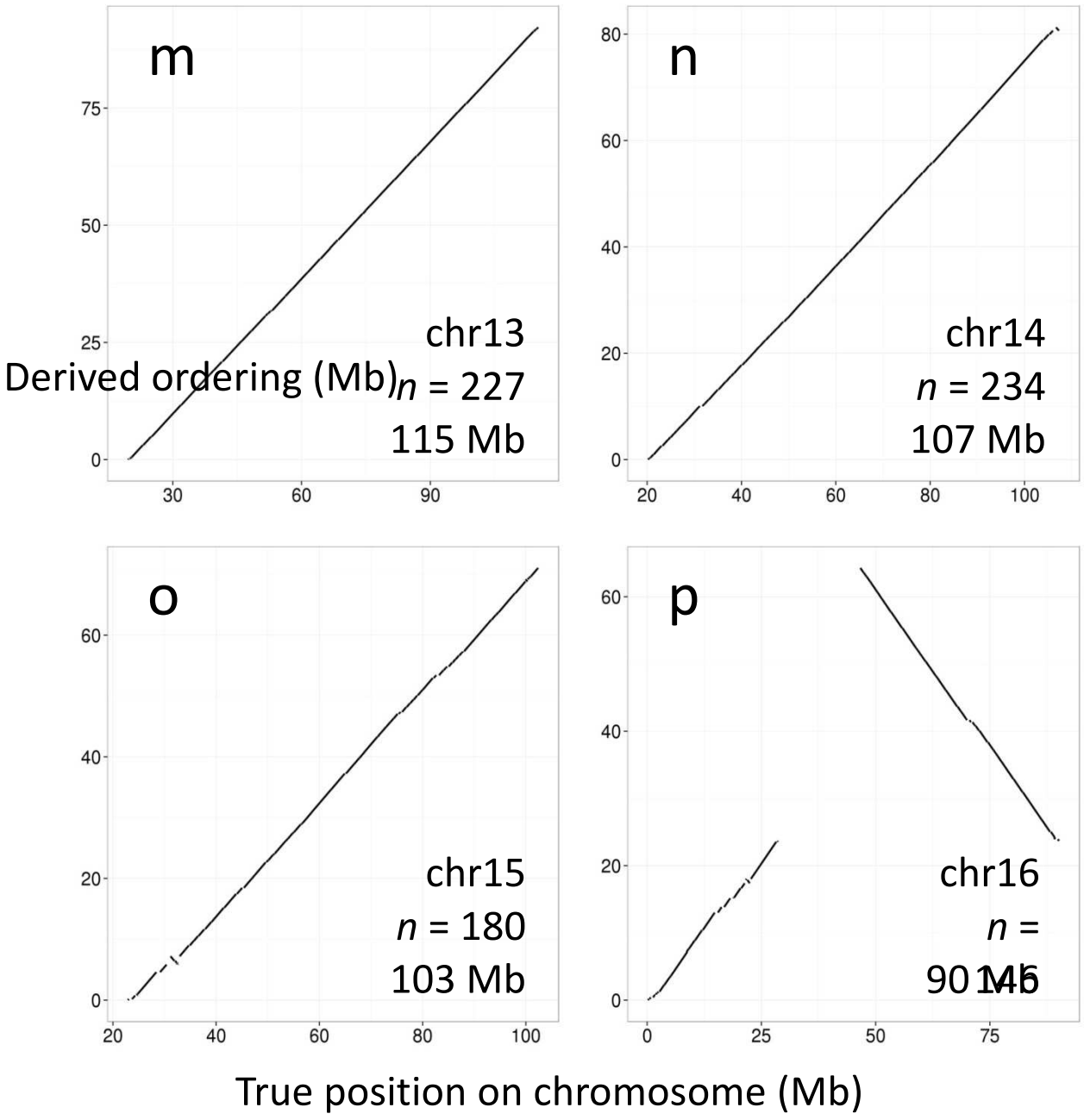


Figure A.3.4 (page 4 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

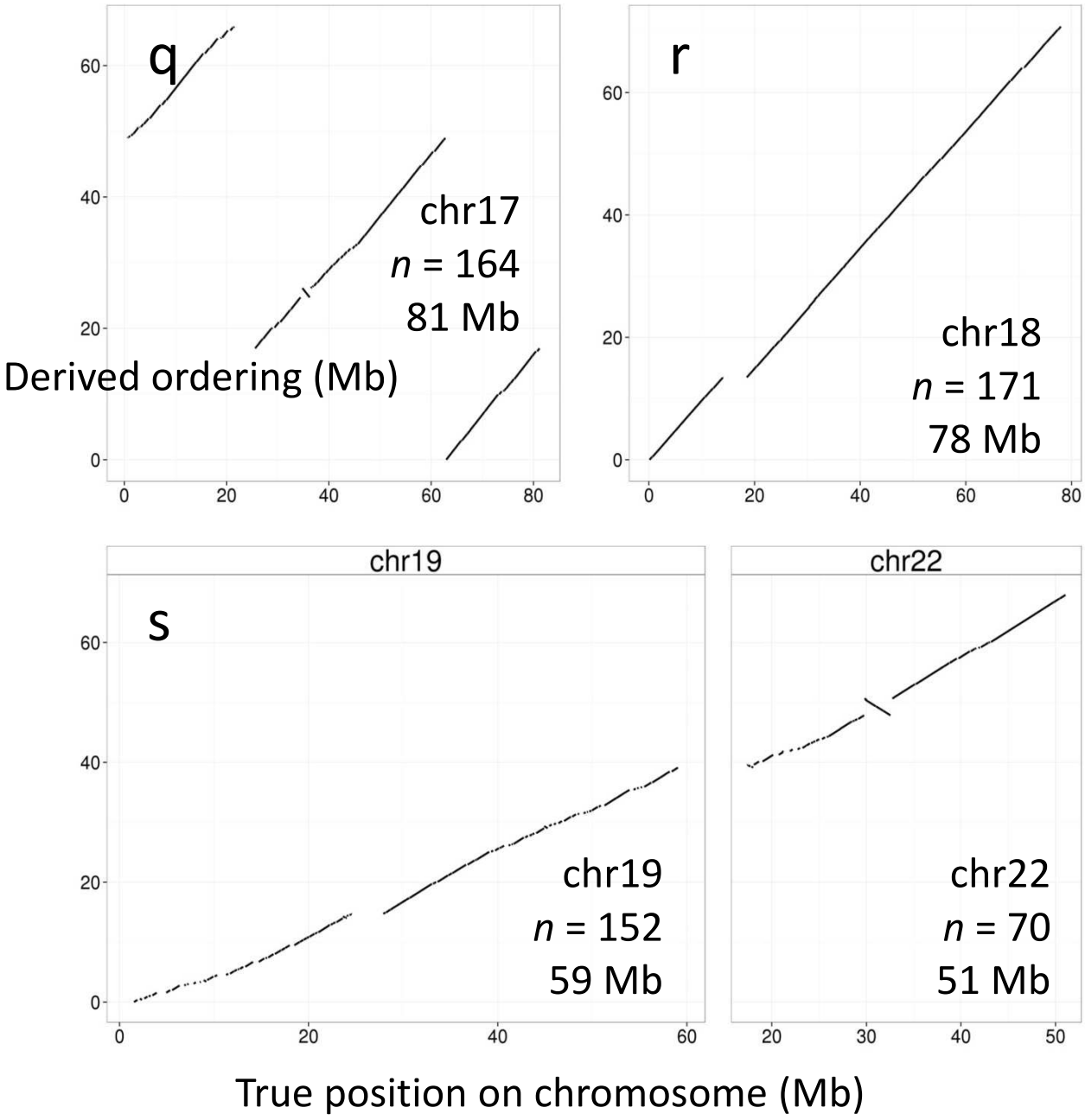


Figure A.3.4 (page 5 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

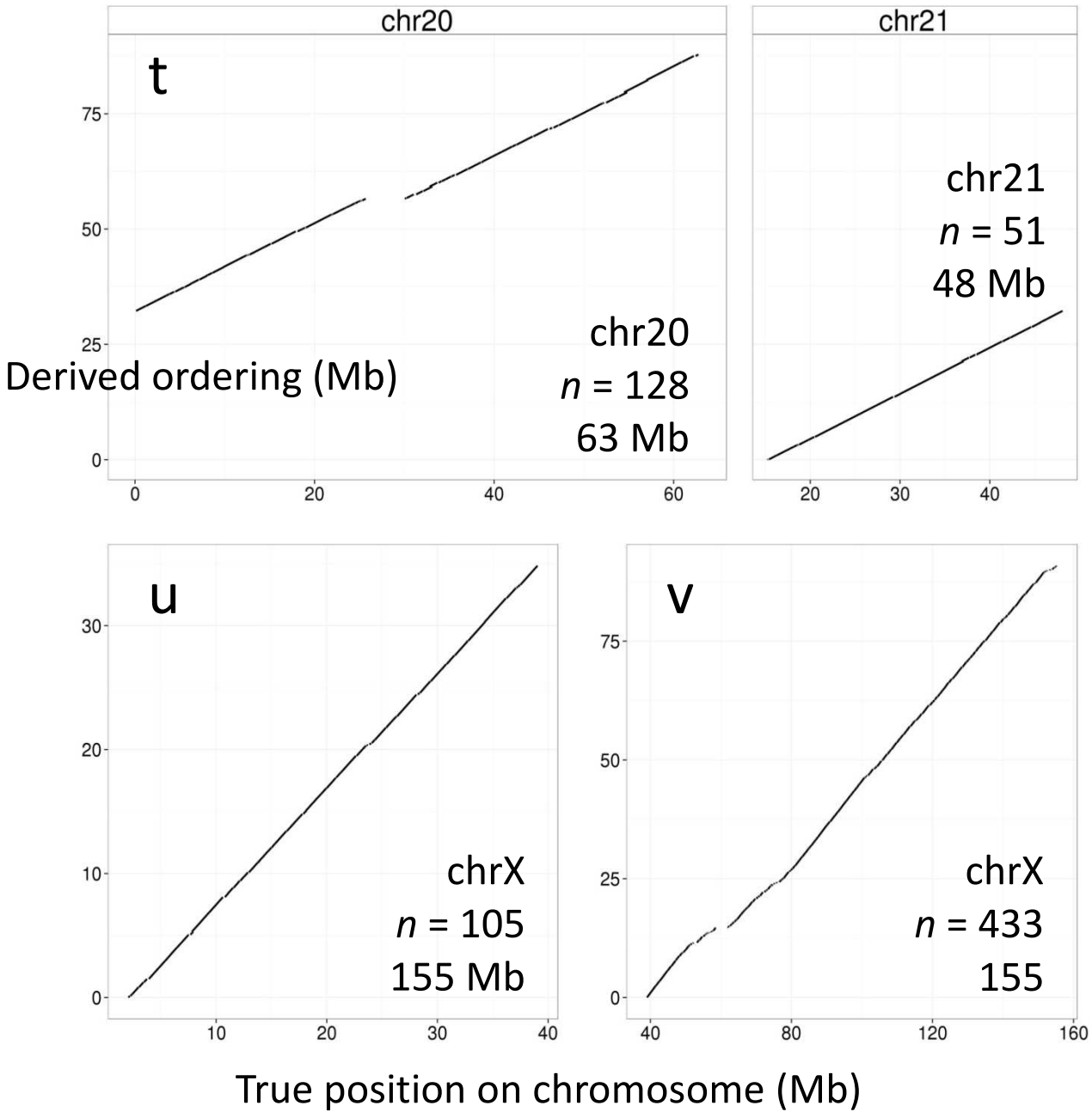


Figure A.3.4 (page 6 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

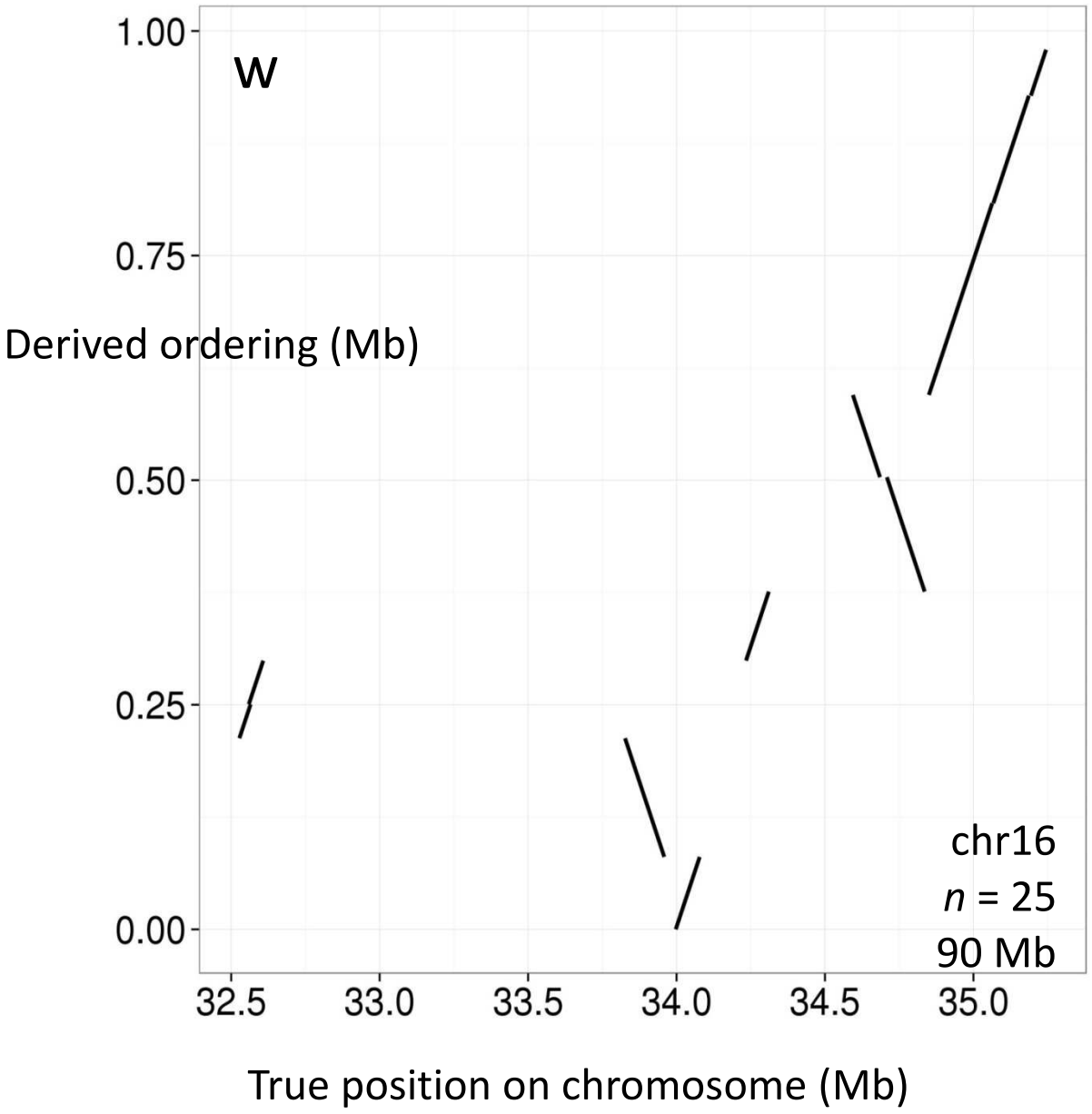


Figure A.3.4 (page 7 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3.3**. **w**, the chimeric group not shown in **Figure 3.3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

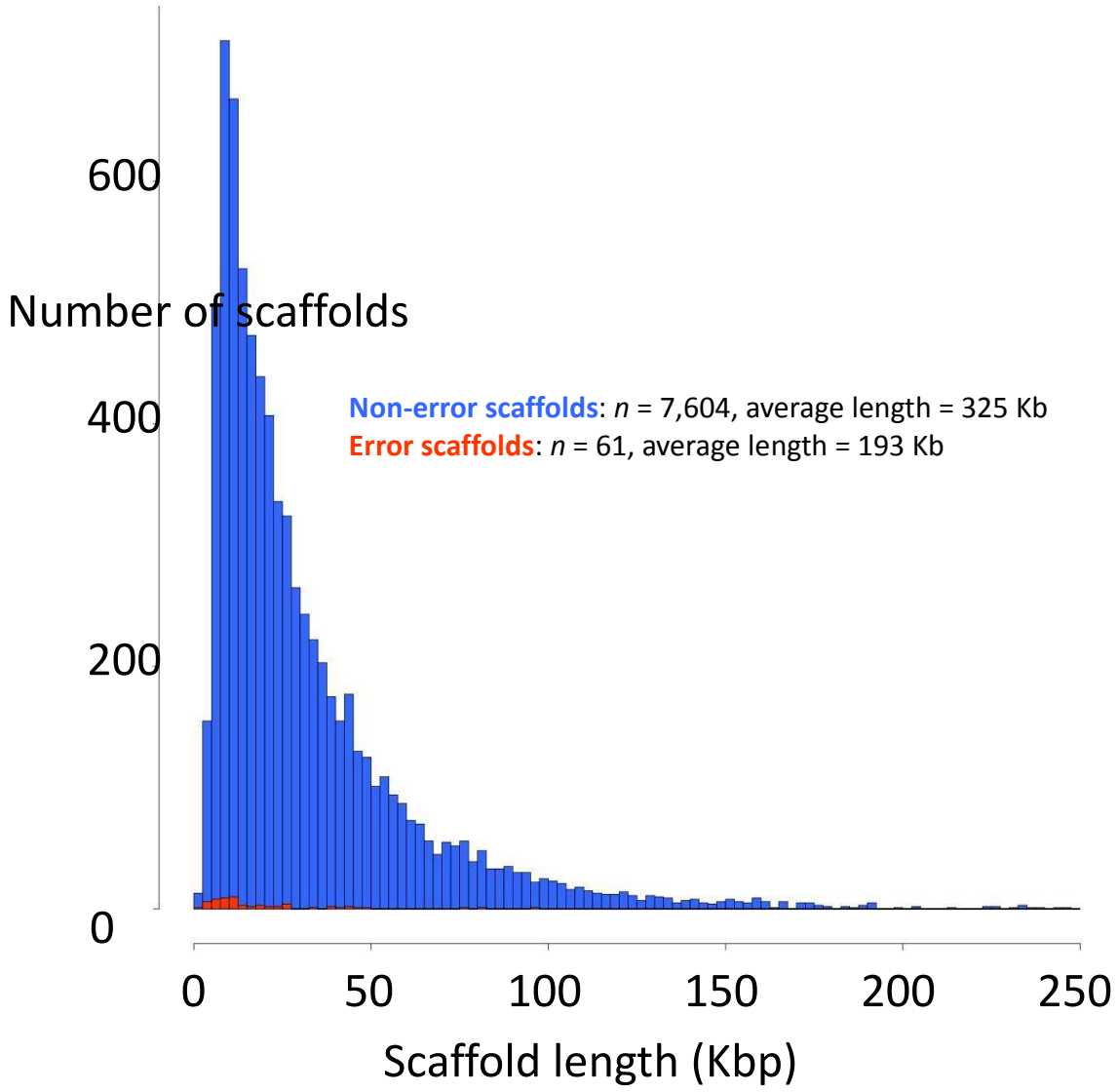


Figure A.3.5 | Scaffolds associated with ordering errors tend to be shorter than correctly ordered scaffolds. A histogram of the lengths of all scaffolds in the *de novo* human assembly which *LACHESIS* places in orderings and which map to the human reference. Scaffolds marked with ordering errors are shown in red; all other scaffolds are shown in blue. For clarity, six scaffolds of length >250 Kbp (none of which have ordering errors) are not shown.

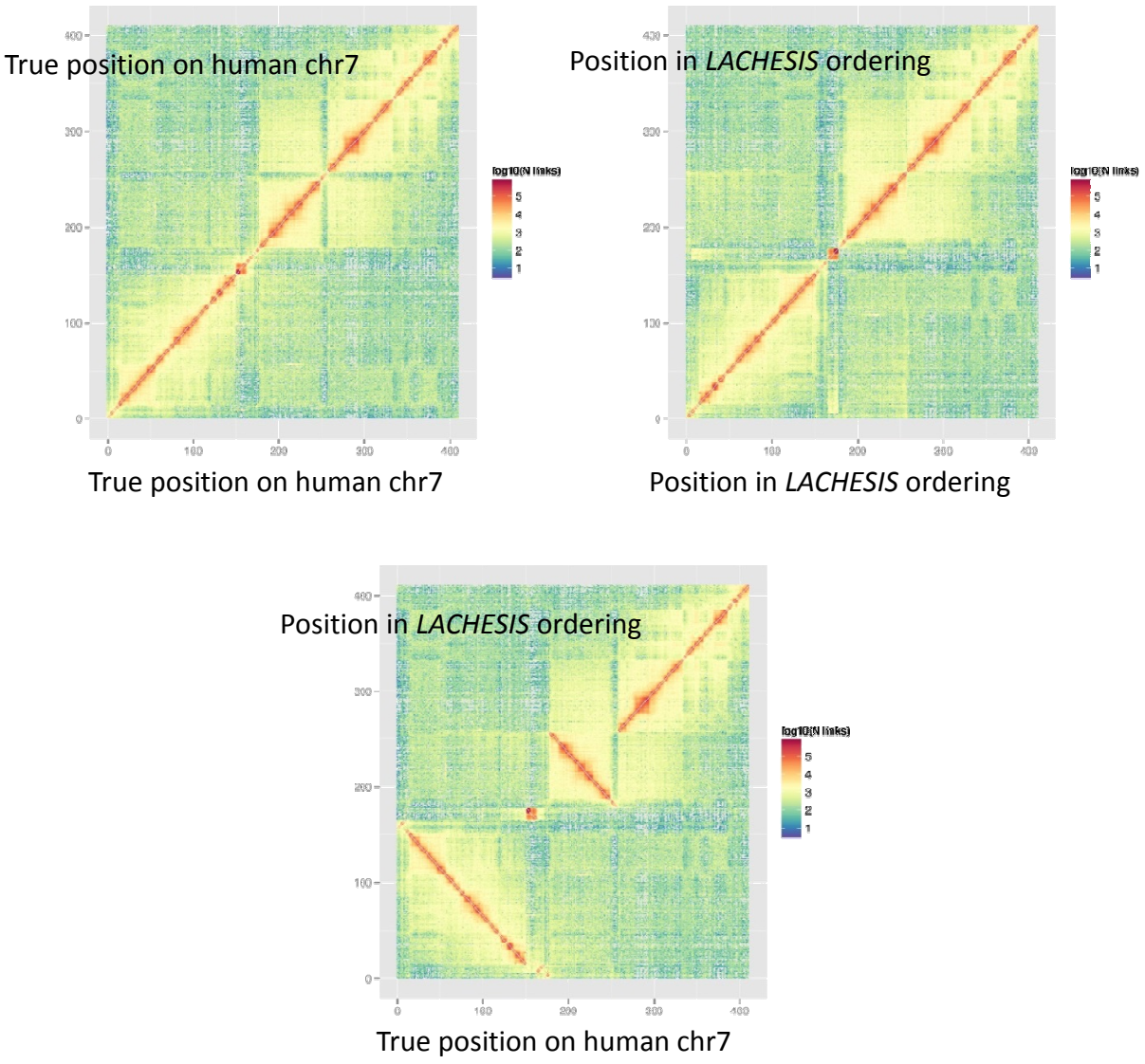


Figure A.3.6 | Example of *LACHESIS* assembly errors due to long-range chromatin interactions. Shown are three heatmaps of the density of Hi-C links between scaffolds of the *de novo* human assembly for chromosome 7. Only mapping contigs of length ≥10 Kb are shown. **a.** The scaffolds are ordered on both axes by their true position on chromosome 7. Note the presence of large domains with long-range internal interactions (squares along diagonal). **b.** The scaffolds are ordered on both axes by their position in the *LACHESIS* ordering in the group corresponding to chromosome 7. **c.** The scaffolds are ordered on the *x*-axis by their true position, and on the *y*-axis by their position in the *LACHESIS* ordering, revealing incorrect fusions of domains. Compare to **Figure A.3.4g**.

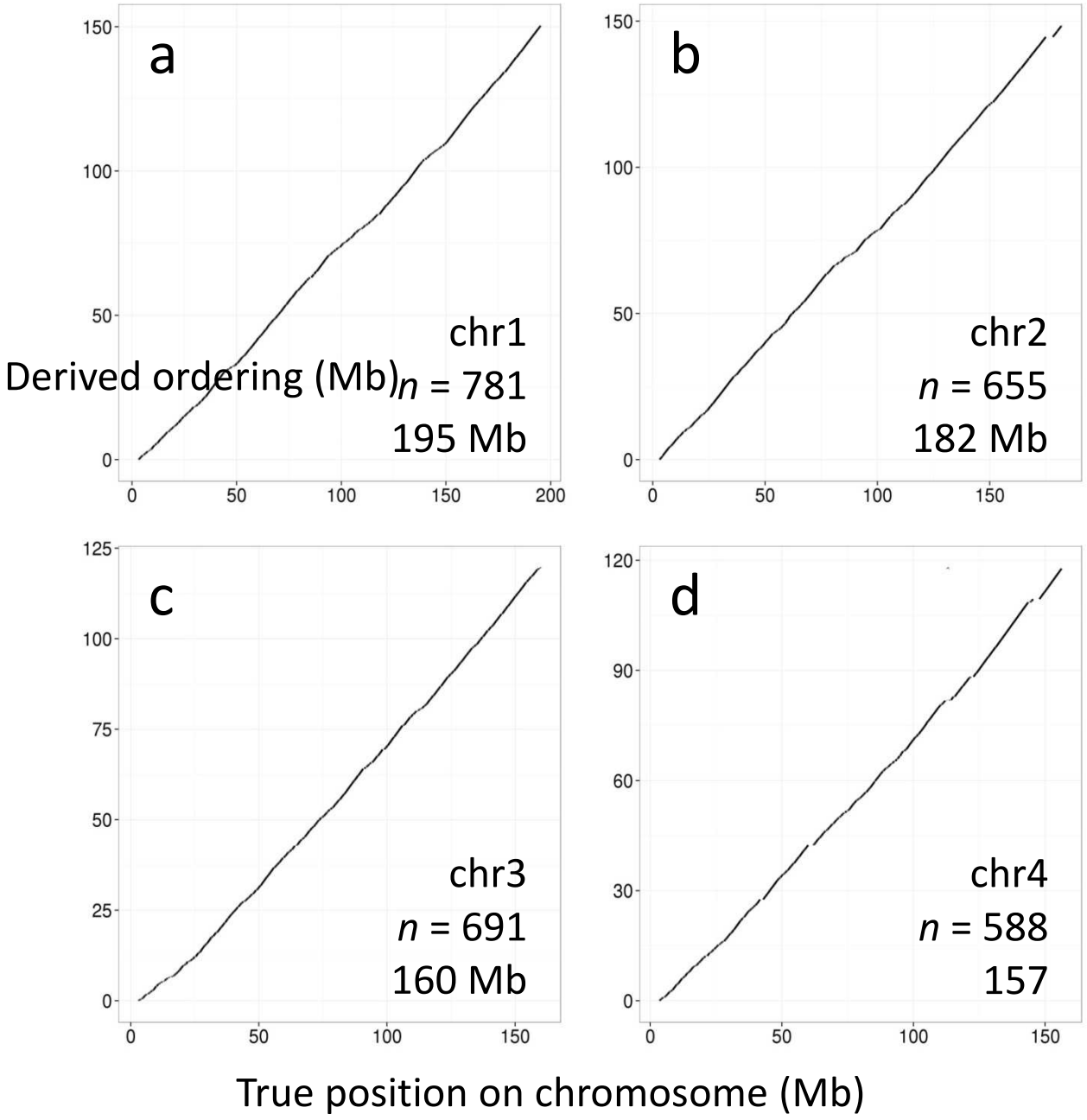


Figure A.3.7 (page 1 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.

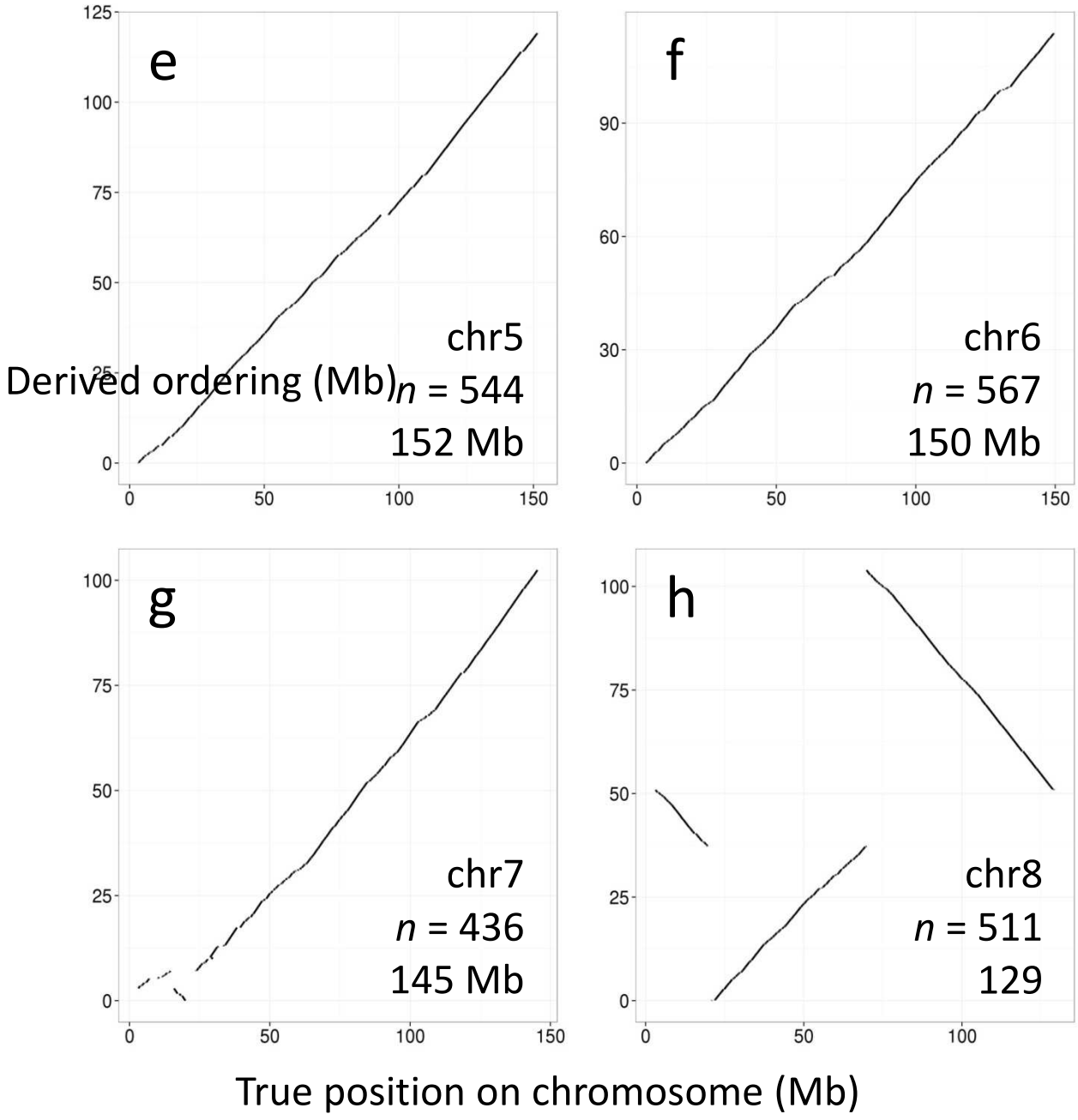


Figure A.3.7 (page 2 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.

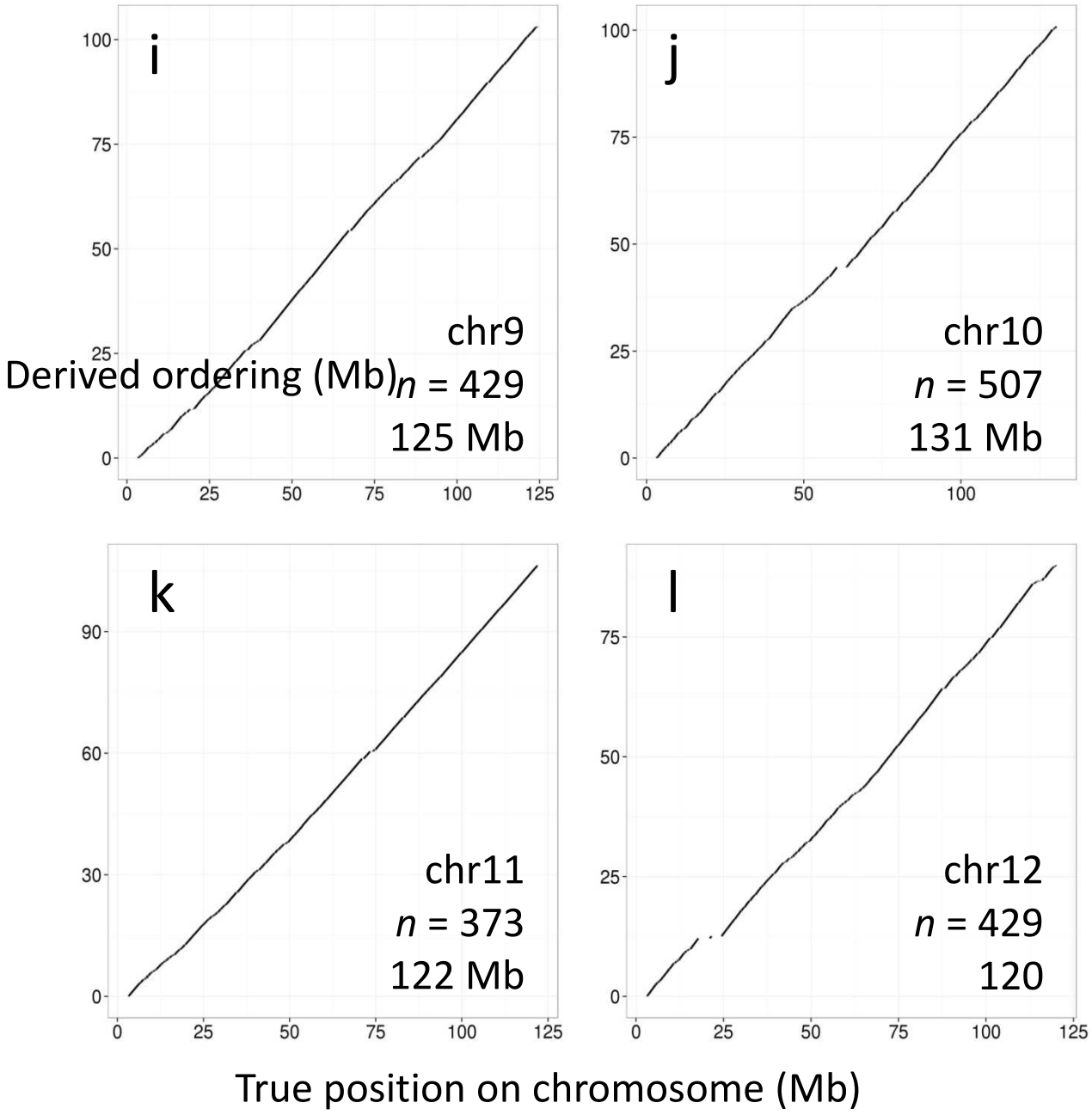


Figure A.3.7 (page 3 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.

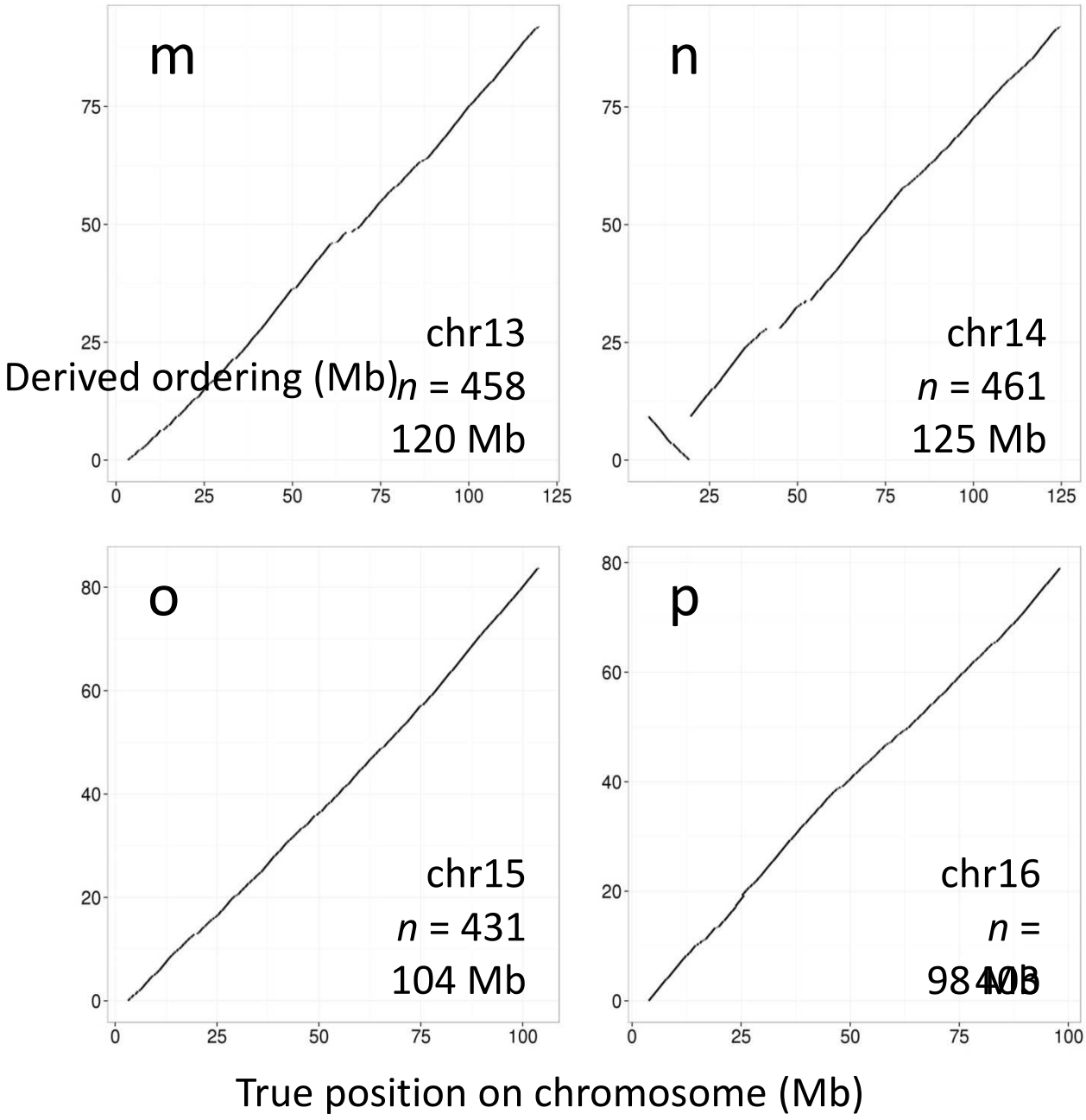


Figure A.3.7 (page 4 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.

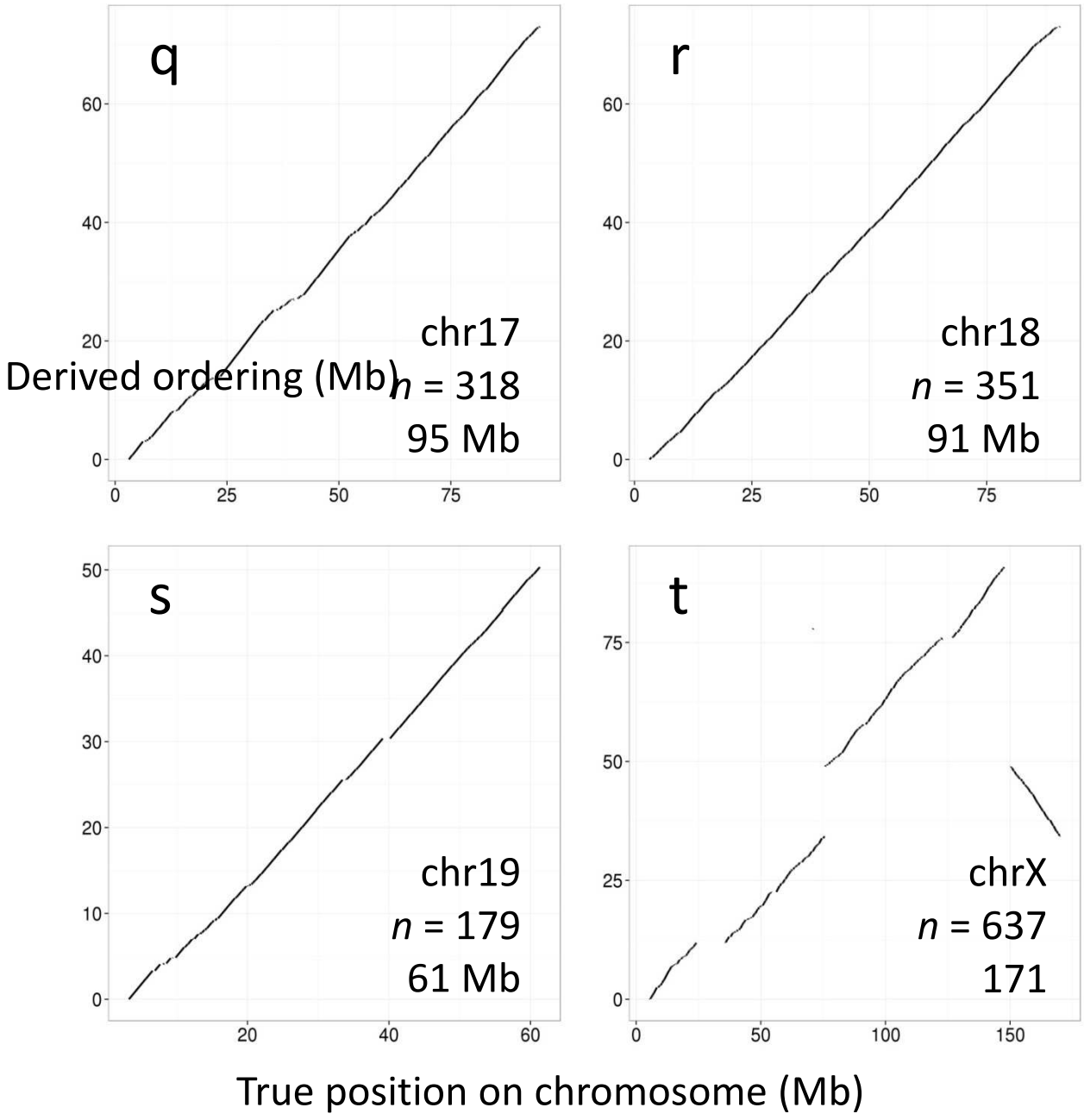


Figure A.3.7 (page 5 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.

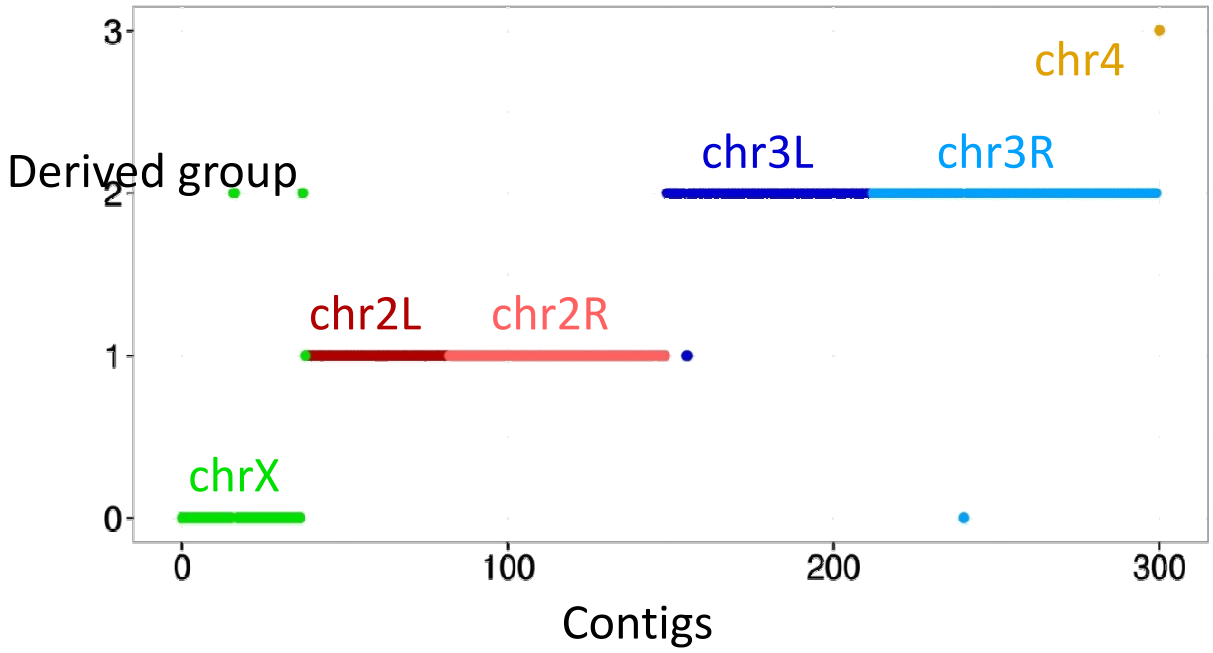


Figure A.3.8 | *LACHESIS* clustering results on the *Drosophila de novo* assembly. Shown on the *x*-axis are the 301 contigs (of 4,568 total contigs; total length: 43.8 Mb) that are long (≥250 GATC restriction sites) and not repetitive (Hi-C link density less than 2 times average), which *LACHESIS* used as informative for clustering. The *y*-axis shows the four groups created by *LACHESIS*, with the order chosen for the purposes of clarity. Each contig is shown as a dot, with a color indicating the chromosome to which the contig truly aligns, including the chromosome arm in the case of chromosomes 2 and 3.

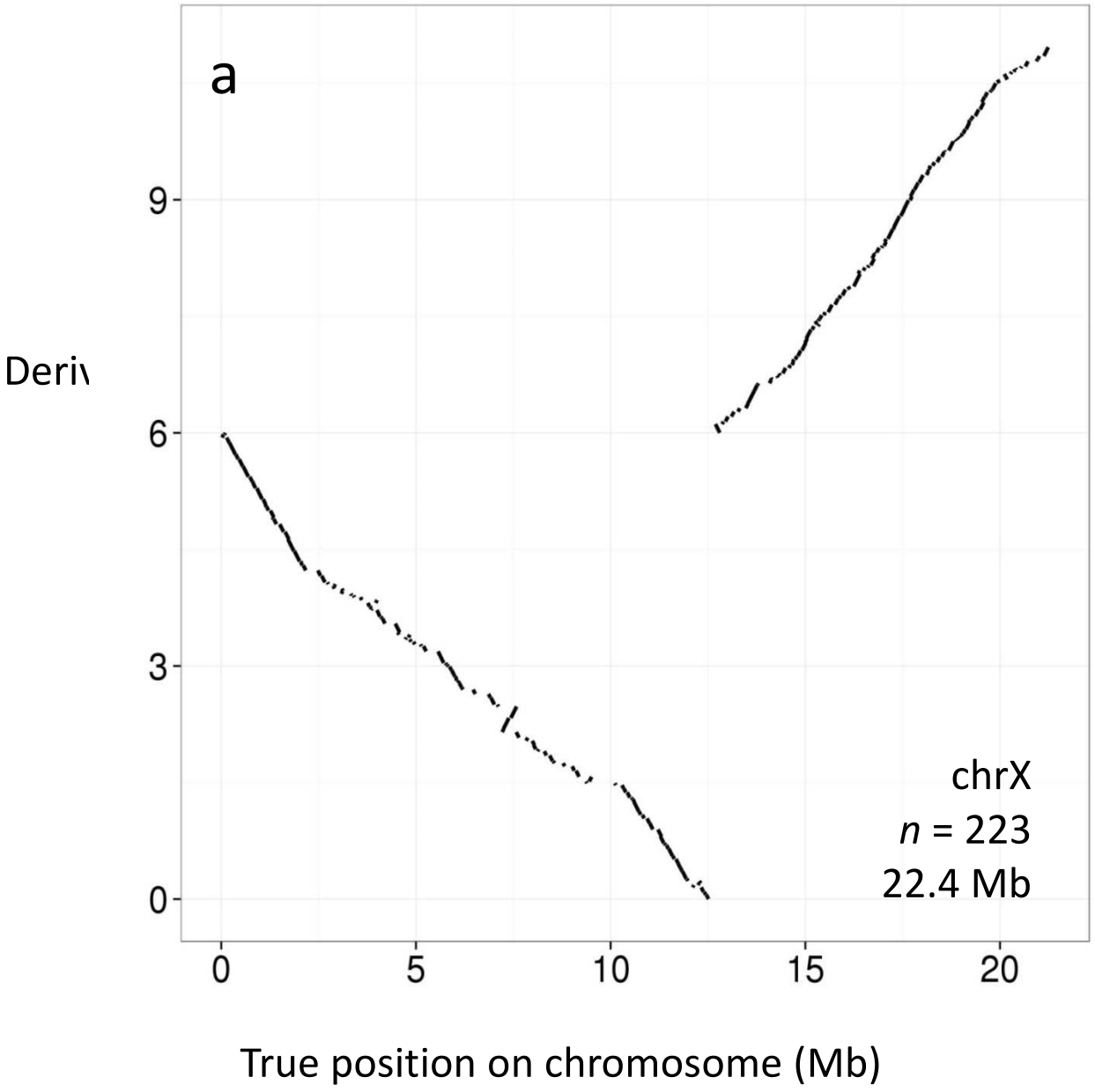


Figure A.3.9 (page 1 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

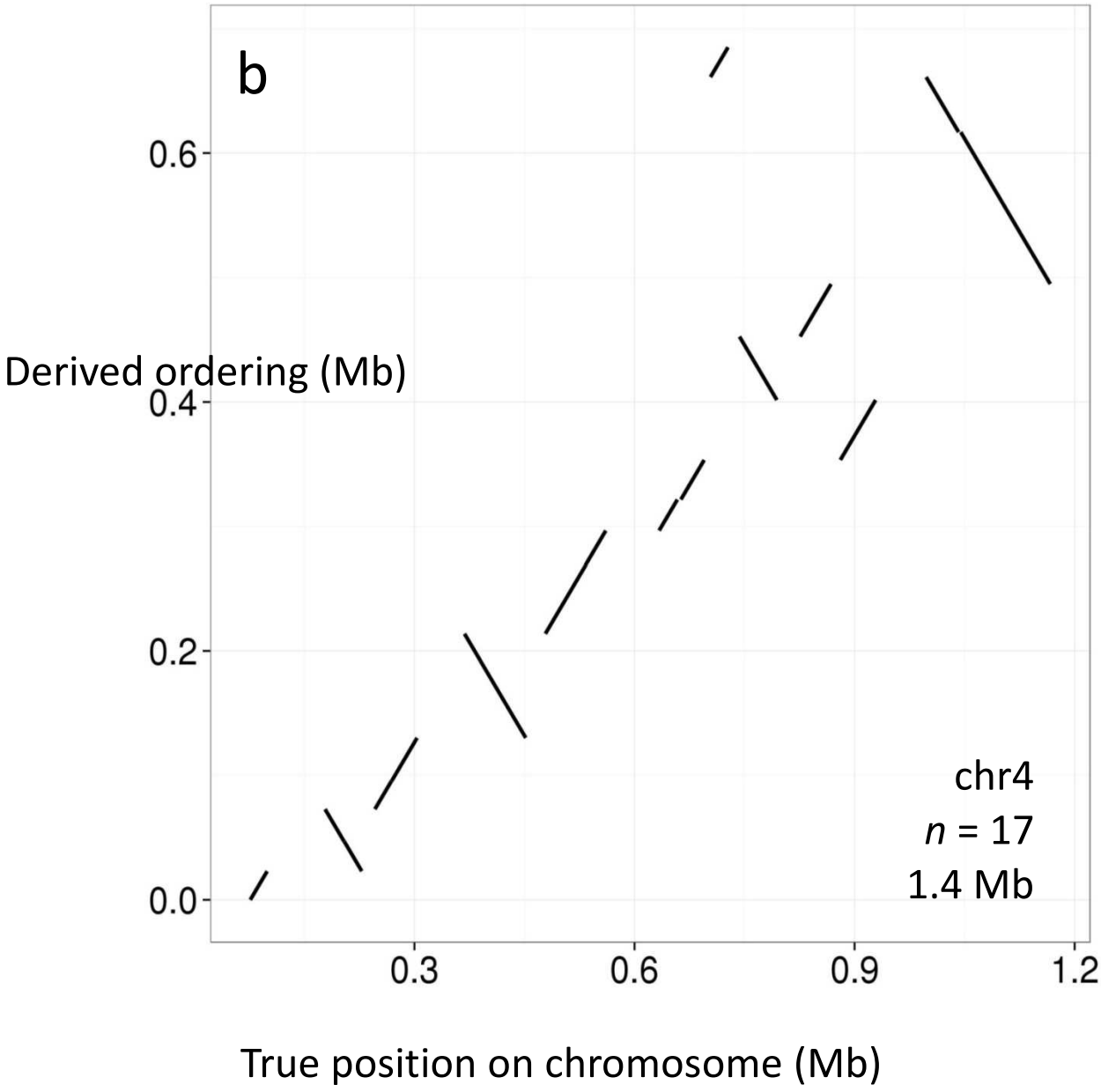


Figure A.3.9 (page 2 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

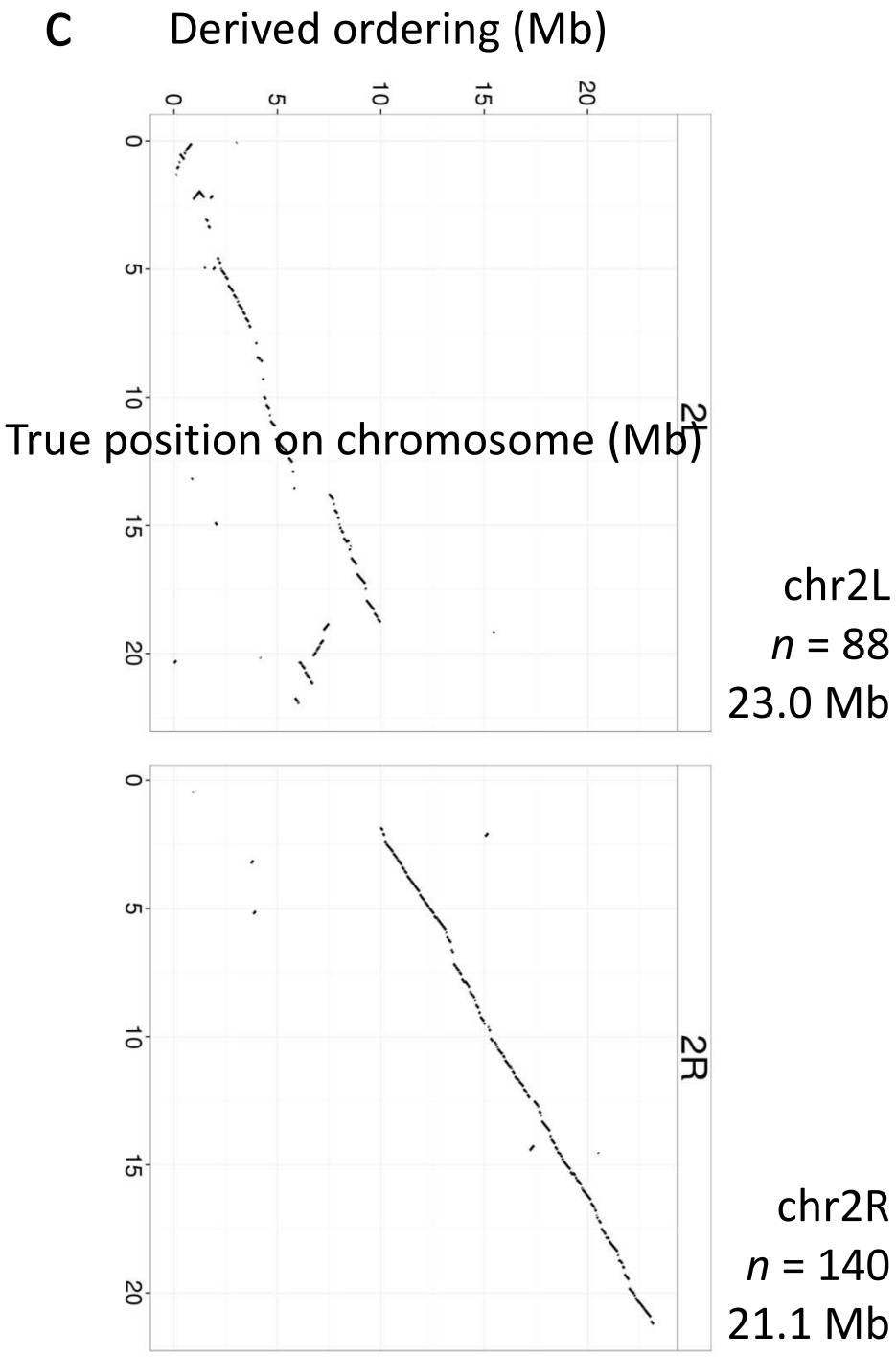


Figure A.3.9 (page 3 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

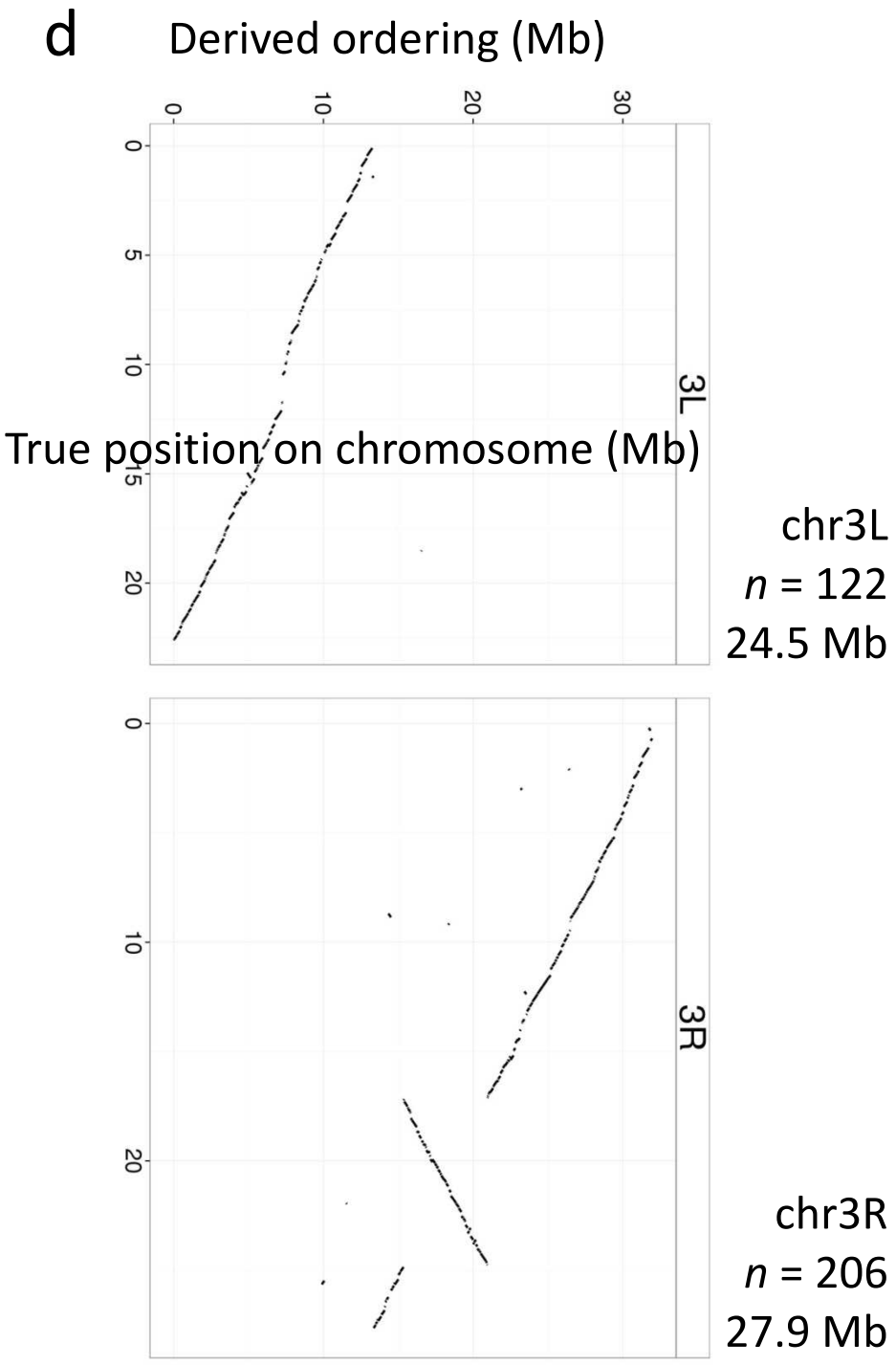


Figure A.3.9 (page 4 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Table A.2.5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

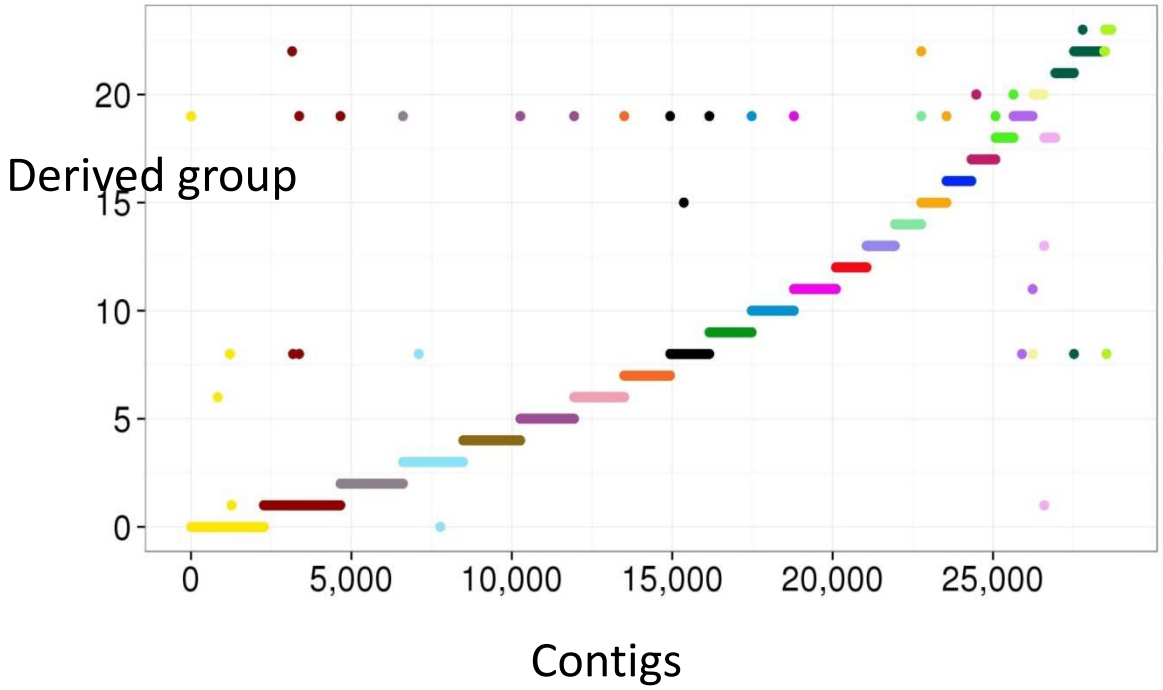


Figure A.3.10 | *LACHESIS* clustering results on simulated 100 Kb contigs of the human reference genome. The human genome was split into simulated 100 Kb contigs and *LACHESIS* was used to cluster these contigs into groups. The 28,689 clustered contigs (total length: 2.87 Gb) are ordered on the *x*-axis in order of ascending chromosome number and then position on the chromosome. The *y*-axis represents the 24 groups created by LACHESIS, with the order chosen for the purposes of clarity. Each 100 Kb contig is shown as a dot, with a color indicating the chromosome on which it belongs. The color scheme is the standard SKY (spectral karyotyping) color scheme for human. Not shown are the 2,281 contigs (7.4%) not placed into groups due to lack of unique sequence content, mostly corresponding to centromeres.

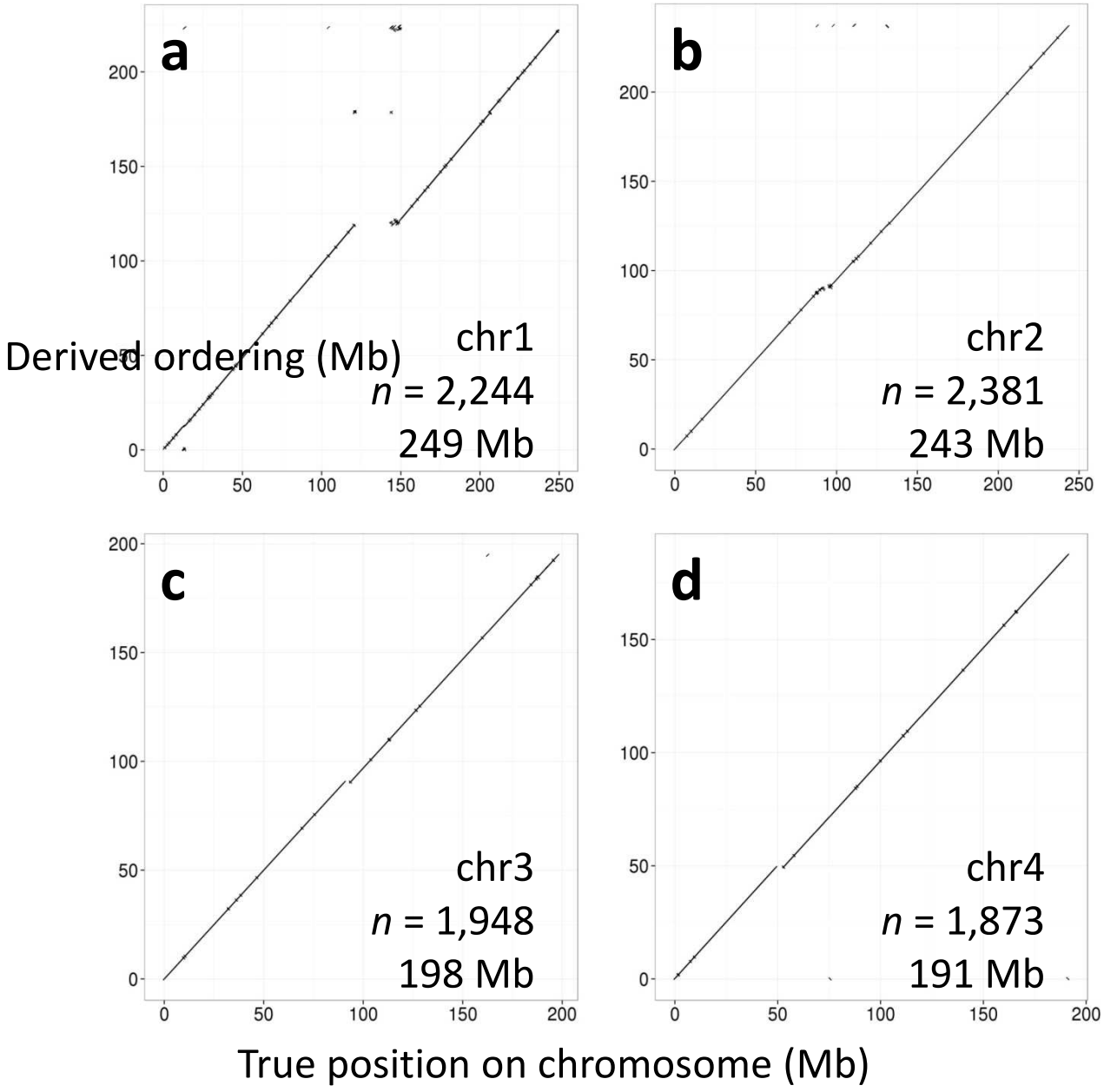


Figure A.3.11 (page 1 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

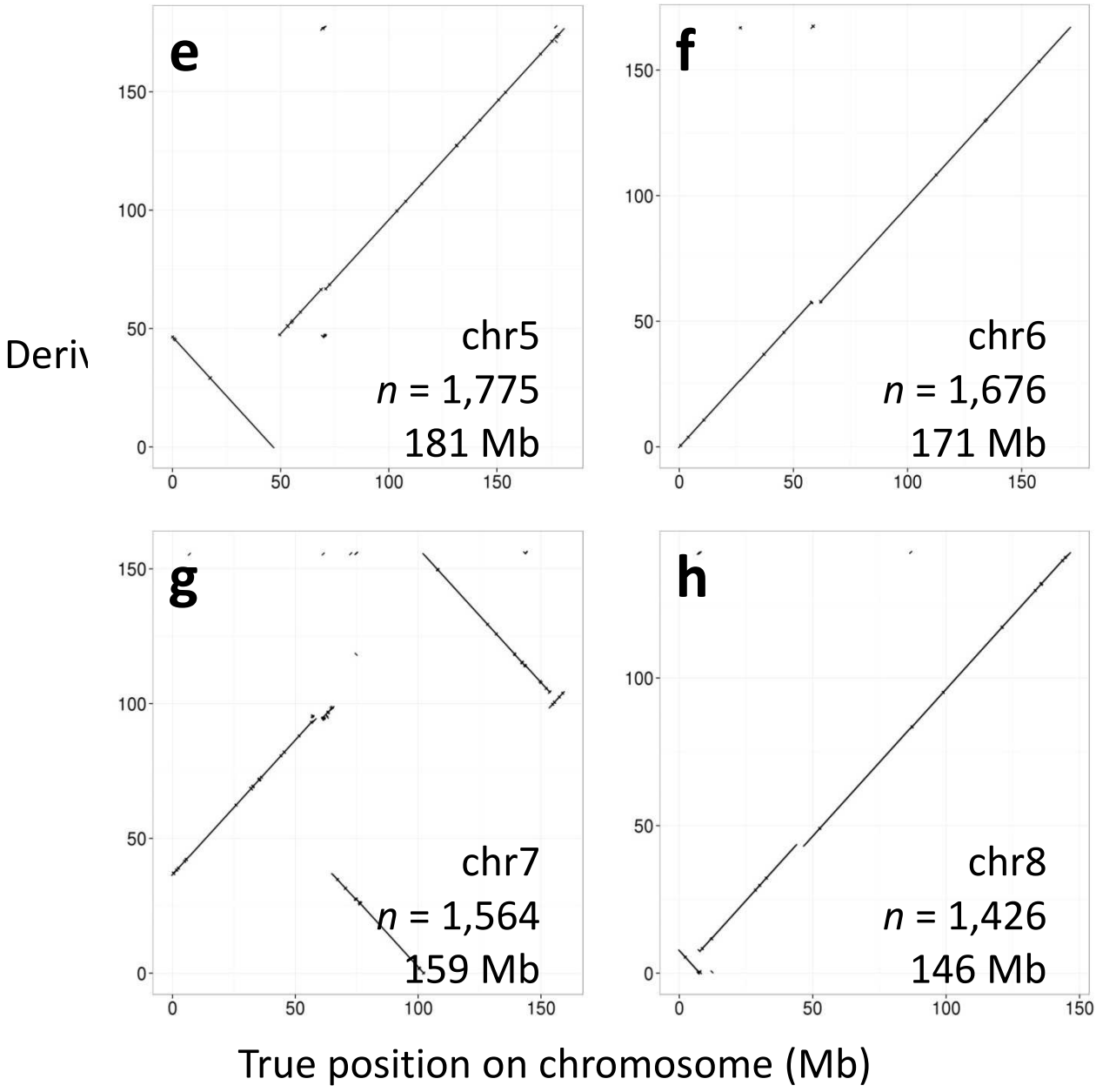


Figure A.3.11 (page 2 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

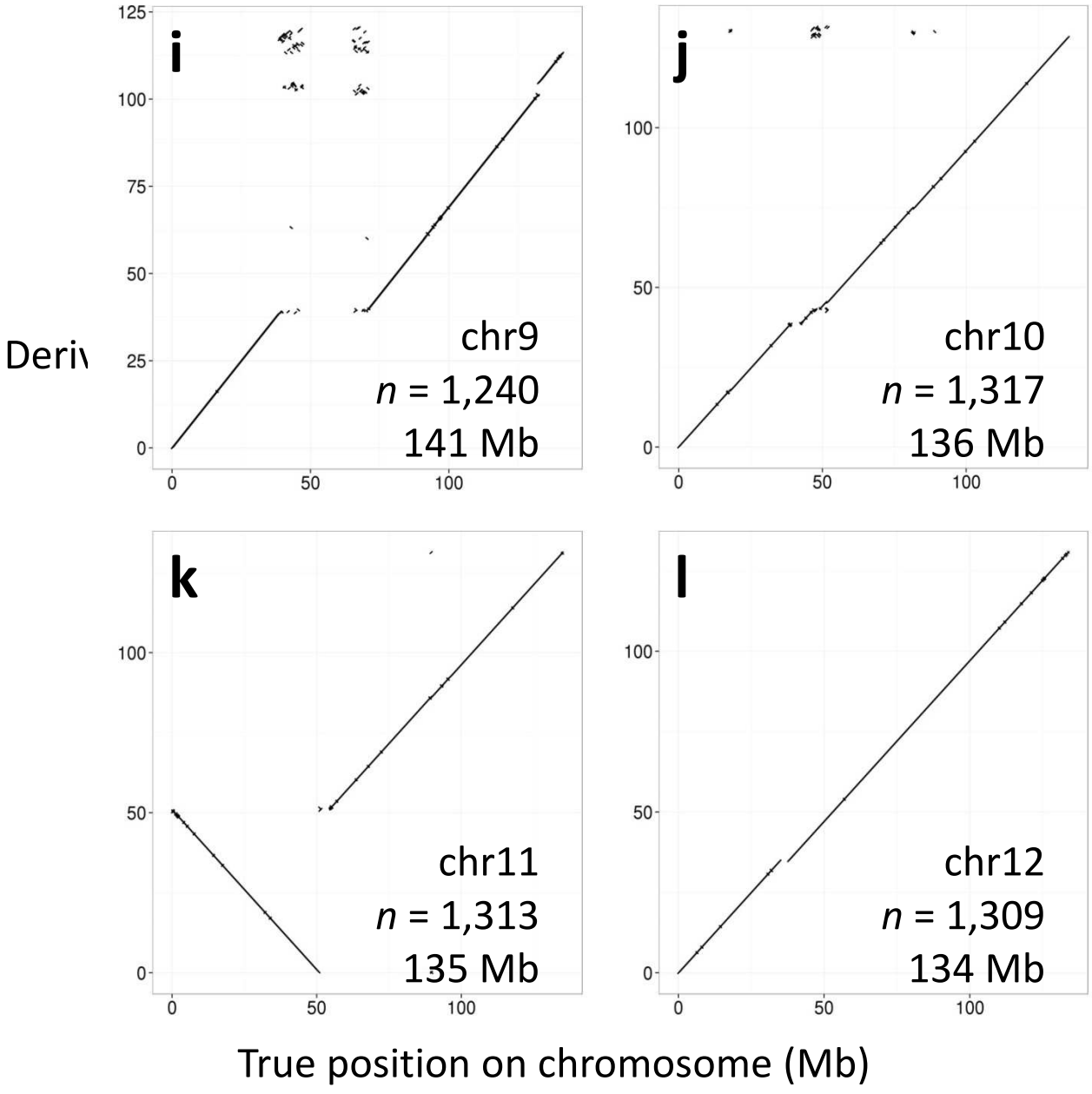


Figure A.3.11 (page 3 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

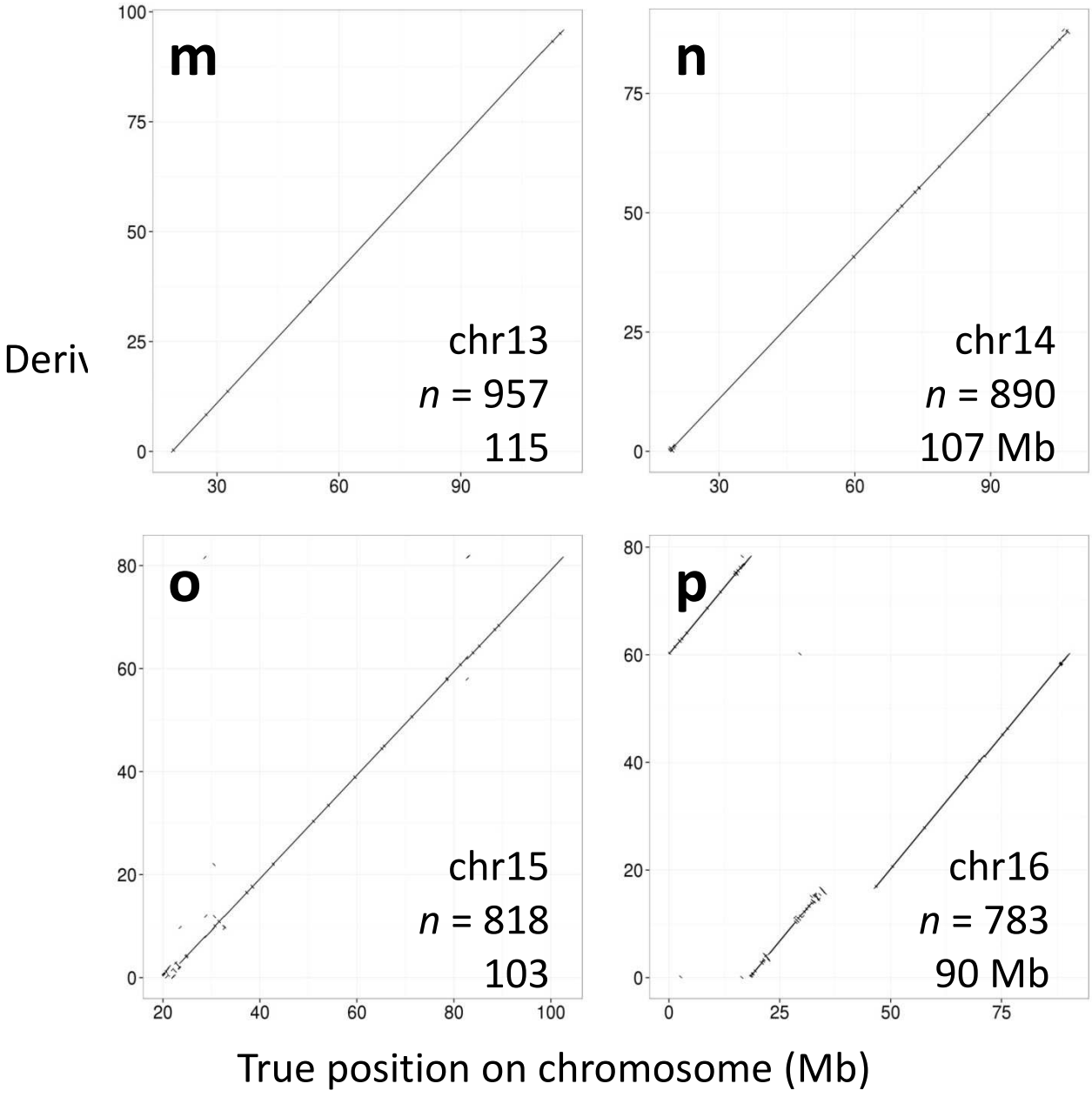


Figure A.3.11 (page 4 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

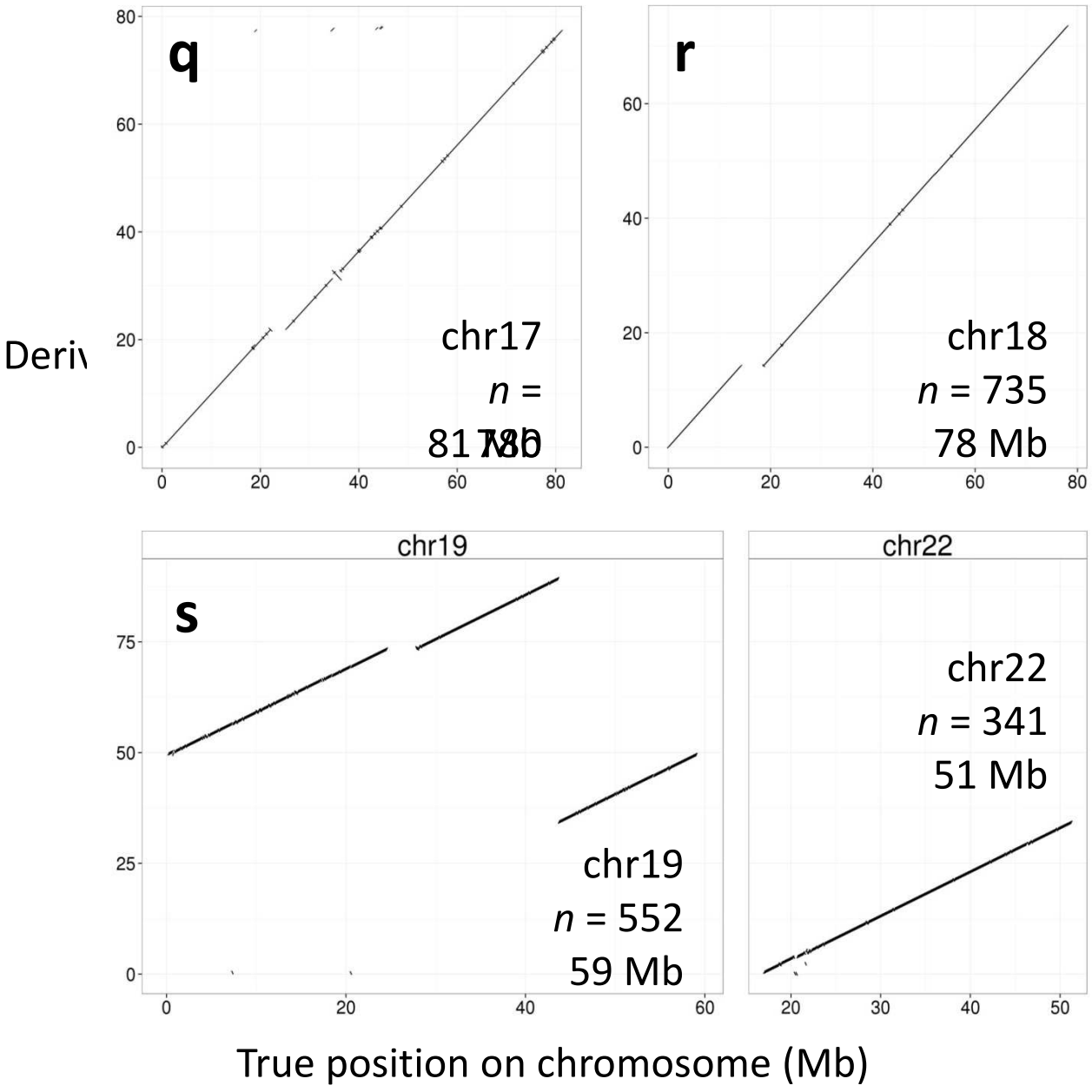


Figure A.3.11 (page 5 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

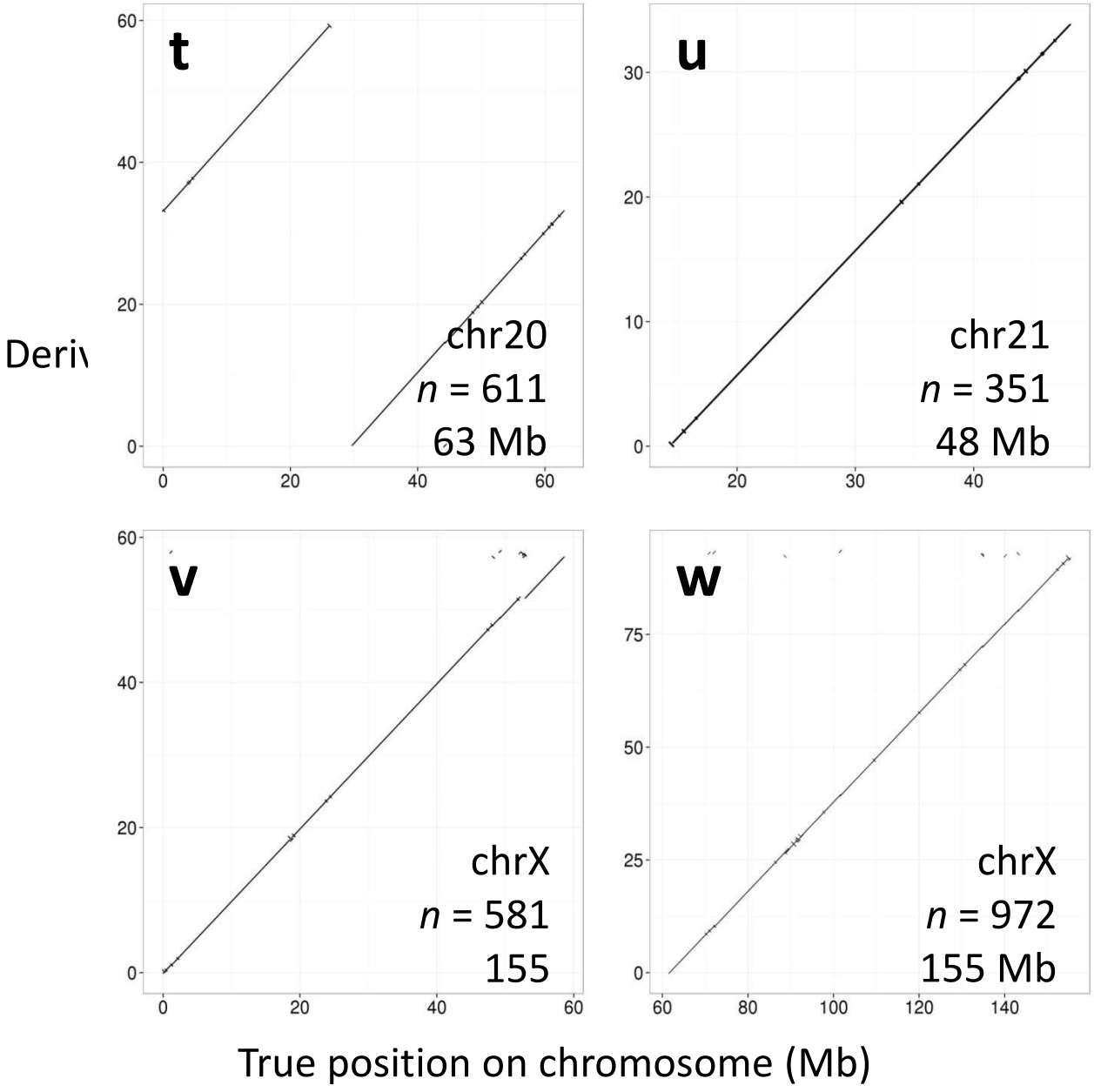


Figure A.3.11 (page 6 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.

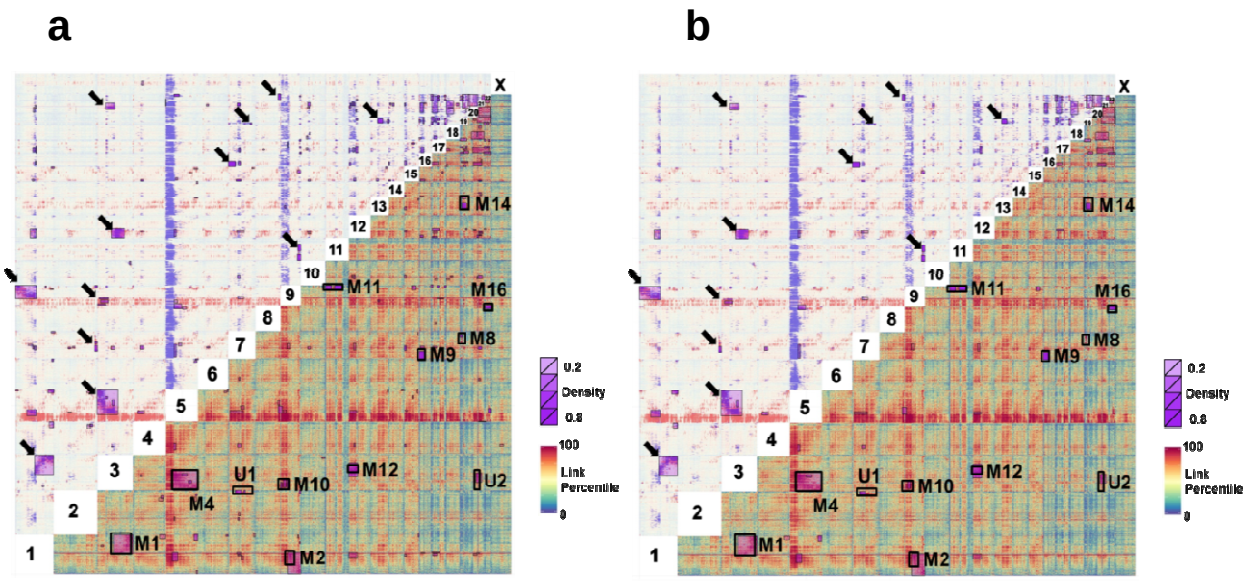


Figure A.3.12 | Using Hi-C to detect interchromosomal rearrangements in HeLa with high sensitivity. In the top left half of each image, blue horizontal lines represent outlying stretches of link scores with ≥10 windows, of which ≥80% of windows are ≥1 standard deviation above the mean of the row. Likewise, vertical red lines represent similar outlying stretches with respect to columns. Windows called as both row and column outliers are designated "outlier windows". Regions with excessive outlier stretch calls (*e.g.*, chromosome 5p) are only called as rows or columns and not likely both, thus reducing noise from globally high-scoring regions of the genome. Outlier window points are then clustered and called as potential fusions (purple boxes) and scored according to the density of outlier points within the window. **a.** An inclusive approach yields 100% sensitivity for detecting previously identified marker chromosomes, but only 8% specificity (assuming no additional marker chromosomes beyond those previously identified). False positive calls are largely due to increased interchromosomal contact among the smaller, gene-rich chromosomes, known to occur in healthy cells. **b.** Specificity can be increased by filtering based on cluster area. Specificity increases to 31% but sensitivity drops to 92%, with a bias towards rearrangements involving larger chromosomes or large regions of chromosomes.

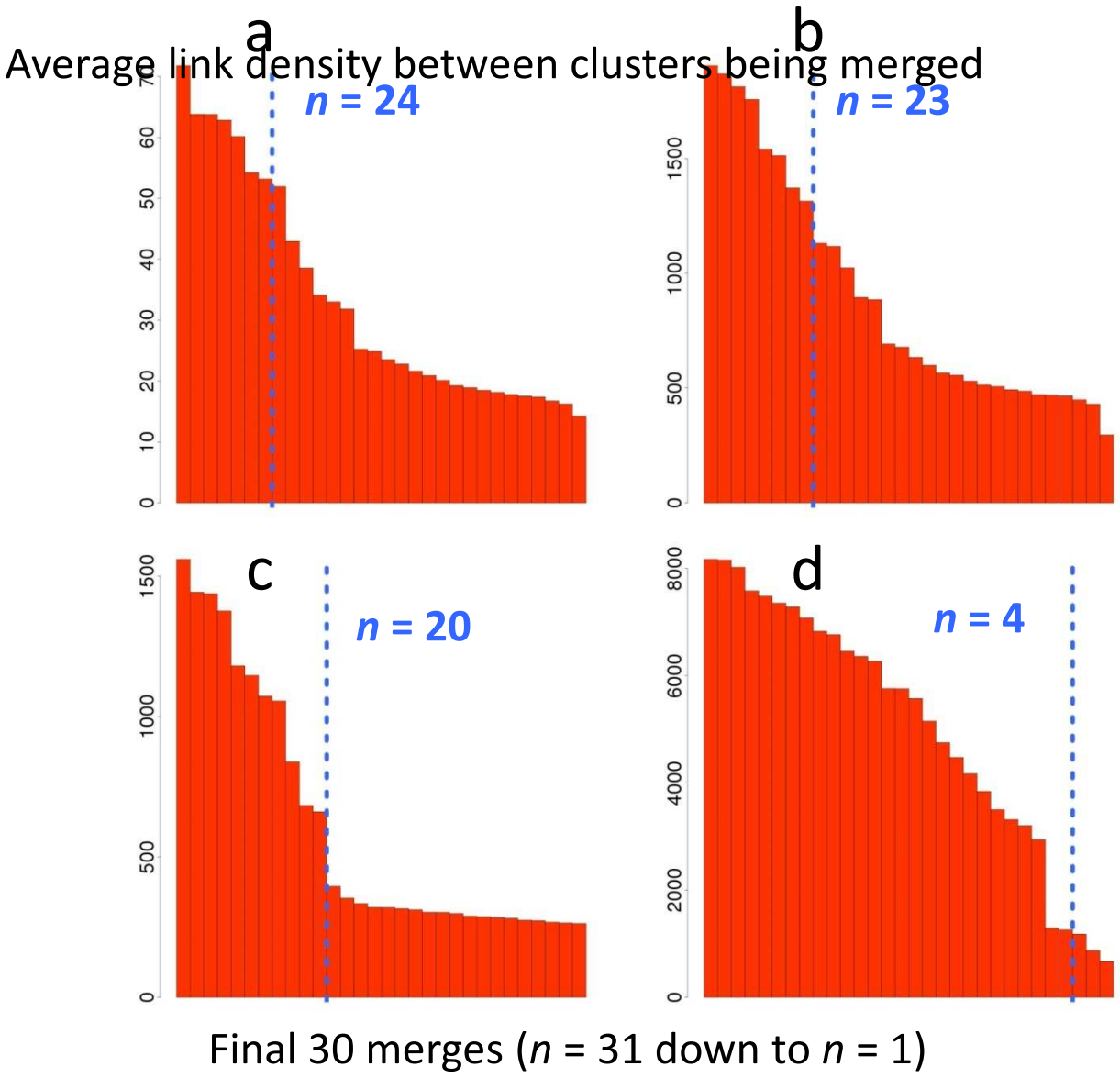


Figure A.3.13 | Difficulty of calling the number of chromosomes from Hi-C link data alone. At each step of the *LACHESIS* clustering algorithm, the two clusters with the highest average link density are found and merged (**Figure A.3.1**). *x*-axis, the final 30 merges; *y*-axis, average link densities of each merge. The average link density decreases monotonically as merges are made. In practice, merging stops when a predetermined number of clusters is reached (blue line); importantly, this number is determined from *a priori* knowledge of the chromosome number rather than from the link densities shown here. **a.** The human simulated assembly with 100 Kb bins. **b.** The human *de novo* assembly. Note that the first several merges beyond $n = 23$, corresponding to fusions of the small chromosomes, have fairly high link densities. **c.** The mouse *de novo* assembly. **d.** The *Drosophila de novo* assembly. Note that the link densities imply $n = 6$, corresponding to a split of the arms of fly chromosomes 2 and 3, is a better solution than $n = 4$.

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

The material in this Appendix, like that of **Chapter 4**, is based on the following peer-reviewed publication[99]:

Joshua N. Burton, **Ivan Liachko**, Maitreya J. Dunham and Jay Shendure. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3: Genes | Genomes | Genetics* **4**(7), 1339-1346 (2014).

Boldface indicates authors who contributed equally to this work.

Supplementary methods are in section B.1. Supplementary tables are in section B.2. Supplementary figures are in section B.3.

B.1 Supplementary methods for Chapter 4

Sample collection. Cultures of individual strains listed in **Table 4.1**, Table C.2.1 (M-Y) and Table C.2.2 (M-3D) were grown to saturation in rich media (YPD for yeasts, LB for bacteria, McCas media for *M. maripaludis*, PMsul media for *R. palustris*). Culture densities were measured by spectrophotometry and FACS. After mixing the strains, cultures were diluted with YPD media (M-Y) or with LB media (M-3D) to a final OD600 of 1.0 in a final volume of 500 mL. Formaldehyde was added to a final concentration of 1% and cultures were incubated at room temperature for 30 min. To quench the crosslinker, 5 g of glycine was added to each 500 mL of culture, and the cultures were incubated for 30 min at room temperature. Cultures were centrifuged to pellet all cells. Cell pellets were frozen at −20° until further processi ng.

Shotgun and mate-pair libraries. Total DNA was isolated from cultures using a standard phenol/chloroform glass bead purification followed by ethanol precipitation and subsequent cleanup using the DNA Clean and Concentrator-5 Kit (Zymo Research). Shotgun libraries were prepared using the Nextera DNA Sample Preparation Kit (Illumina). Mate-pair libraries were constructed using the Nextera Mate Pair Sample Preparation Kit (Illumina).

Hi-C libraries. Cell pellets (~100 µL volume each) were resuspended in 2 mL of 1× TBS buffer containing 1% Triton-X and Protease Inhibitors (cOmplete, EDTA-free; Roche) and split equally into two separate 2-mL tubes; 300–500 µL of 0.5-mm diameter glass beads were added to each tube and tubes were vortexed on the highest setting in four 5-min increments, each separated by 2-min incubations on ice. Lysate was transferred to fresh tubes. Crosslinked chromatin was recovered by centrifugation at 13 KRPM in an accuSpin Micro17 centrifuge (Fisher) and rinsed with 1× TBS buffer. Chromatin from each tube was digested overnight with 100 units of either *Hin*dIII or *Nco*I restriction endonuclease (NEB) at 37° in a total volume of 200 µL. To enrich for long-range interactions (M-3D library only), digested chromatin was centrifuged for 10 min at 13 KRPM, rinsed in 200 µL of 1× NEBuffer 2 (NEB), centrifuged again, and resuspended in 200 µL of 1× NEBuffer 2 (NEB). Restriction fragment overhangs were filled in using biotinylated dCTP (Invitrogen) and Klenow (NEB) as described[201]. DNA concentration within the chromatin suspension was quantitated using the QuBit fluorometer (Invitrogen), and for each sample an

8-mL ligation reaction was set-up at a final DNA concentration of 0.5 ng/μL using T4 DNA Ligase (NEB). Ligation reactions were incubated at room temperature for 4 hr and then overnight at 70° to reverse crosslinks. DNA was purified using a standard phenol/chloroform purification followed by ethanol precipitation and resuspended in 600 μL of water with 1× NEBuffer 2 (NEB) and 1× BSA (NEB). To remove biotin from unligated DNA ends, 20 units of T4 Polymerase (NEB) were added to each 600 μL DNA sample and incubated at 25° for 10 min followed by 12° for 1 hr. DNA was purified using the DNA Clean and Concentrator-5 Kit (Zymo Research). Illumina libraries were constructed as described[201] using reagents from the Illumina Mate Pair Sample Preparation Kit. Paired-end sequencing was performed using the HiSeq and MiSeq Illumina platforms (**Table 4.2**).

Draft metagenome assembly for M-Y and M-3D. To create draft metagenome assemblies for the synthetic samples, we assembled the fragment library using the IDBA-UD assembler[86]. We ran IDBA-UD with the –read option set to the fragment reads and the following additional parameters: '–pre_correction –mink 20 –maxk 60 –step 10'. We used the assembly in contig.fa rather than scaffold.fa to reduce the risk of false joins made at the scaffolding stage.

Aligning Hi-C reads. We aligned the Hi-C reads to the draft metagenome assembly in a multi-step process. First, the reads were aligned using BWA[199] with the option '-n 0', requiring a perfect match of the entire 100-bp read. For read pairs in which an alignment was not found for both reads, the reads were trimmed from 100 bp to 75 bp and were aligned using '-n 0' again. For read pairs in which alignment was still not found for both reads, the reads were trimmed to 50 bp and aligned using '-n 0' again. All read pairs for which no alignment was found were discarded from further analysis. Read pairs were also discarded if the reads did not both align within 500 bp of a restriction site, as recommended by Yaffe and Tanay[116].

Clustering contigs by species. To cluster the contigs of the draft metagenome assembly into individual species, we used a hybrid clustering algorithm. A graph was built, with each node representing one contig and each edge between nodes having a weight equal to the number of Hi-C read pairs linking the two contigs, normalized by the number of restriction sites on the contigs. Only the single largest component in the graph was used; the other components, generally comprising isolated contigs containing a small

fraction of the total sequence length, were discarded and the contigs were not clustered. Within this component, the Jarvis-Patrick nearest-neighbor clustering algorithm[144] was applied with k = 100, removing some edges and reweighting all other edge weights by the frequency of their nodes' shared nearest neighbors. This nearest-neighbor approach accounts for the likely possibility that the clusters representing each species will have different internal densities of Hi-C links due to species' differing abundances in the sample or differing susceptibility to the cell lysis step of Hi-C. Finally, the nodes were merged together using hierarchical agglomerative clustering with an average-linkage metric[114], which was applied until the number of clusters was reduced to the expected or predicted number of individual species (12 for M-Y, not including *P. pastoris*; 18 for M-3D).

Scaffolding of genomic content within individual clusters. To scaffold the individual species' genomes represented in each cluster of contigs, we aligned the Hi-C reads to these contigs and ran them through our Lachesis software[97] to create chromosome-scale scaffolds. The number of chromosomes in each species (7 for *K. wickerhamii*[202]; 8 for *S. stipitis*[203]) was provided as an input to Lachesis.

Validation**.** To determine the true species identity of the contigs in the draft metagenome assembly, we aligned them to a combined reference genome that included the reference genomes of all strains known to be in the metagenome sample (16 strains for M-Y; 18 species for M-3D). The alignment was performed by BLASTn[200] with the following stringent parameters: '-perc_identity 95 -evalue 1e-30 -word_size 50'. A contig was defined as aligning to a species if any alignment of the contig to the species' reference genome was found; the placement of the alignment was ignored.

B.2 Supplementary tables for Chapter 4

Genus	Species	In sample			Reference				
		Strain	Source	Ploidy	Optical density	Strain	Size (Mb)	Download source	Finished?
Saccharomyces	*cerevisiae*	FY4H	M. Dunham	1	0.079088	FY	12.2	downloads.yeastgenome.org	Yes
Saccharomyces	*cerevisiae*	CEN.PK	P. Kotter	1	0.071645	CEN.PK	11.5	downloads.yeastgenome.org	No
Saccharomyces	*cerevisiae*	RM11-1A	L. Kruglyak	1	0.084903	RM11-1A	11.7	www.broadinstitute.org	Yes
Saccharomyces	*cerevisiae*	SK1	A. Deutschbauer	2	0.075366	SK1	11.9	cbio.mskcc.org/public/SK1_MvO/	Yes
Saccharomyces	*paradoxus*	YDG613	D. Greig	2	0.076762		11.7	saccharomycessensustricto.org	Yes
Saccharomyces	*mikatae*	FM356	M. Johnston	2	0.08188	IFO 1815	11.5	saccharomycessensustricto.org	Yes
Saccharomyces	*kudriavzevii*	FM527	M. Johnston	2	0.008141	IFO 1802	11.3	saccharomycessensustricto.org	Yes
Saccharomyces	*bayanus* var. *uvarum*	YZB5-113	Y. Zhang	1	0.055827	CBS 7001	11.5	saccharomycessensustricto.org	Yes
Naumovozyma (*Saccharomyces*)	*castellii*	4310	D. Bartel	1	0.082577	NRRL Y-12630	11.2	downloads.yeastgenome.org	No
Lachancea	*waltii*	Kwaltii ura3	B. Brewer	1	0.086067	NRRL Y-8285	10.2	fangman-brewer-gbrowse.gs.washington.edu	Mostly
Lachancea (*Saccharomyces*)	*kluyveri*	FM628	M. Johnston	1	0.096534	CBS 3082	11.3	genolevures.org/sakl.html	Yes
Kluyveromyces	*lactis*	MW98-8C	C. Newlon	1	0.055827	NRRL Y-1140	10.7	genolevures.org/klla.html	Yes
Kluyveromyces	*wickerhamii*	Y-8286	USDA/ARS	1	0.062805	UCD 54-210	9.81	www.ncbi.nlm.nih.gov	No
Ashbya (*Eremothecium*)	*gossypii*	WT	S. Jaspersen	1	Can't measure	ATCC 10895	8.74	genolevures.org/ergo.html	Yes
Scheffersomyces (*Pichia*)	*stipitis*	Y-11545	USDA/ARS	1	0.080251	CBS 6054	15.4	www.ncbi.nlm.nih.gov	Yes
Pichia (*Komagataella*)	*pastoris*	JC308	J. Cregg	1	0.002326	GS115	9.21	www.ncbi.nlm.nih.gov	Yes

Table B.2.1: M-Y species list and abundances in sample.

Domain	Genus	Species	In sample		Reference				
			Strain	Source	Optical density	Strain	Size (Mb)	Download source	Finished?
Eukaryota	*Saccharomyces*	*cerevisiae*	FY4H	M. Dunham	1.02	FY	12.2	downloads.yeastgenome.org	Yes
Eukaryota	*Zygosaccharomyces*	*rouxii*	Y-229	USDA/ARS	0.66	CBS 732	9.76	genolevures.org/zyro.html	Yes
Eukaryota	*Lachancea* (*Kluyveromyces*)	*thermotolerans*	Y-8284	USDA/ARS	0.98	CBS 6340	9.39	genolevures.org/klth.html	Yes
Eukaryota	*Kluyveromyces*	*aestuarii*	YB-4510	USDA/ARS	1.04	ATCC 18862	9.91	www.ncbi.nlm.nih.gov	No
Eukaryota	*Hansenula* (*Ogataea*)	*polymorpha*	Y-5445	USDA/ARS	1.21	DL-1	8.86	www.ncbi.nlm.nih.gov	Yes
Eukaryota	*Pichia* (*Komagataella*)	*pastoris*	JC 308	J. Cregg	0.61	GS115	9.22	www.ncbi.nlm.nih.gov	Yes
Eukaryota	*Schizosaccharomyces*	*pombe*	YFS 103	N. Rhind	0.52	ASM294	12.6	www.pombase.org	Yes
Eukaryota	*Schizosaccharomyces*	*japonicus*	YFS 760	N. Rhind	0.19	yFS275	11.7	www.broadinstitute.org	No
Archaea	*Methanococcus*	*maripaludis*	S2	J. Leigh	~0.1	S2	1.67	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Escherichia*	*coli*	AG 111	H. Merrikh	0.26	K-12	4.69	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Vibrio* (*Aliivibrio*)	*fischeri*	ES114	P. Greenberg	0.25	ES114	4.27	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Pseudomonas*	*fluorescens*	Pf-5	C. Harwood	0.4	Pf0-1	6.44	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Acinetobacter*	*baylyi*	ADP1	C. Harwood	0.12	ADP1	3.60	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Burkholderia*	*thailandensis*	E264	C. Harwood	0.6	E264	6.72	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Agrobacterium*	*tumefaciens*	P4	C. Queitsch	0.37	P4	6.33	www.ncbi.nlm.nih.gov	Mostly
Bacteria	*Rhodopseudomonas*	*palustris*	CGA009	C. Harwood	0.32	CGA 009	5.47	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Flavobacterium*	*johnsoniae*	UW 101	C. Harwood	0.55	UW 101	6.10	www.ncbi.nlm.nih.gov	Yes
Bacteria	*Bacillus*	*subtilis*	HM1/168	H. Merrikh	0.35	168	4.22	www.ncbi.nlm.nih.gov	Yes

Table B.2.2: M-3D species list and abundances in sample.

M-Y (total sequence length = 135206617)				
	Sequence clustered	% clustered	Seq misclustered	% misclustered
Main result	111112059	82.18%	922932	0.83%
Bootstrap 1	111136126	82.20%	4146798	3.73%
Bootstrap 2	111102736	82.17%	4500167	4.05%
Bootstrap 3	111106339	82.18%	4655083	4.19%
Bootstrap 4	111101816	82.17%	4389061	3.95%
Bootstrap 5	111106542	82.18%	4425448	3.98%
Bootstrap 6	111158559	82.21%	4356095	3.92%
Bootstrap 7	111089140	82.16%	4173561	3.76%
Bootstrap 8	110777343	81.93%	1294345	1.17%
M-PE (total sequence length = 133169811)				
	Sequence clustered	% clustered	Seq misclustered	% misclustered
Main result	118841530	89.24%	461626	0.39%
Bootstrap 1	117677687	88.37%	1748183	1.49%
Bootstrap 2	117818421	88.47%	737953	0.63%
Bootstrap 3	117636660	88.34%	1834184	1.56%
Bootstrap 4	117604654	88.31%	497732	0.42%
Bootstrap 5	117695244	88.38%	509778	0.43%
Bootstrap 6	117566728	88.28%	1600895	1.36%
Bootstrap 7	117679867	88.37%	1870031	1.59%
Bootstrap 8	117760573	88.43%	1748183	1.48%

Table B.2.3: Clustering results on bootstrapped Hi-C link datasets. We ran the MetaPhase clustering algorithm on the M-Y and M-PE datasets, producing the results given in the main Results section. We also re-ran the clustering algorithm in each of these cases and applied randomized bootstrapping (that is, re-sampling with replacement of N data points) to the Hi-C link data. Shown are the results of eight bootstrapping runs for each sample.

B.3 Supplementary figures for Chapter 4

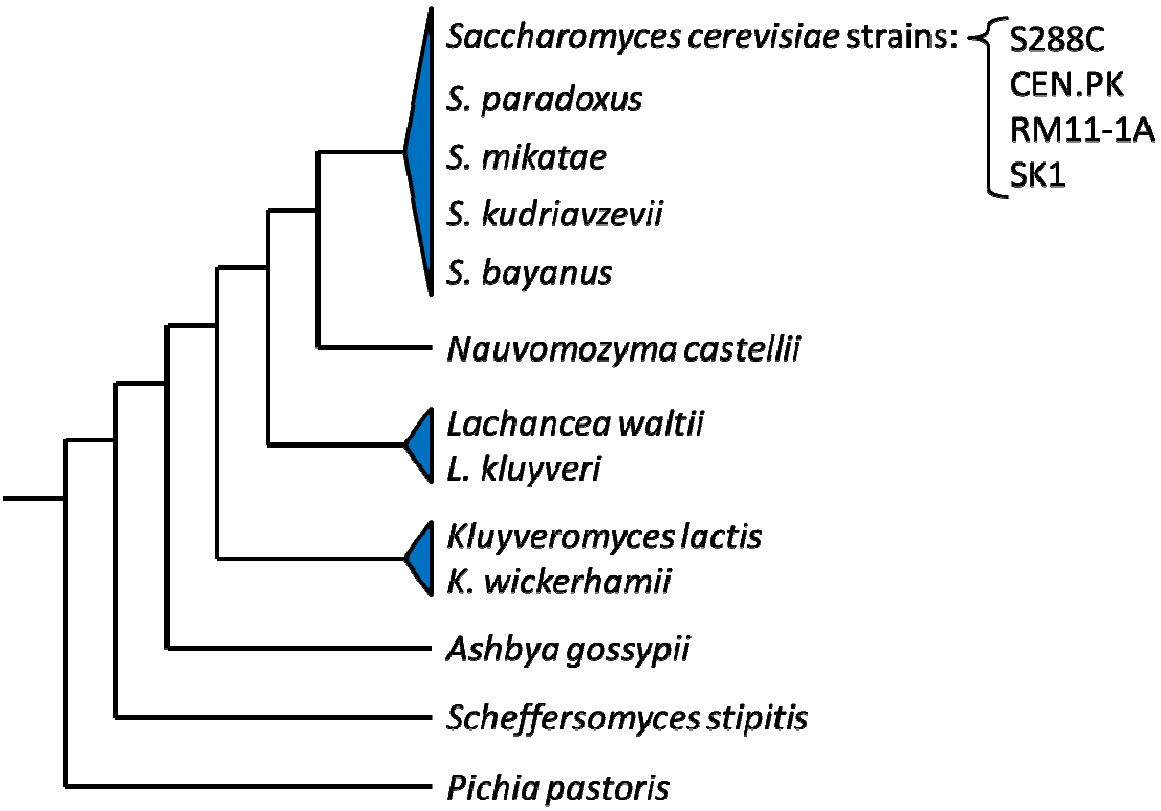


Figure B.3.1: M-Y species phylogeny. Phylogenetic tree of the 16 Ascomycetes yeast strains used in the M-Y sample (Table B.2.1).

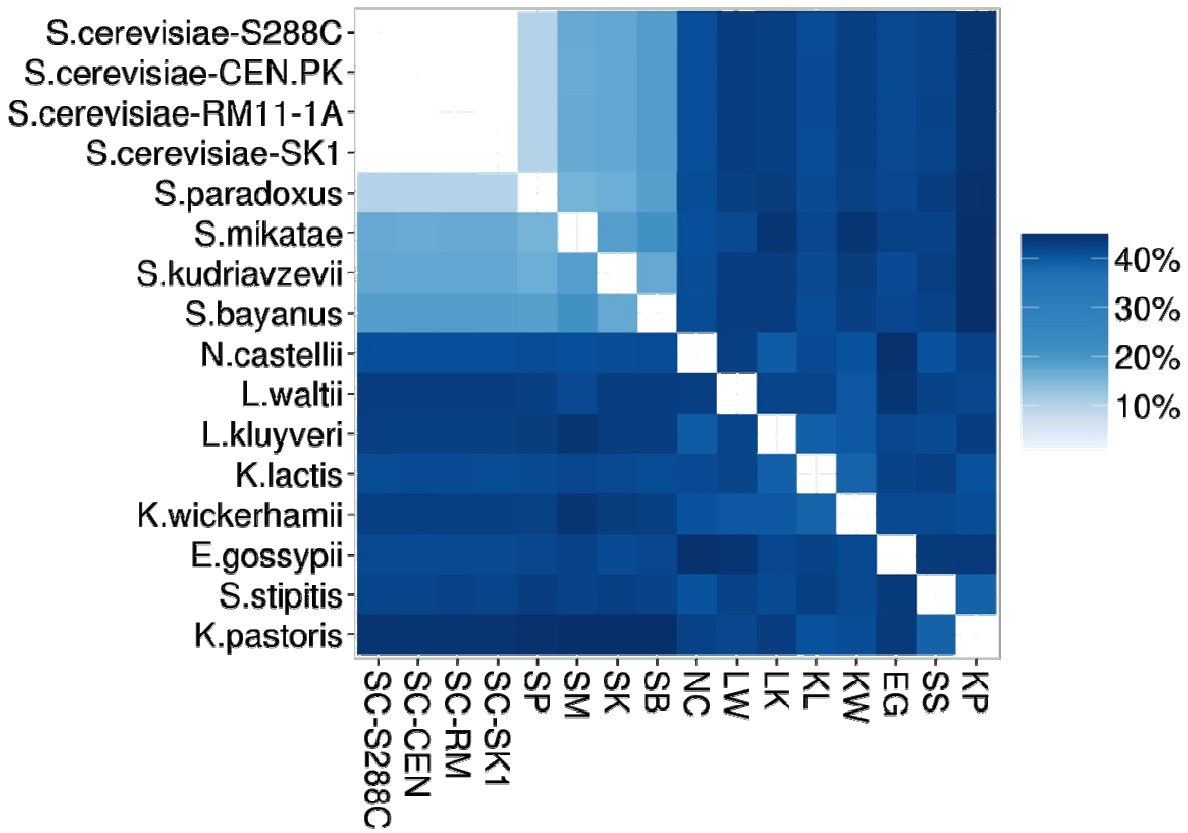


Figure B.3.2: M-Y sequence divergences between species. Divergence rates were calculated as follows: First, a set of essential ORFs in the *Saccharomyces cerevisiae* genome was downloaded from the Yeast Deletion Website. For each essential ORF, orthologous sequences in every other species were found via BLASTn alignment[200], and these sequences were all aligned together using Clustal Omega[204]. Pairwise divergences were calculated by counting the frequency of mismatches among aligned base pairs in the Clustal Omega alignments. This analysis was repeated using essential ORFs from *K. lactis* instead of *S. cerevisiae*, with very similar results (data not shown).

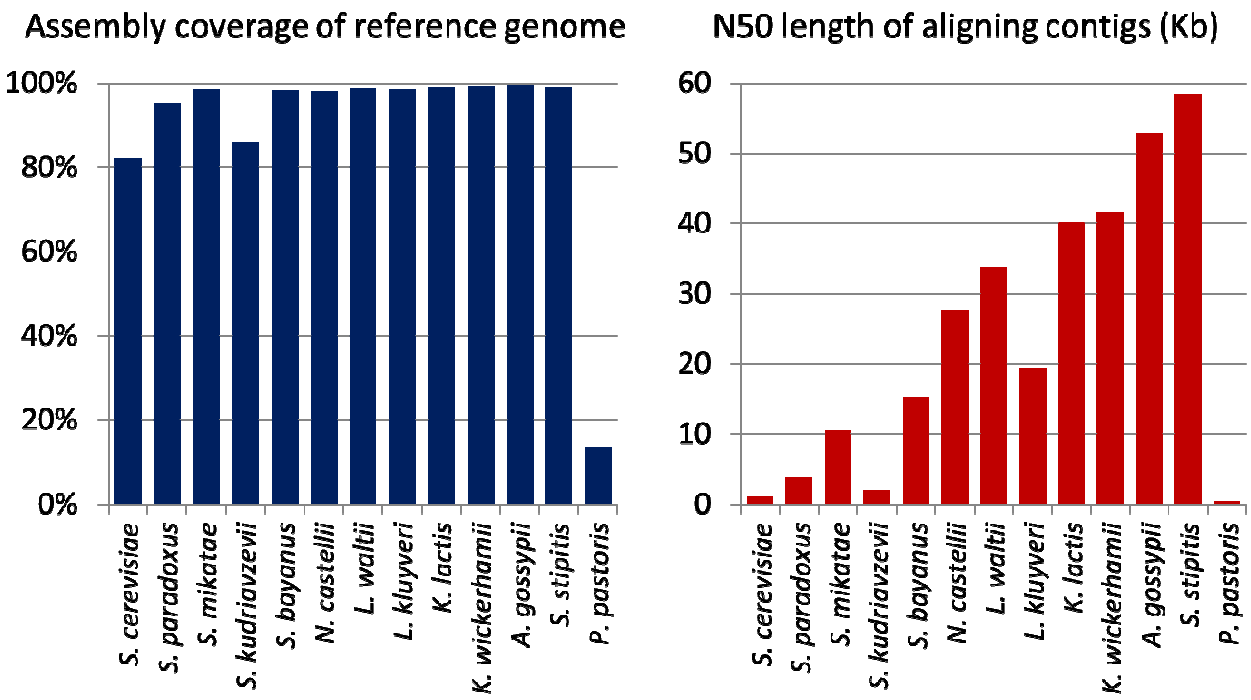


Figure B.3.3: Coverage of M-Y reference genomes by draft metagenome assembly. Contigs from the M-Y draft metagenome assembly were aligned to the reference genomes of each species with BLASTn[200] using the following parameters: `-perc_identity 95 -evalue 1e-30 -word_size 50`. The restrictiveness of these parameters ensured that all alignments generated were greater than 70 bp. Left: The fraction of each reference genome covered by BLASTn alignments. Right: For each reference genome, the N50 length of draft contigs aligning to that genome.

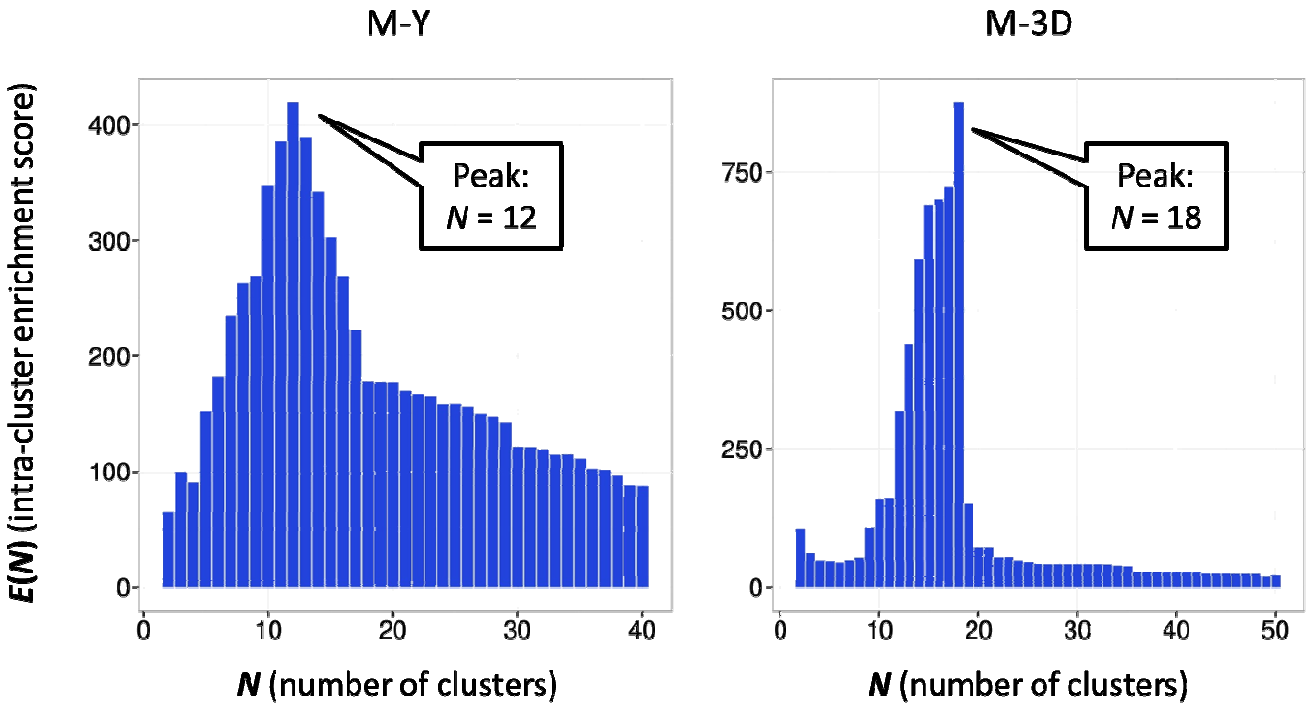


Figure B.3.4: Intra-cluster link enrichment as a function of cluster number in M-Y and M-3D. We ran the hierarchical agglomerative clustering algorithm on the M-Y and M-3D datasets. In this algorithm, the number of clusters gradually decreases as clusters are merged together; to generate this data, we continued clustering all the way down to $N = 1$. Shown is the metric E, or intra-cluster link enrichment, at each value of N. Note that for both M-Y and M-3D the maximum value of $E(N)$ occurs when N is equal to the true number of distinct species present in the draft assembly.

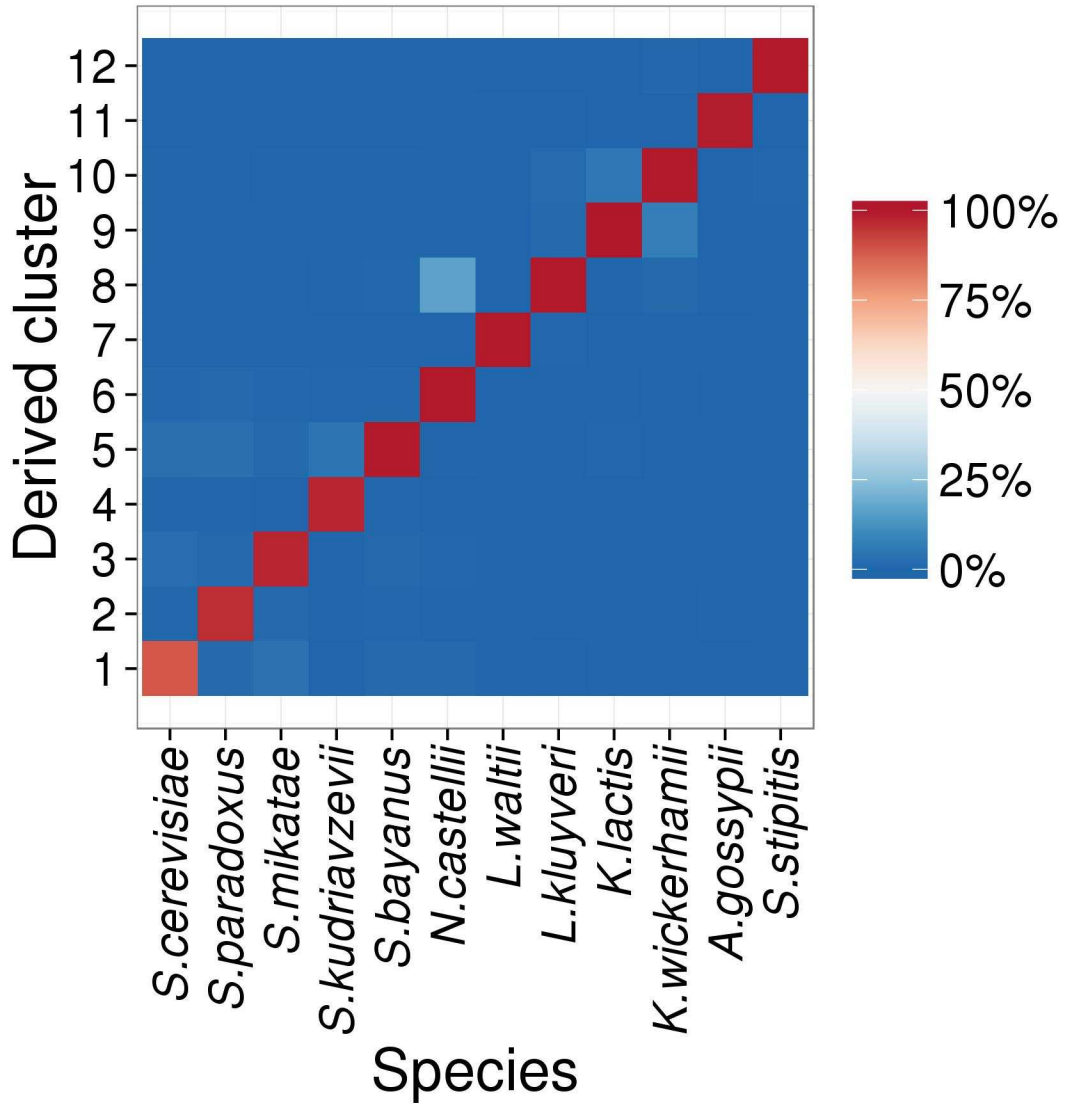


Figure B.3.5: Heatmap of non-unique reference alignments of contigs in each M-Y cluster. This is identical to **Figure 4.2b**, except that all contig alignments to all genomes are shown here, whereas in **Figure 4.2b** only contigs that align uniquely to a single reference genome are shown.

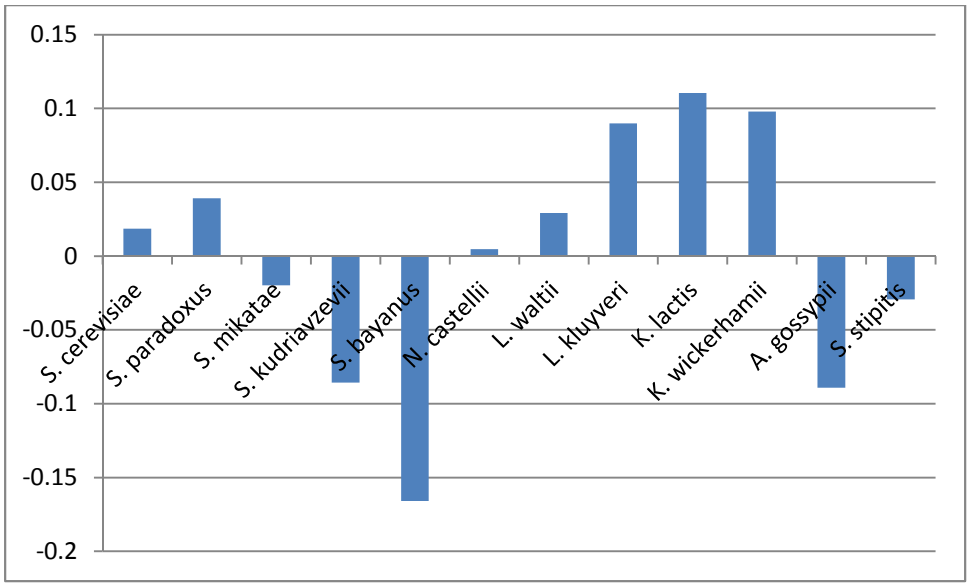


Figure B.3.6: Differential Hi-C efficiency rates by species for the M-Y sample. For each species, the Hi-C efficiency rate was calculated as $E_{species} = f_{species}(\text{Hi-C}) / f_{species}(\text{shotgun})$, where $f_{species}(\text{library})$ is the fraction of reads from a sequencing library that align to the given species' reference genome. These efficiency rates were log-scaled and then normalized to create an average of 0 over all species.

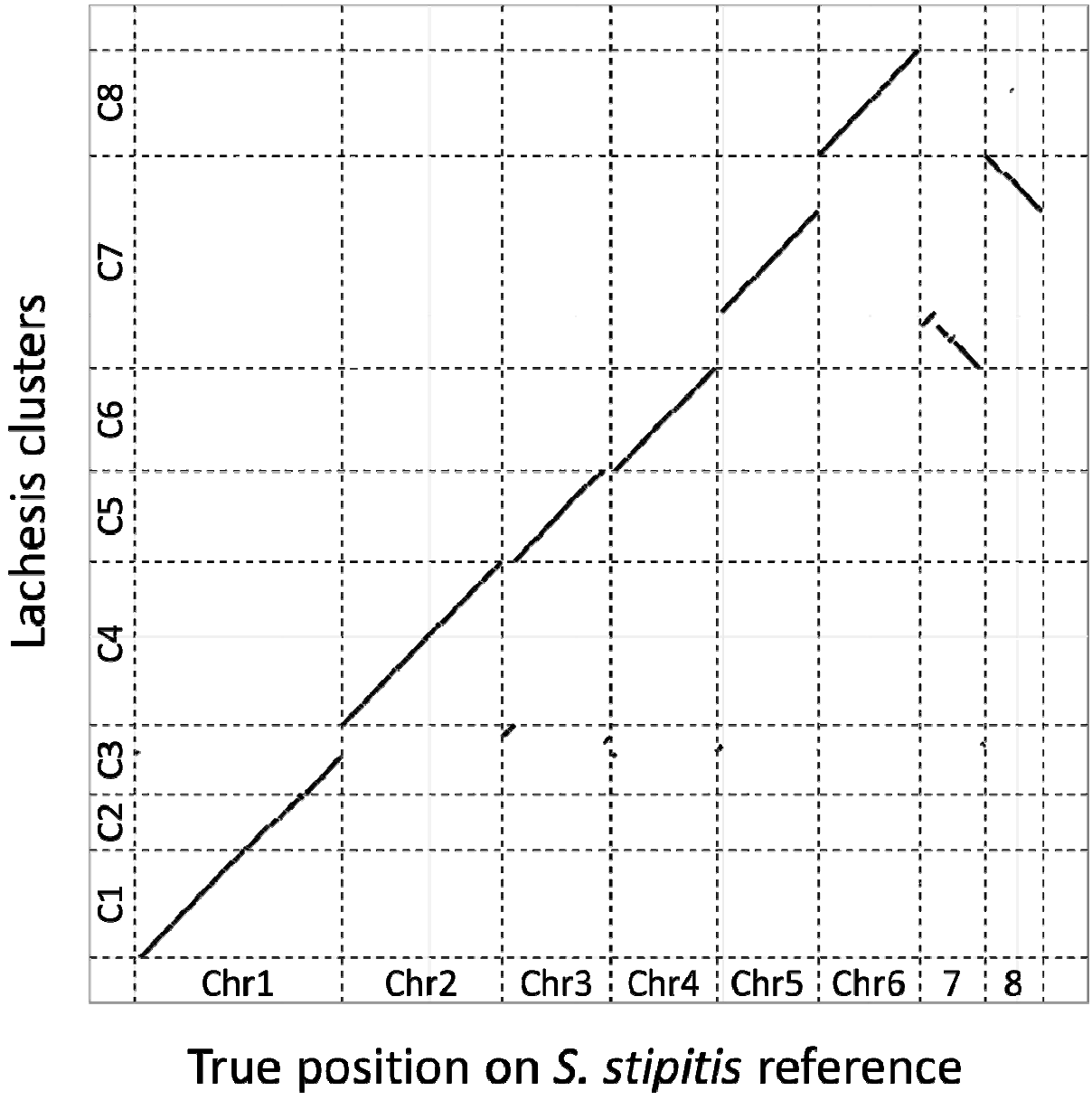


Figure B.3.7: Accuracy of Lachesis assembly of *Scheffersomyces stipitis*. The contigs in the MetaPhase cluster corresponding to *S. stipitis* were clustered, ordered, and oriented with Lachesis[97] (**Figure 4.2c**). Shown here is a validation of the Lachesis assembly. Every contig that is placed by Lachesis and which aligns to the *S. stipitis* reference genome is shown. *x*-axis: the contig's true position in the *S. stipitis* reference. *y*-axis: the contig's placement in the Lachesis assembly (note that both the order of the clusters on the *y*-axis and the overall orientation of each cluster are arbitrary; they are chosen here for visual clarity and are not the same as in **Figure 4.2c**.)

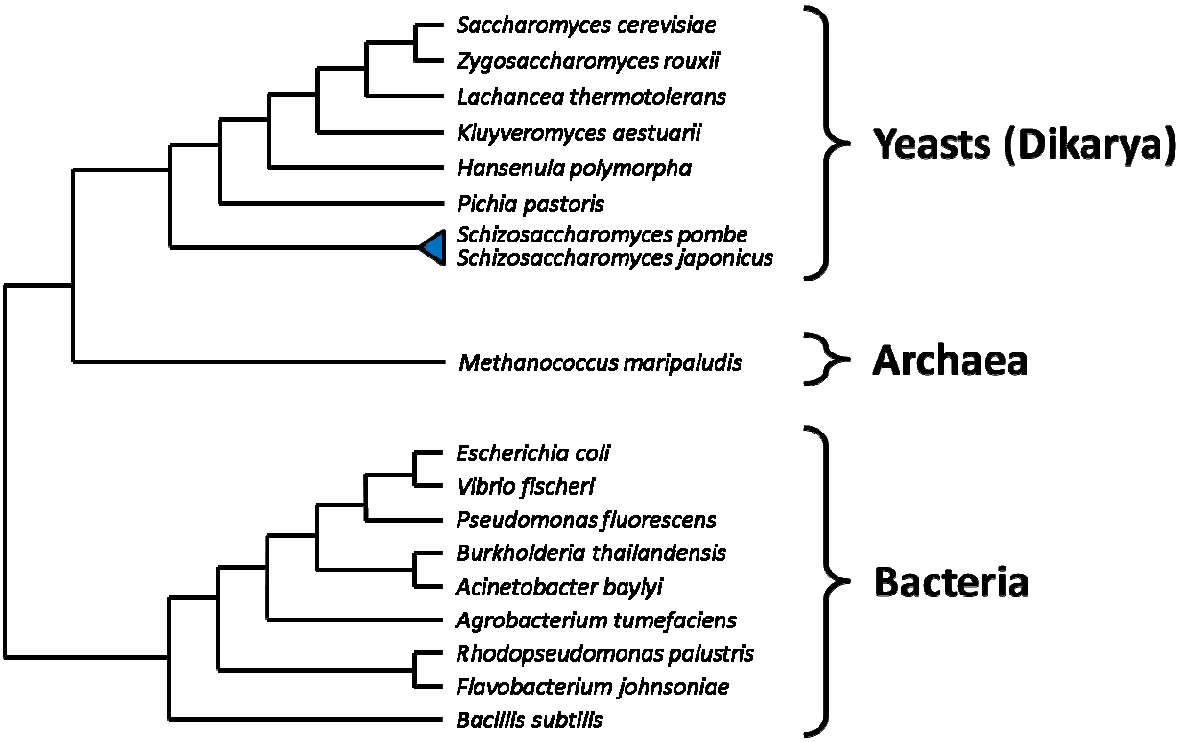


Figure B.3.8: M-3D species phylogeny. Phylogenetic tree of the 18 yeast, archaeal, and bacterial strains used in the M-3D sample (Table B.2.2).

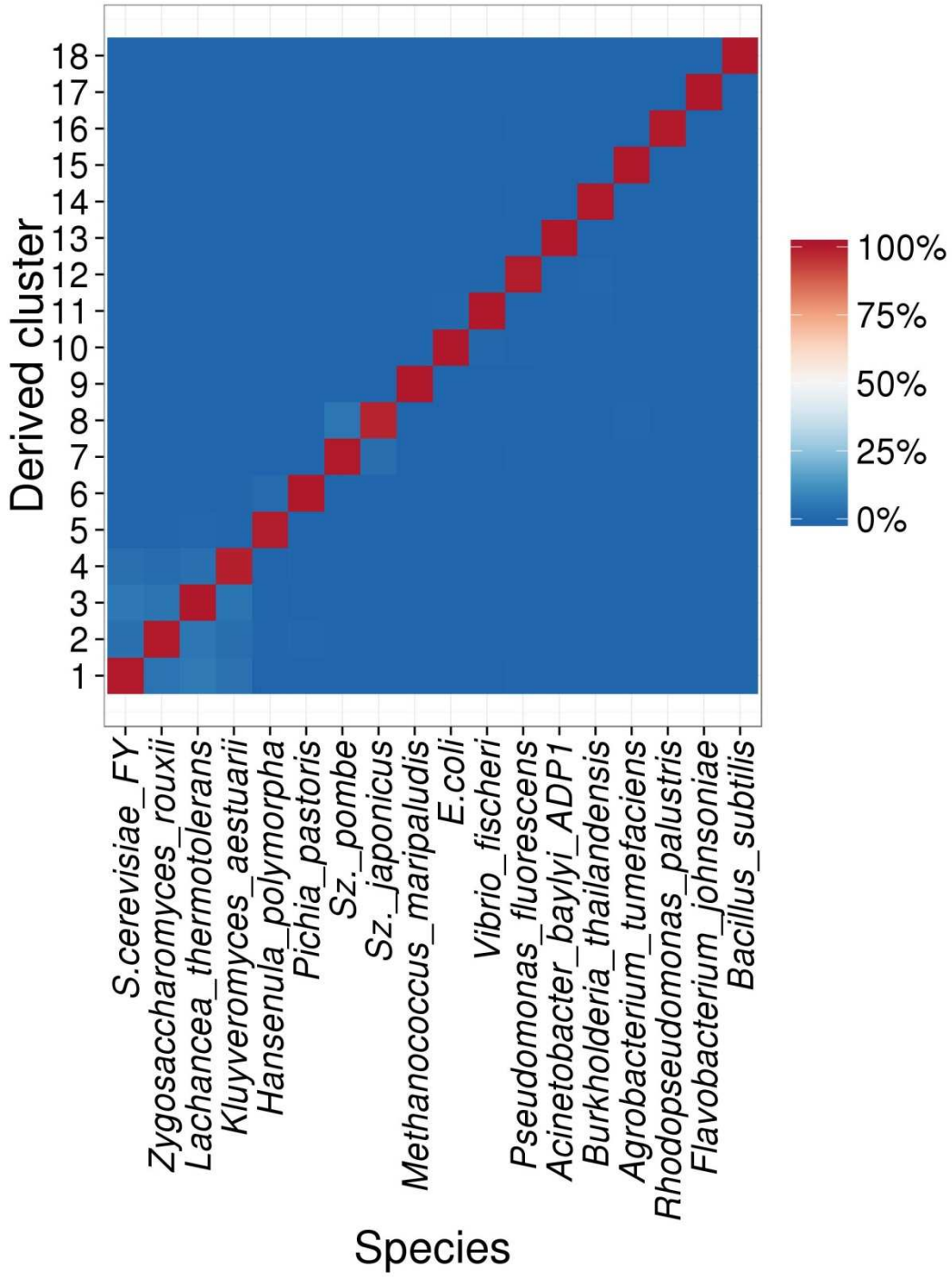


Figure B.3.9: Heatmap of non-unique reference alignments of contigs in each M-3D cluster. This is identical to **Figure 4.3b**, except that all contig alignments to all genomes are shown here, whereas in **Figure 4.3b** only contigs that align uniquely to a single reference genome are shown.

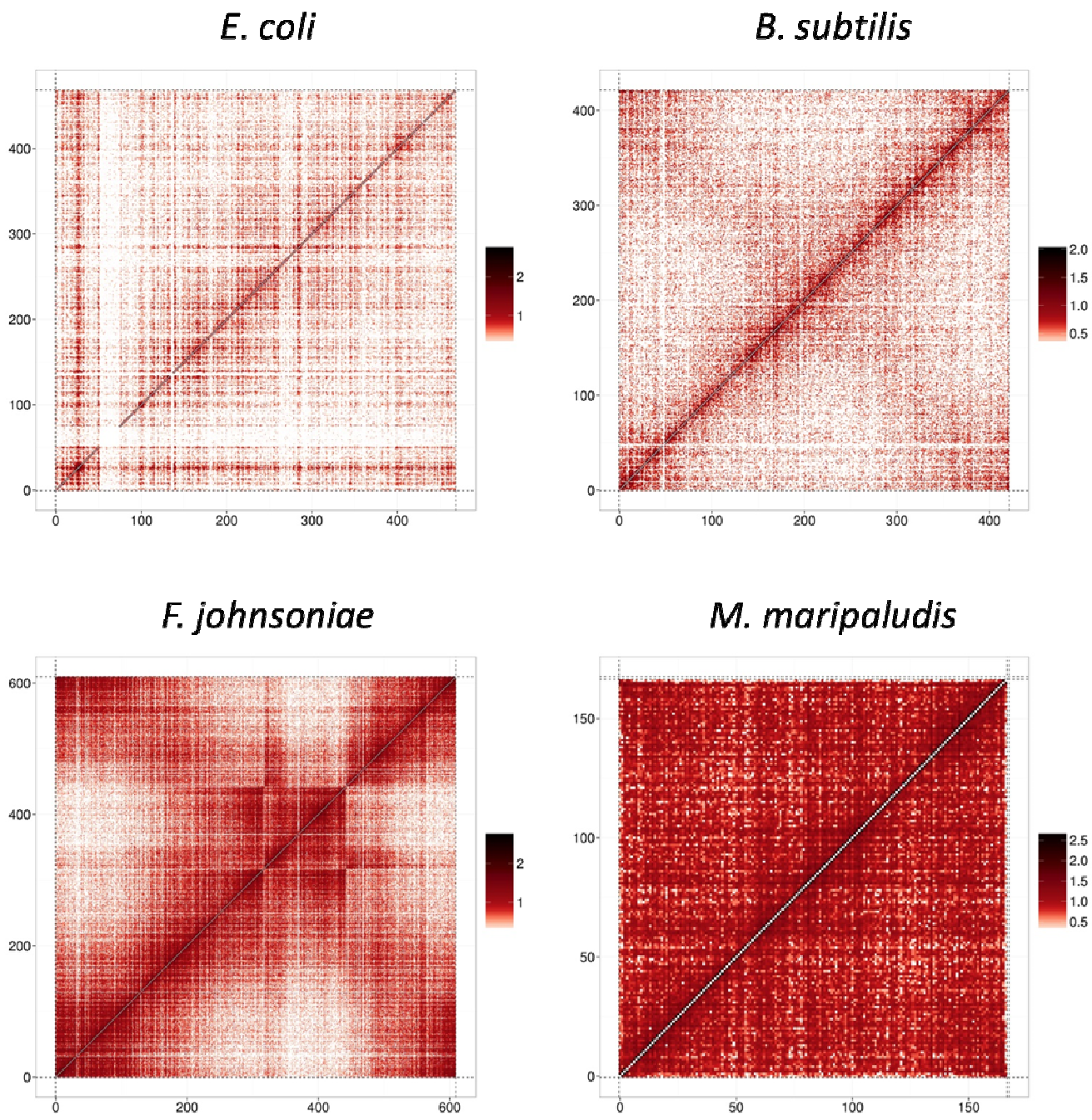


Figure B.3.10: Heatmaps of M-3D Hi-C links aligned to prokaryotic reference genomes. Reads from the M-3D *Hin*dIII non-resuspended library (Table B.2.3) were aligned to the draft assemblies of four prokaryotic species present in the M-3D sample. Each heatmap has a resolution of 10 Kb, and the legend indicates the $\log_{10}$ of link density.

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

The material in this Appendix, like that of **Chapter 5**, is based on the following peer-reviewed publication[100]:

Andrew Adey, **Joshua N. Burton**, **Jacob O. Kitzman**, Joseph B. Hiatt, Alexandra P. Lewis, Beth K. Martin, Ruolan Qiu, Choli Lee and Jay Shendure. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).

Boldface indicates authors who contributed equally to this work.

Supplementary methods are in section C.1. Supplementary tables are in section C.2. Supplementary figures are in section C.2.

C.1 Supplementary methods for Chapter 5

HeLa cell culture: HeLa cell cultures (HeLa ATCC, CCL-2 (laboratory stock); HeLa S3 ATCC, CCL-2.2 (laboratory stock); Chang liver ATCC, CCL-13; L132 ATCC, CCL-5; KB ATCC, CCL-17; HEp-2 ATCC, CCL-23; WISH ATCC, CCL-25; Intestine 407 ATCC, CCL-6; FL ATCC, CCL-62; AV-3 ATCC, CCL-21) were maintained in DMEM F-12, HEPES (Gibco) media supplemented with fetal bovine serum (FBS) to 10% and a 1× final concentration of pen-strep antibiotic (Gibco).

Shotgun sequencing, alignment and variant calling: All shotgun libraries were constructed using standard ligation chemistry methods and sequenced on an Illumina HiSeq 2000. Reads were aligned to the human reference genome (hg19, b37) using BWA[199] followed by duplicate removal, quality score recalibration and local indel realignment using GATK[74]. SNVs were called using samtools[205], indel variants were called using GATK[74] and short tandem repeats (STRs) were called using LobSTR[206]. Gene ontology term analysis was carried out using DAVID[207]. Data sets used for each analysis are depicted as a flow chart in Figure C.3.48.

Read depth copy number analysis: Shotgun reads for HeLa and Human Genome Diversity Project (HGDP) control genomes[153] along with a similarly prepared control library with a matched G + C profile were aligned using mrsFAST[208], processed as described previously[209] to generate read depth-based copy number predictions within non-overlapping windows of singly unique nucleotide *k*-mers (SUNK windows). Copy-number calling in HeLa was carried out at high (approximately 1.5-kb) and low (approximately 77-kb) resolution using an HMM, and a recalibration process was then used to account for widespread aneuploidy. Short amplifications and deletions were identified using a sliding-window approach. Copy-number calling was also carried out on HeLa S3 at both high and low resolutions, as well as on the eight additional HeLa strains at low resolution, and profiles were compared between strains. Regions of LOH were identified using a two-state HMM that used the fraction of homozygous SNVs in non-repetitive regions across low-resolution copy-number windows described above.

Mate-pair library construction, sequencing and analysis: Library construction for 40-kb mate-pair libraries was carried out starting with fosmid clone DNA pooled within each original fosmid preparation, using a protocol similar to one described previously[67]. Libraries of approximately 3-kb inserts were constructed following protocols described previously[210]. After read trimming and alignment, reads were split into classes based on aligned orientation and insert size, and processed using sliding windows to identify regions of probable structural rearrangement.

Fosmid pool construction, sequencing and haplotype phasing: Three replicate fosmid libraries were prepared as described previously[67], and then partitioned by limited dilution into 96 sub-libraries. This was followed by outgrowth, barcoded transposase-based library preparation[211], sequencing and alignment. Clone boundaries were inferred as described previously[82], and base calls were made at all heterozygous variant positions as ascertained from whole-genome shotgun sequencing. Overlapping clones were merged to consensus haplotype blocks using an implementation of the ReFHap algorithm[103]. Within the majority of the HeLa genome in which haplotypes are unequally amplified, adjacent blocks were merged to create scaffolds, using an HMM that finds the most likely phase of neighbouring blocks given their shotgun allele frequencies of inherited variants (those found within the 1000 Genomes Project. This produced a final set of haplotype scaffolds with an N50 size of 44.8 Mb, which was then used in conjunction with copy-number calls to estimate haplotype-resolved copy number for HeLa. Haplotype scaffolds were analyzed for variant population frequencies to investigate the ancestral origin of phased blocks. Finally, overall copy numbers were compared among all HeLa strains sequenced in this study.

Long-read phase validation: Genomic DNA from HeLa CCL-2 was mechanically sheared using a Covaris G-tube column and standard microcentrifuge following the manufacturer's instructions, and this produced a mean fragment size of approximately 10 kb. Single-molecule real-time sequencing libraries for the Pacific Biosciences RS sequencer were prepared using the Pacific Biosciences DNA Template Prep Kit (3–10 kb), and the resulting library was sequenced across eight cells using a 90-min movie. Resulting base calls were aligned to the genome with bwasw (using parameters '-b5 -q2 -r1 -z1'). Reads that overlapped at least two phased SNPs were considered, excluding those within ±10 bp of an insertion or deletion in the alignment.

Identification of putative post-aneuploidy mutations: We searched for candidate somatic post-aneuploidy mutations by taking the initial set of SNVs called from the shotgun sequencing data and filtering to remove probable germline variants. SNVs that were phased on a duplicated haplotype but that were polymorphic between the two duplicated copies were identified. Common polymorphisms and sequencing artefacts were removed by filtering against repeat annotations and control genomes.

HPV-18 insertion characterization: The HPV-18 integration locus was characterized by aligning all fosmid libraries to a modified genome that included the HPV-18 reference genome as an additional chromosome. Interchromosomal read pairs, fosmid-pool coverage profiles, and copy-number calls were used to determine the repeat structure of the chromosome 8q24.21–HPV-18 integration locus. Polymerase-chain-reaction primers were then designed to amplify the proposed breakpoints, and then sequencing for base-pair resolution was carried out.

ENCODE and RNA-seq phasing: Directional, PolyA$^+$ RNA-seq data generated in-house on HeLa S3 were analyzed in parallel with publically available ENCODE epigenomics and transcriptomics data downloaded from the online data portal for HeLa S3, and RNA-seq data on HeLa CCL-2[151]. RNA-seq reads were aligned using TopHat[212] and transcript quantification was carried out using Cufflinks[213]. Haplotype phasing was performed by genotyping aligned-sequence data for all phased SNVs and assigning haplotype contributions to either peaks (epigenomics data sets) or RPKM (RNA-seq data sets), and then carrying out copy-number normalization. Reference bias was investigated in all tracks and removed in a subset to identify its impact on outlier calling. Haplotype-specific peaks were then identified in all data tracks. Finally, a meta-analysis of all data tracks was used to identify large regions of haplotype imbalance.

C.2 Supplementary tables for Chapter 5

a **Shotgun Sequencing**

ID	Unique Read Pairs	Unique reads (%)	Insert Size (bp)	Aligned Bases (Gbp)	Fold Coverage*
HELA					
HELA.s.1	297,931,188	97.5	142 +/- 29	48.36	17.27
HELA.s.2	155,257,336	96.3	131 +/- 28	32.80	11.71
HELA.s.3	527,420,302	90.5	206 +/- 39	94.32	33.69
HELA.s.4	430,404,271	85.3	196 +/- 45	72.28	25.81
			TOTAL	**247.76**	**88.48**
HELA S3					
S3.s.1	241,974,865	93.8	267 +/- 119	29.15	10.41
S3.s.2	234,719,279	94.2	270 +/- 116	28.09	10.03
S3.s.3	32,257,707	97.1	289 +/- 122	6.62	2.36
S3.s.4	42,792,269	98.2	300 +/- 114	8.76	3.13
			TOTAL	**72.62**	**25.93**

b **Fosmid Clone Pool Sequencing**

ID	Called Clones	Average Clone Size (bp)	Physical Coverage*
Hapfos1	171,580	33,495	2.05
Hapfos2	228,667	34,851	2.85
Hapfos3	118,046	33,204	1.40
TOTAL	**518,293**	**-**	**6.30**

c

Mate Pair Sequencing

ID	Type	Unique Concordant Pairs	Insert Size (bp)	Physical Coverage*	Unique Discordant Pairs (Intra)	Unique Discordant Pairs (Inter)	Unique Discordant Pairs (Inversion)
Matepair1	circularization	85,311,942	2,862 +/- 453	87.20	356,696	10,522,598	300,785
Matepair2	circularization	46,363,211	2,861 +/- 453	47.38	209,998	6,041,976	178,559
			TOTAL	**134.58**			
Matefos1	fosmid end	86,462	34,992 +/- 4,309	1.08	248	3,279	207
Matefos2	fosmid end	163,115	35,969 +/- 4,211	2.10	321	4,895	400
Matefos3	fosmid end	94,503	35,064 +/- 3,321	1.18	6,367**	94,588**	6,471**
			TOTAL	**4.36**			

d **RNA-Seq**

ID	Unique Read Pairs	Insert Size (bp)	Correct Strand (%)	Ribosomal (%)	Aligned to Coding (%)	Aligned to UTR (%)
S3.RNA	227,472,084	173 +/- 46	99.42	0.07	45.68	41.73

*Assuming 2.8 Gbp alignable reference
**Higher discordant rate due to increased intermolecular ligation events

e **HeLa 8 Strains**

ID	Name	Total Aligned Bases (Gb)	Insert Size (bp)	Coverage*
CCL-13	Chang Liver	11.3	138 +/- 100	4.0
CCL-5	L132	12.1	138 +/- 109	4.3
CCL-17	KB	10.5	145 +/- 110	3.8
CCL-23	HEp-2	10.2	147 +/- 107	3.7
CCL-25	WISH	10.1	138 +/- 109	3.6
CCL-6	Intestine 407	10.7	140 +/- 70	3.8
CCL-62	FL	10.8	146 +/- 95	3.9
CCL-21	AV-3	9.8	144 +/- 107	3.5

f **PacBio RS Long Read Sequencing**

ID	Total Reads	Aligned Reads	Aligned Informative (Inf) Reads**	Aligned Read Length (all, bp)	Aligned Read Length (Inf, bp)
HELA.PB5KB	601,217	114,584	6,746	1,428 +/- 1488	2970 +/- 1645

*Assuming 2.8 Gbp alignable reference
** Reads overlapping at least 2 heterozygous, phased SNVs with aligned positions >= 10bp from nearest alignment indel

Table C.2.1 | Sequencing data obtained.
Six major types of sequence data were obtained. **a.** Shotgun sequencing data obtained for HeLa (CCL-2) and HeLa S3. **b.** Haplotype specific fosmid clone pool sequencing. **c.** Mate pair sequencing for both 3kb jumping libraries as well as fosmid based 40kb jumping libraries. **d.** Directional PolyA RNA-Seq library for HeLa S3 generated in house. **e.** Shotgun sequencing of 8 additional HeLa strains. **f.** PacBio RS long read sequencing of HeLa CCL-2 for haplotype phasing validation.

	ID	iSize Mean	iSize Stdev	Total Aligned Bases (Gb)	Fold Coverage*
Dinka	DNK02	258	117	92.37	32.99
French	HGDP00521	283	124	95.35	34.05
Papuan	HGDP00542	260	119	92.83	33.15
Sardinian	HGDP00665	270	119	88.48	31.60
Han	HGDP00778	269	120	97.74	34.91
Yoruban	HGDP00927	279	120	113.82	40.65
Karitiana	HGDP00998	267	128	96.69	34.53
San	HGDP01029	272	126	122.75	43.84
Madenka	HGDP01284	264	123	91.14	32.55
Dai	HGDP01307	256	125	97.40	34.79
Mbuti	HGDP0456	265	120	85.89	30.68

Table C.2.2 | HGDP control genomes. Sequencing data summary for 11 HGDP control individuals from Rohland, N. and Reich, D (2012)[214]. Assumes 2.8 Gbp alignable reference.

	AVERAGES			
	5 African	2 Euro	Reich 11	HELA CCL-2
Number of SNVs	4,699,881	3,782,752	4,178,701	4,068,395
Number of indels	358,218	320,632	334,885	417,471*
Number of 1kG SNVs	4,072,563	3,393,182	3,663,541	3,670,543
Number of non-1kG SNVs	627,319	389,571	515,159	397,852
Number of 1kG indels	212,327	181,597	193,522	195,613
Number of non-1kG indels	145,891	139,035	141,363	221,858
% SNVs that are homozygous	36.11%	39.45%	40.42%	43.99%
Ti/Tv for SNVs in 1kG	2.15	2.15	2.15	2.14
Ti/Tv for SNVs not in 1kG	1.69	1.60	1.65	1.55
Private Protein-Altering (PPA) SNVs	508.6	258	390.9	269
PPA SNVs in COSMIC	3.2	0.5	2.6	1
PPA SNVs in Cancer Genes	13.4	4	8.7	4
PPA indels	24.8	10.5	17.5	35*
PPA indels in COSMIC	0	0	0	0
PPA indels in Cancer Genes	0	0	0.1	1
Total bases in homozygous tracts	49,360,241	52,718,117	83,552,819	374,139,228

Table C.2.3 | Summary of variants and regions of homozygosity for HeLa and control genomes.

Variants with a minimum of 8X coverage were annotated as protein-altering using the SeattleSeq annotation server. Private protein-altering ("PPA") variants were those not observed among the 1000 Genomes Project ("1kG") or the Exome Sequencing Project 6500 call set, and found outside regions annotated for excessive sequence depth (HiSeq top 5%ile coverage track from the UCSC genome browser). For comparison to COSMIC database, the variant allele was required to match exactly. Comparison to CGP used gene-level overlap. * HeLa CCL-2 has an increased indel call rate due to higher depth of coverage.

Copy Number	Size (kbp)	% of Genome*	HRCN	Size in HRCN (kbp)	% of Genome*
1	5,638	0.22%	1:0	5,638	0.22%
2	691,955	27.21%	2:0	256,419	10.08%
			1:1	435,536	17.13%
3	1,556,135	61.19%	3:0	200,761	7.89%
			2:1	1,355,373	53.30%
4	140,458	5.52%	4:0	31,388	1.23%
			3:1	31,060	1.22%
			2:2	78,010	3.07%
5	42,249	1.66%	5:0	18,081	0.71%
			4:1	15,719	0.62%
			3:2	8,449	0.33%
6	15,702	0.62%	6:0	13,292	0.52%
			5:1	689	0.03%
			4:2	1,721	0.07%
			3:3	0	0.00%
7	3,197	0.13%	7:0	2,225	0.09%
			6:1	0	0.00%
			5:2	971	0.04%
			4:3	0	0.00%

*Of high-quality, alignable regions

Table C.2.4 | Haplotype-Resolved Copy Number (HRCN) profile of HeLa CCL-2. Proportion of the genome (UCSC hg19/GRC37h, excluding assembly gaps and sgemental duplications) at each haplotype-resolved copy number (HRCN) state.

Chromosome-arm sized LOH regions

Chromosome	Region	Size (Mb)	CN=1?
2q	106,690,345-qter	136.4	No
3q	94,582,003-qter	103.3	No
5p	pter-centromere	46.1	No
6p,6q	pter-qter	170.7	No
11q	102,239,620-qter	32.7	No
13q	19,167,980	95.9	No
19p	pter-12,893,034	12.9	No
22q	16,385,650-qter	34.8	No
Xp,Xq	pter-qter	152.1	No

Short LOH regions

Chromosome	Region	Size (kb)	CN=1?
2	40,339,750-41,992,745	1,653	No
3	80,281,400-81,385,274	1,104	Yes
4	158,267,826-161,280,735	3,013	Yes
4	172,475,207-173,703,673	1,228	No
7	15,684,943-16,895,503	1,211	No
7	123,832,242-126,478,678	2,646	No
11	184,961-2,876,557	2,692	Yes
11	22,372,309-24,503,054	2,131	No

Table C.2.5 | Large regions of LOH in HeLa CCL-2.

IN 1000 GENOMES?	Yes	No
Allele not observed (depth >= 2) in clones	179000	107501
Allele observed only in unphased clones	3342	23603
Unphased due to inconsistency (observed in A and B clones with equal scores)	3732	8510
Phased by majority rule among clones, with conflicting phase calls between clones	30496	32326
Phased unanimously among clones, only one allele observed	613890	69806
Phased unanimously among clones, both alleles observed	1143908	62709

IN 1000 GENOMES?	Yes	Yes	Yes	Yes	No	No	No	No
IN SEGDUP?	No	No	Yes	Yes	No	No	Yes	Yes
REPEAT-MASKED?	No	Yes	No	Yes	No	Yes	No	Yes
Allele not observed (depth >= 2) in clones	67404	100728	4997	6871	9051	474'0	22787	28253
Allele observed only in unphased clones	827	1637	479	399	6612	10130	3647	3214
Unphased due to inconsistency (observed in A and B clones with equal scores)	1'47	1948	315	322	1808	3559	1501	1642
Phased by majority rule among clones, with conflicting phase calls between clones	12708	14986	1394	1408	8071	12333	5540	6382
Phased unanimously among clones, only one allele observed	268193	323668	'0230	1'799	14379	30783	11705	12942
Phased unanimously among clones, both alleles observed	545114	563384	17552	17858	21966	29179	5541	6023

Table C.2.6 | Phasing status of heterozygous SNVs in HeLa CCL-2.

Counts of heterozygous SNVs are shown by phasing status (phased or unphased, and reason) and overlap with 1000 Genomes Project data and genomic repeats (segmental duplications or regions identified by Repeat Masker). For unphased variants, the reason for lack of phase assignment is indicated (does not appear among clones, or alleles are inconsistent among phased clones). Phased variants are separated by the degree of support among clone data (both alleles observed with no inconsistency between clones, or only one allele observed with no inconsistency between clones, or inconsistencies between clones resolved by majority rule).

HRCN (Total : HapA : HapB)	Total genomic extent (bp)	Total bp of duplicated haplotype(s) (extent x copy)	Number clone-confirmed mutations	Clone-confirmed somatic mutation frequency (per bp x 10^6)	Expected frequency given 61% sensitivity (per bp x 10^6)
2:2:0	**369,962,202**	**739,924,404**	**1022**	**1.38**	**2.26**
3:3:0	**328,232,258**	**984,696,774**	**1437**	**1.46**	**2.39**
4:4:0	**54,229,430**	**216,917,720**	**287**	**1.32**	**2.17**
5:5:0	**11,618,893**	**58,094,465**	**39**	**0.67**	**1.10**
6:6:0	**33,636,597**	**201,819,582**	**98**	**0.49**	**0.79**
3:2:1	1,395,662,889	2,791,325,778	4128	1.48	2.42
4:2:2*	221,271,991	885,087,964	891	1.01	1.65
4:3:1	64,697,997	194,093,991	216	1.11	1.82
5:3:2*	2,368,480	11,842,400	7	0.59	0.97
5:4:1	19,813,202	79,252,808	30	0.38	0.62
6:4:2*	5,861,548	35,169,288	10	0.28	0.47
TOTAL	**2,507,355,487**	**6,198,225,174**	**8165**	**1.32**	**2.16**

Table C.2.7 | Clone-confirmed somatic mutation frequency.

Counts and frequencies of somatic mutations in the HeLa CCL-2 genome. The total number of bases in the genome at each haplotype-resolved copy number (HRCN) state (total copies:haplotype A copies:haplotype B copies) are listed, as well as the number of somatic mutations observed and confirmed by clone pool sequencing. Mutations occurring on duplicated haplotypes could arise on any of the haplotype copies, so mutation rate is taken as (# sites in given C.N.) / ([total bases within reference at C.N.] x [copies of duplicated haplotype(s)]). Shaded rows indicate regions of LOH (haplotype B copies = 0).

*In these regions, both haplotypes are duplicated, so mutations on either were considered; in all other cases, only mutations occurring on the major haplotype were counted.

ID	Genotyped for:	Num. >= 8X (both)	Num. Shared	Percent Shared
S3 DNA	**CCL-2 SNVs**	**204,800**	**194,416**	**94.93**
S3 DNA	CCL-2 protein-altering SNVs	301	249	82.72
S3 RNA	**CCL-2 SNVs**	**22,772**	**22,129**	**97.12**
S3 RNA	Shared S3 & CCL-2 protein-altering SNVs	74	65	87.84
CCL-2	**S3 SNVs**	**55,540**	**50,610**	**91.12**
CCL-13	**CCL-2 SNVs**	**47,781**	**43,507**	**91.06**
CCL-5	CCL-2 SNVs	55,696	50,596	90.84
CCL-17	**CCL-2 SNVs**	**45,734**	**41,847**	**91.50**
CCL-23	CCL-2 SNVs	41,668	37,914	90.99
CCL-25	**CCL-2 SNVs**	**44,262**	**40,632**	**91.80**
CCL-6	CCL-2 SNVs	37,119	33,623	90.58
CCL-62	**CCL-2 SNVs**	**42,249**	**38,481**	**91.08**
CCL-21	CCL-2 SNVs	38,906	35,476	91.18

Table C.2.8 | Variants shared between HeLa strains

HeLa S3 shotgun reads, HeLa S3 RNA-Seq reads, and shotgun reads from 8 additional HeLa strains were genotyped at HeLa CCL-2 variant sites for the presence or absence of the HeLa CCL-2 variant allele. Positions were only included if both HeLa CCL-2 and the data set have a coverage of at least 8x, and are not in segmental duplications or at 1000 Genomes Project sites.

C.3 Supplementary figures for Chapter 5

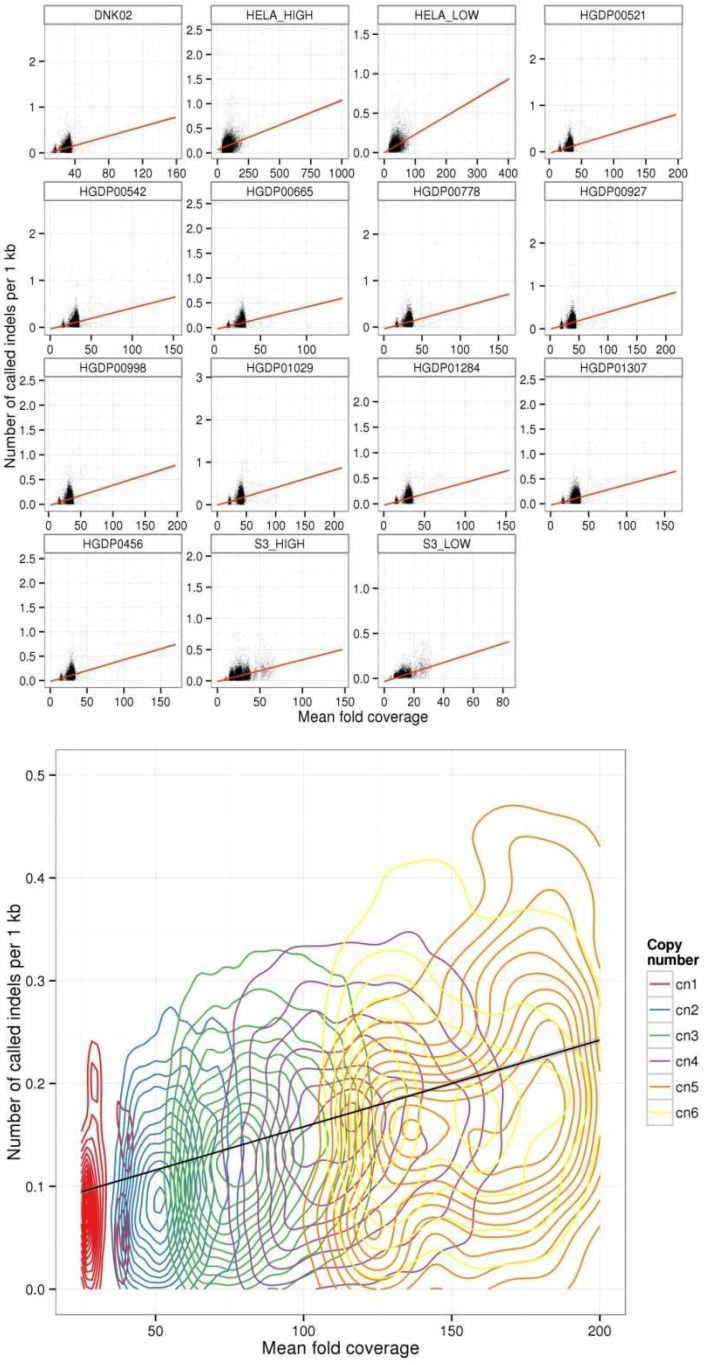


Figure C.3.1 | Indel calling by coverage
a. Counts of indels called is plotted versus read depth at each indel for HeLa and 11 HGDP controls. Shotgun reads from HeLa CCL-2 as well as HeLa S3 were randomly downsampled in order to study the effects of lower coverage upon indel counts in each genome. Mean coverage was HeLa CCL-2 full dataset: ("HELA_HIGH"), ~88X; HeLa CCL-2, subsampled ("HELA_LOW"), ~35X; HeLa S3 full dataset ("S3_HIGH"), ~26X; HeLa S3 subsampled ("S3_LOW"), ~12X; 11 HGDP controls, ~30-45X. Each point represents one of the low resolution SUNK windows (mean size, 77 kbp), and for each window, mean read depth and total number of indel calls per kilobase were determined. In all genomes analyzed, there is a strong correlation between number of calls by read depth. **b.** Indel calls in HeLa (88X) for points as in **a** but shown as a 2d density contour plot, split by underlying copy number. As the mean coverage increases with the copy number so does the ability to call indels, resulting in a higher call count per kilobase at higher copy numbers.

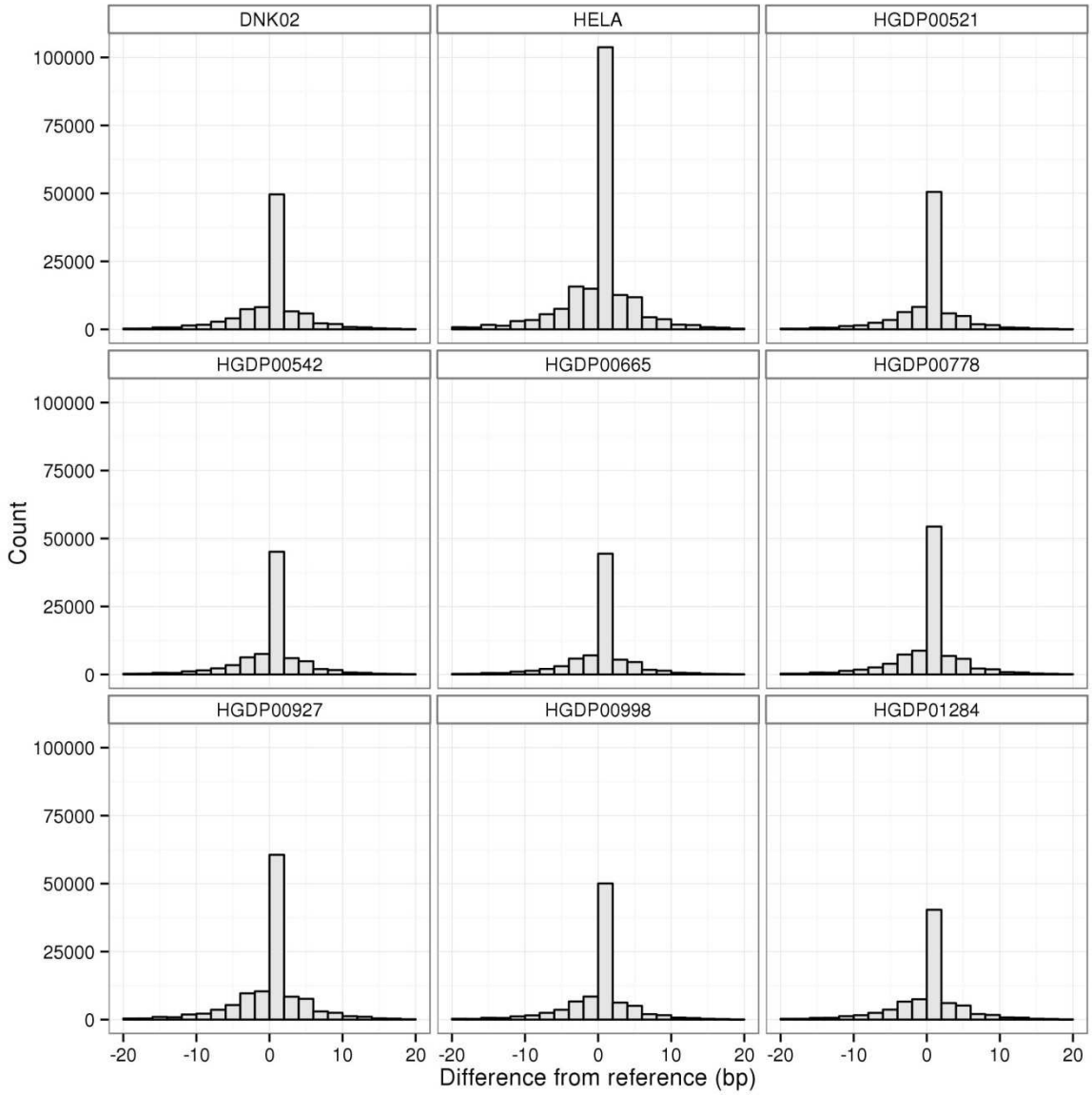


Figure C.3.2 | STR profiling with lobSTR.

Short Tandem Repeats (STRs) were identified using lobSTR[206] for HeLa as well as eight of the diversity panel control individuals[214]. Repeats with a coverage of at least 10 are represented above as a histogram of counts for the length difference in base pairs of called STRs from the reference. While more calls above the coverage threshold are called for HeLa, likely due to having 88X coverage compared to ~30-45X for the control samples, the profile of lengths are comparable between all samples.

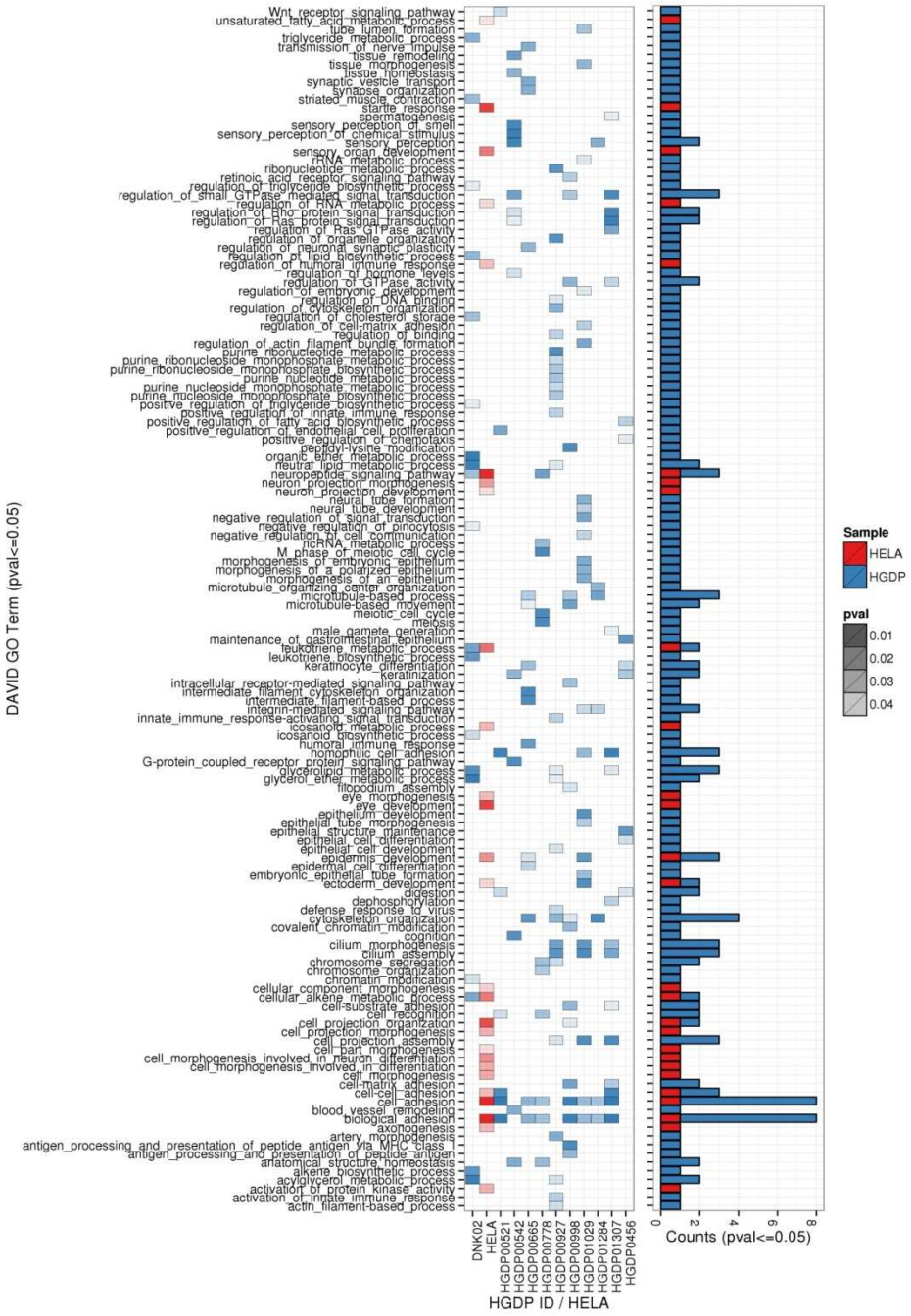


Figure C.3.3 | Gene ontology enrichment analysis for genes with protein-altering variants in Hela CCL-2 and 11 HGDP controls.

For HeLa CCL-2 and the 11 control genomes, a list of genes with protein-altering SNVs, indels, structural rearrangements, or copy number alterations (copy-number <1 or >9) was analyzed by DAVID (Huang *et. al.* (2009)). Gene Ontology terms (GO-terms) were then filtered to retain only those with a p-value <= 0.05 and plotted in the left panel where color indicates the genome (HeLa or control) and shading represents significance. The right panel shows, for each term, the number of genomes with significant enrichment for protein-altering variants in the associated genes. With the exception of the “Startle response” GO-term, all of the terms in HeLa with a p-value <= 0.01 occur in at least one of the control genomes.

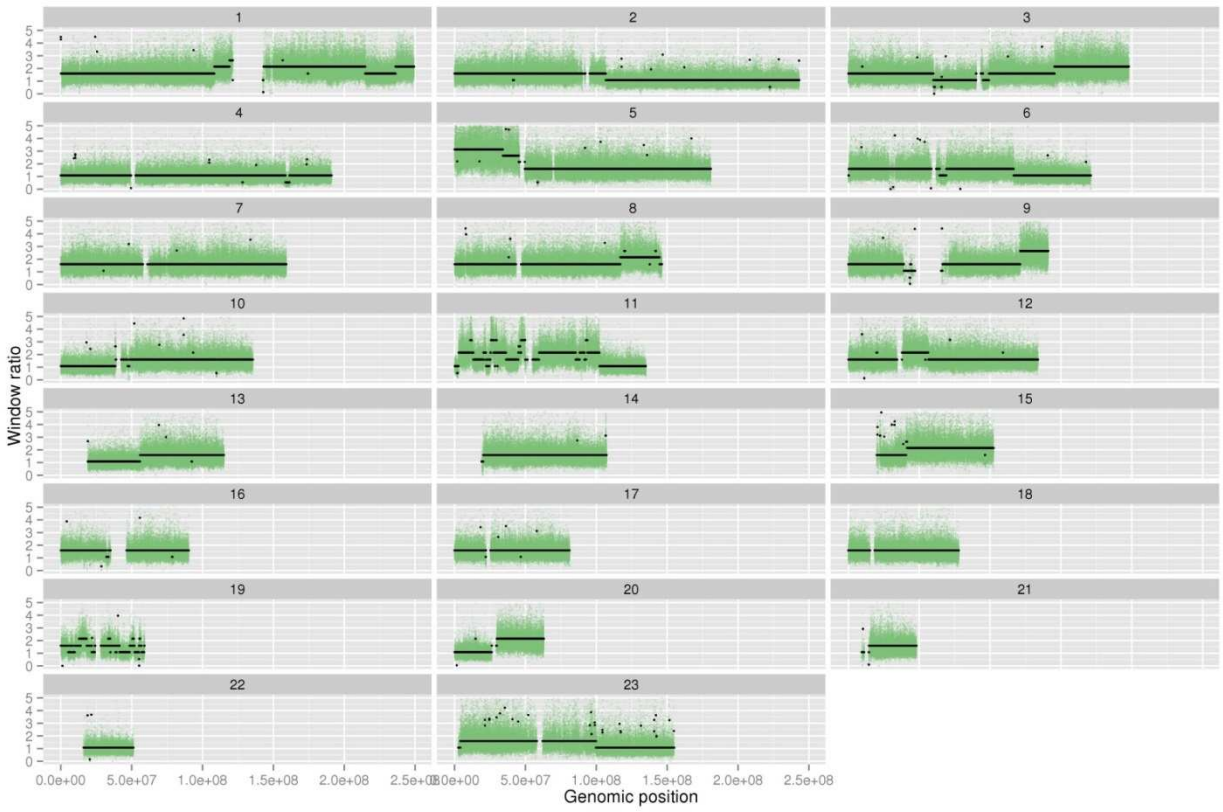


Figure C.3.4 | Hela CCL-2 high resolution copy number calls.

Copy number ratios versus control genomes are plotted within high-resolution SUNK windows (green dots, each window size ~1.5 kb), with predicted copy number state overlaid (black dots).

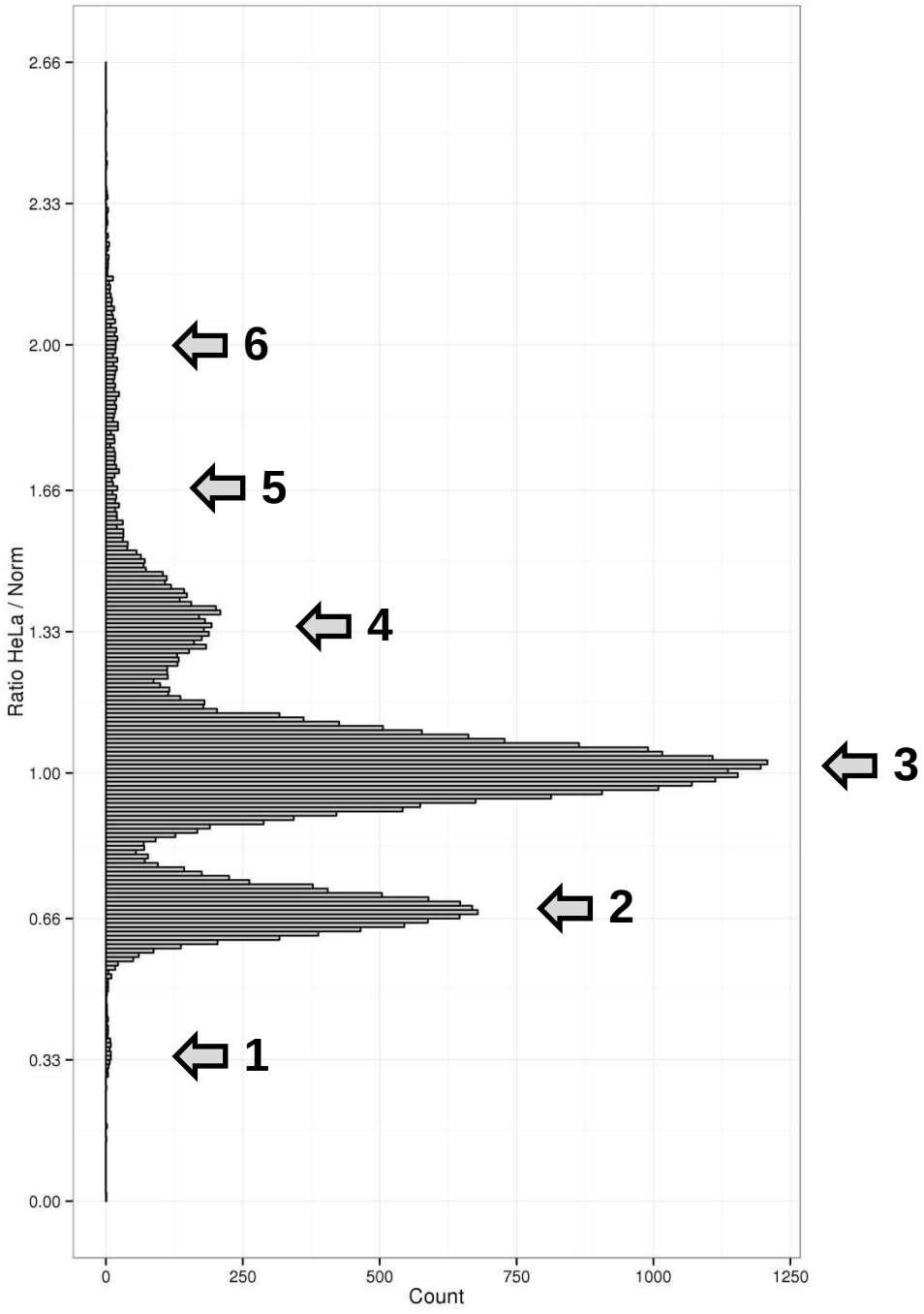


Figure C.3.5 | HeLa over GC-matched control ratio histogram.

SUNK window (500 unique 30mer) resolution ratio scores plotted as a histogram. Distinct peaks are observed at approximately 0.33, 0.66, 1.0, 1.33, consistent with an approximately triploid numerator sample (HeLa) over a diploid denominator sample (GC-matched control). Inferred copy numbers are indicated by arrows.

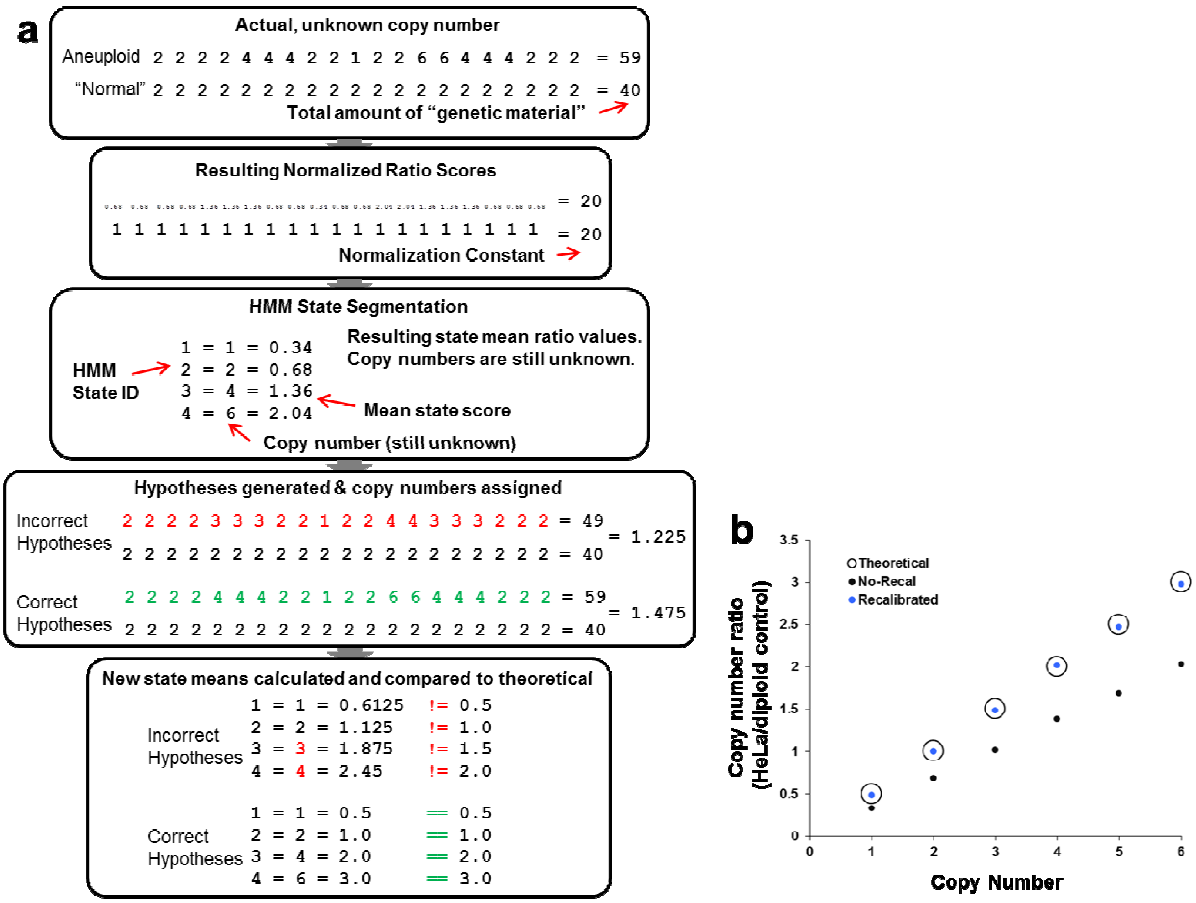


Figure C.3.6 | Copy-number recalibration strategy.
a. Schematic of steps involved in the recalibration process. In order to adjust for differences in total read depth between genomes, window scores are normalized to a constant. Ratios are then taken between a G+C profile matched normal control and states are segmented using an HMM. Resulting ratios are not directly relatable to absolute copy number when the two genomes' chromosomal complements are of unequal size (e.g., one is triploid and the other diploid). Assignments of copy numbers to HMM states ("hypotheses") are exhaustively generated; windowed copy number values then summed to generate a "genetic material ratio" which is used as the normalization constant. The mean across windows from each HMM state is recalculated, and ratios to the diploid control genome are taken, after which the per-state . The hypothesis which minimizes the mean difference between observed and expected per-state ratios is chosen. **b.** HeLa copy number state scores are shown before and after recalibration (black and blue, respectively), with theoretical values shown as open circles.

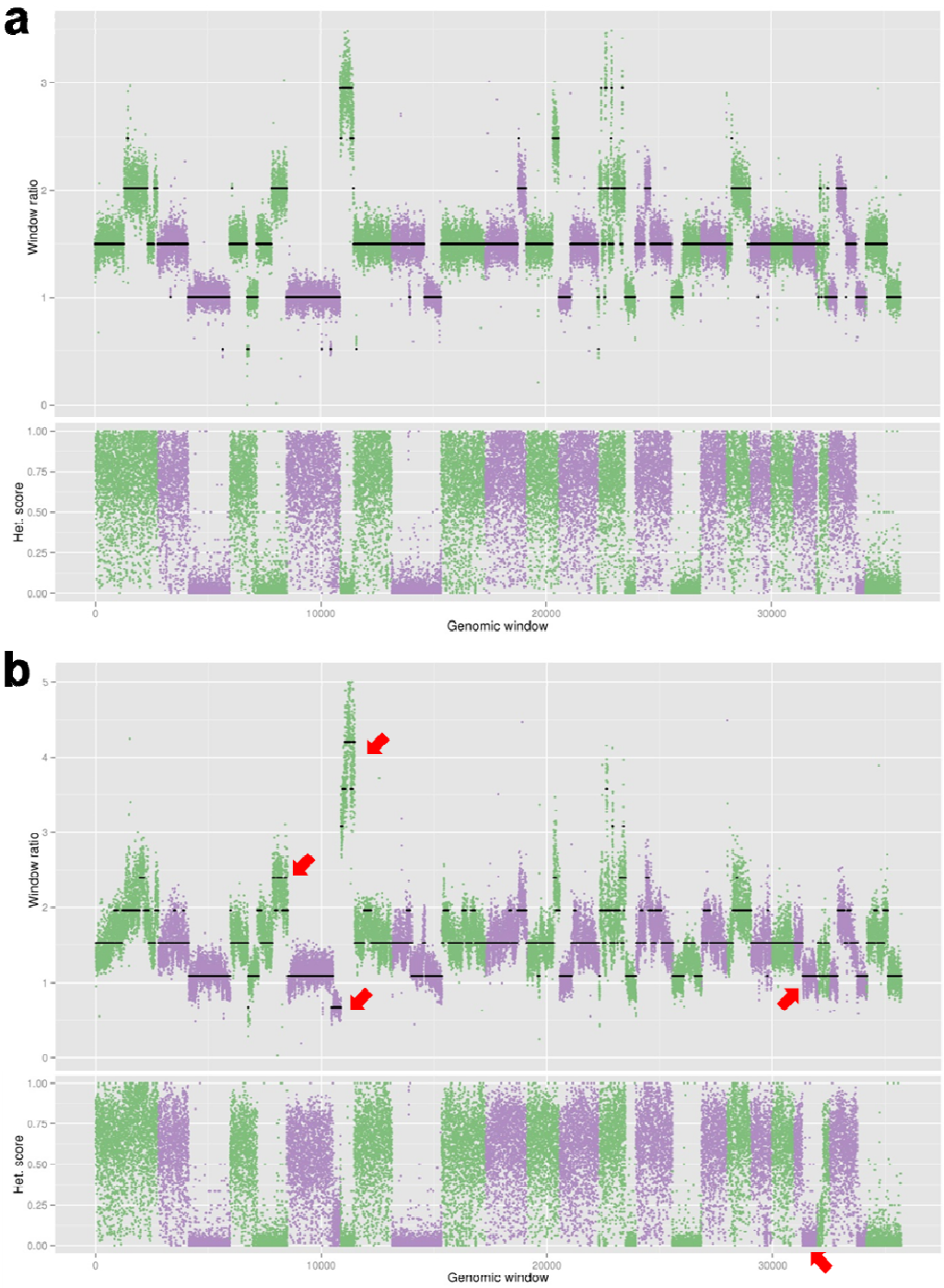


Figure C.3.7 | HeLa CCL-2 and S3 copy number and LOH profiles.
a, Top - Low resolution SUNK window ratio scores (green or purple points) and copy number state calls (black lines) for HeLa CCL-2. Bottom – Loss of heterozygosity scores measured by the fraction of heterozygous variants in each window. **b,** As in **a.** but for HeLa S3. Red arrows indicate notable changes in copy number or LoH.

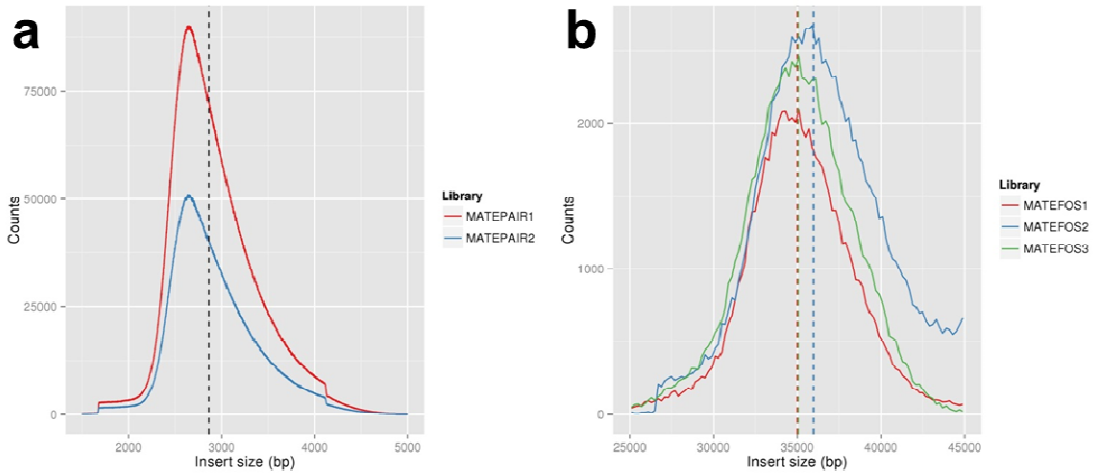


Figure C.3.8 | Mate pair insert size distributions.
a, Insert size distributions of concordant pairs for the two "3 kb" mate-pair libraries constructed using *in vitro* circularization. **b,** Insert size distributions of concordant pairs for the three "40 kb" mate-pair libraries constructed using fosmid cloning.

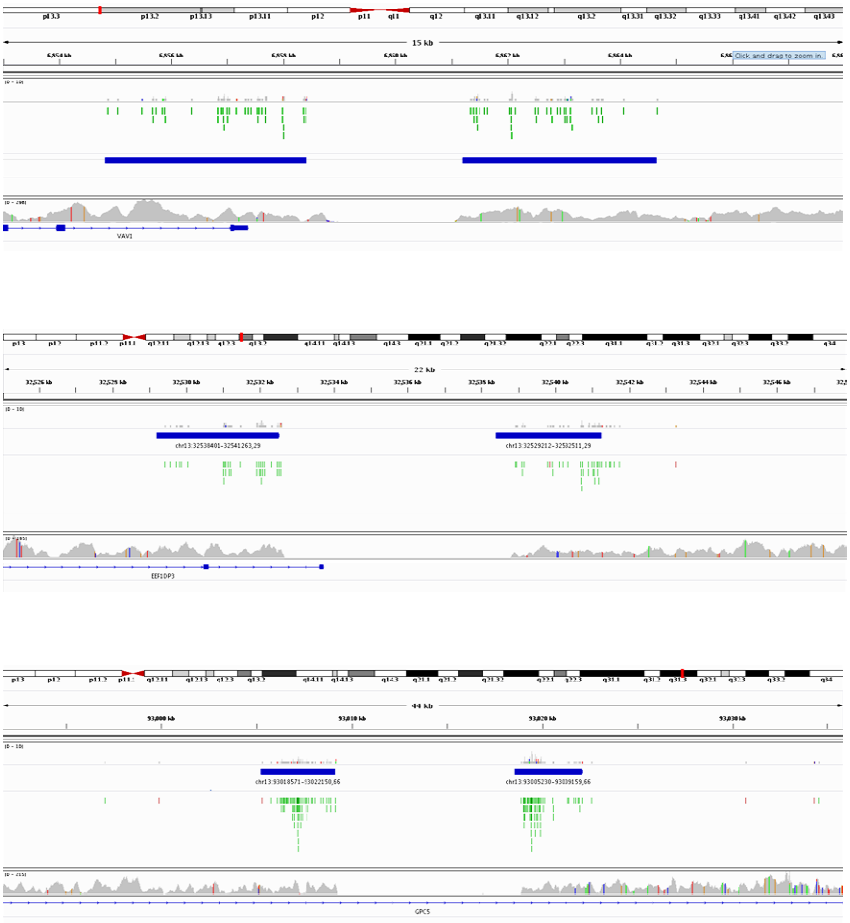


Figure C.3.9 | Examples of deletions in HeLa CCL-2
Three examples of deletions called using a sliding window approach shown in the IGV genome browser. Blue bars denote regions of coverage from supporting 3 kb mate-paired reads (green ticks). Shotgun sequence coverage (gray bars) are plotted beneath each event.

Figure C.3.10 | Examples of inter-chromosomal rearrangements in HeLa CCL-2.
Two examples of inter-chromosomal rearrangements detected by a sliding window approach from discordantly-mapping 3 kb mate-pair reads. The upper example is one of the rearrangements within marker chromosome M14.

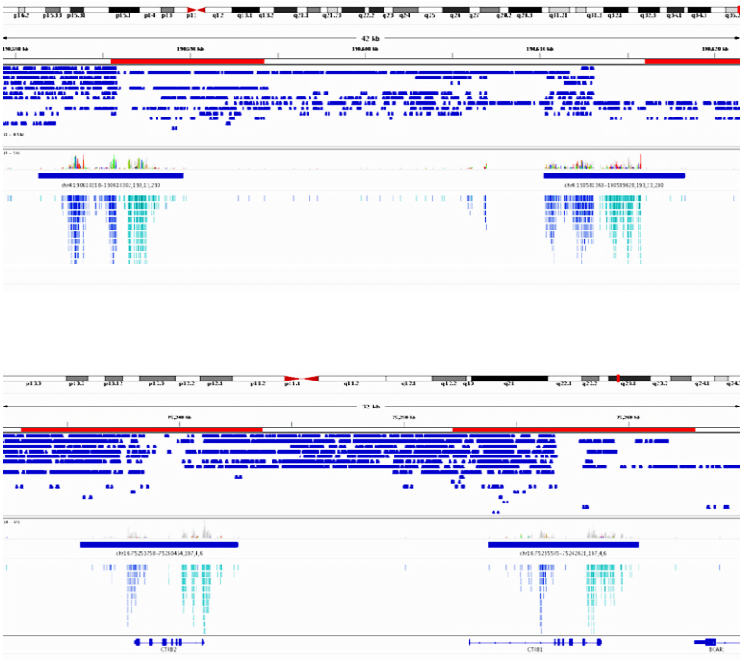


Figure C.3.11 | Called inversion examples in HeLa CCL-2.
Two examples of inversions detected by a sliding window approach from discordantly-mapping 3 kb mate-pair reads. Both inversions are supported by fosmid sequence coverage profiles (blue tracks

shown below chromosome ideograms), with overlapping clones showing discontinuous patterns of coverage near each inversion breakpoint.

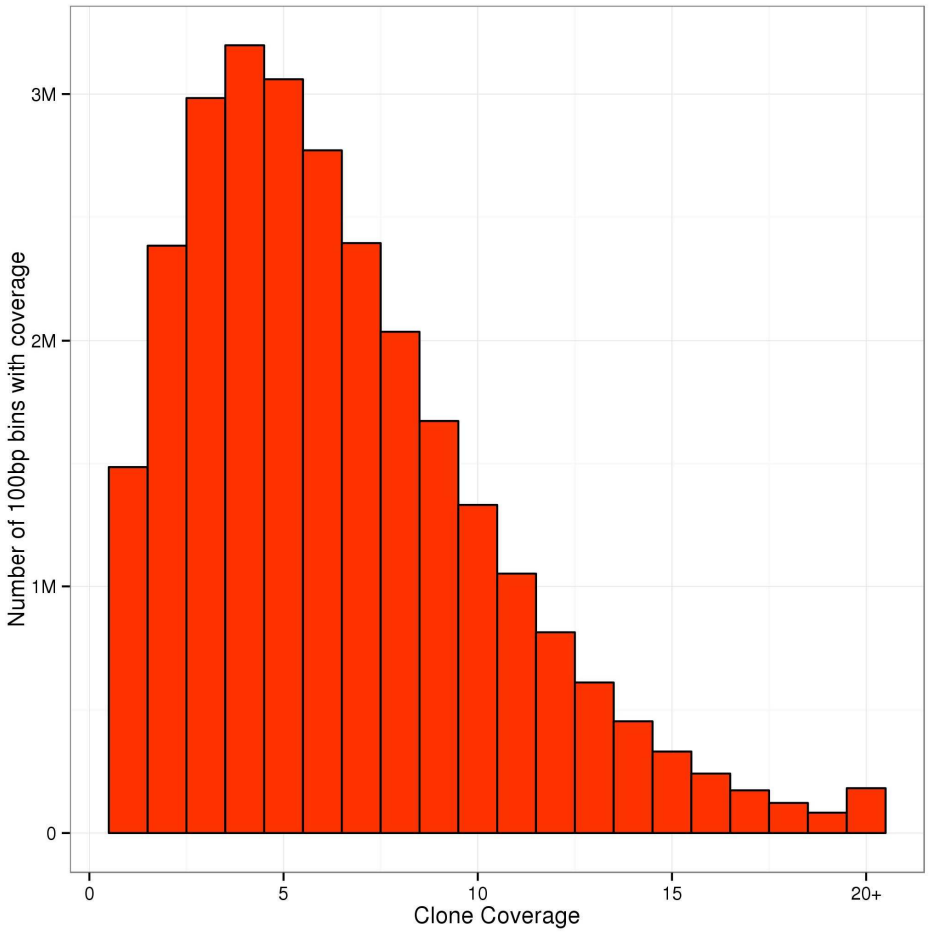


Figure C.3.12 | Histogram of clone coverage.

Histogram of the physical coverage by fosmid clone inserts. Overall, 3.5% of the genome is not covered (coverage=0, excluding chromosome Y and assembly gaps).

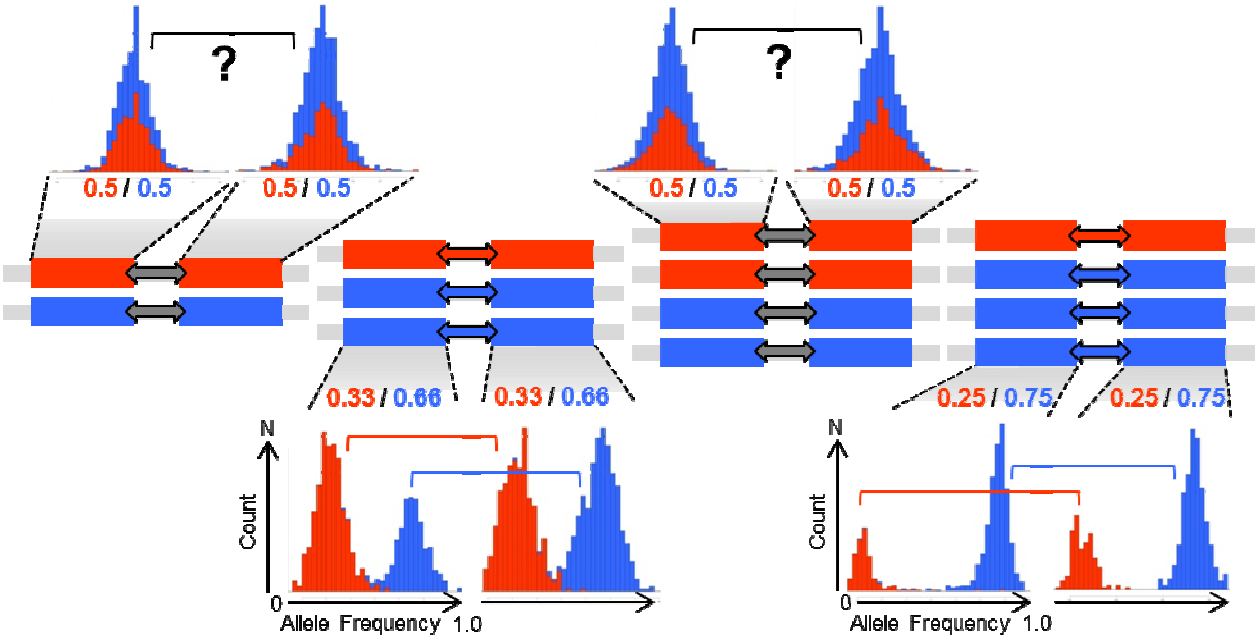


Figure C.3.13 | Schematic of haplotype scaffolding approach using allele imbalance.

Consecutive haplotype blocks in regions containing imbalanced copy numbers (green boxes) between haplotypes can be merged using an HMM to form a haplotype scaffold based on the allele frequencies of phased variants within the blocks (histograms of with (haplotype A) and blue (haplotype B) distributions representing allele frequencies for the respective haplotypes). For haplotype blocks in regions of imbalanced haplotype these histograms are distinct (histograms on bottom of figure), whereas haplotype blocks in regions of even copy number overlap and can not be distinguished (histograms at the top of the figure).

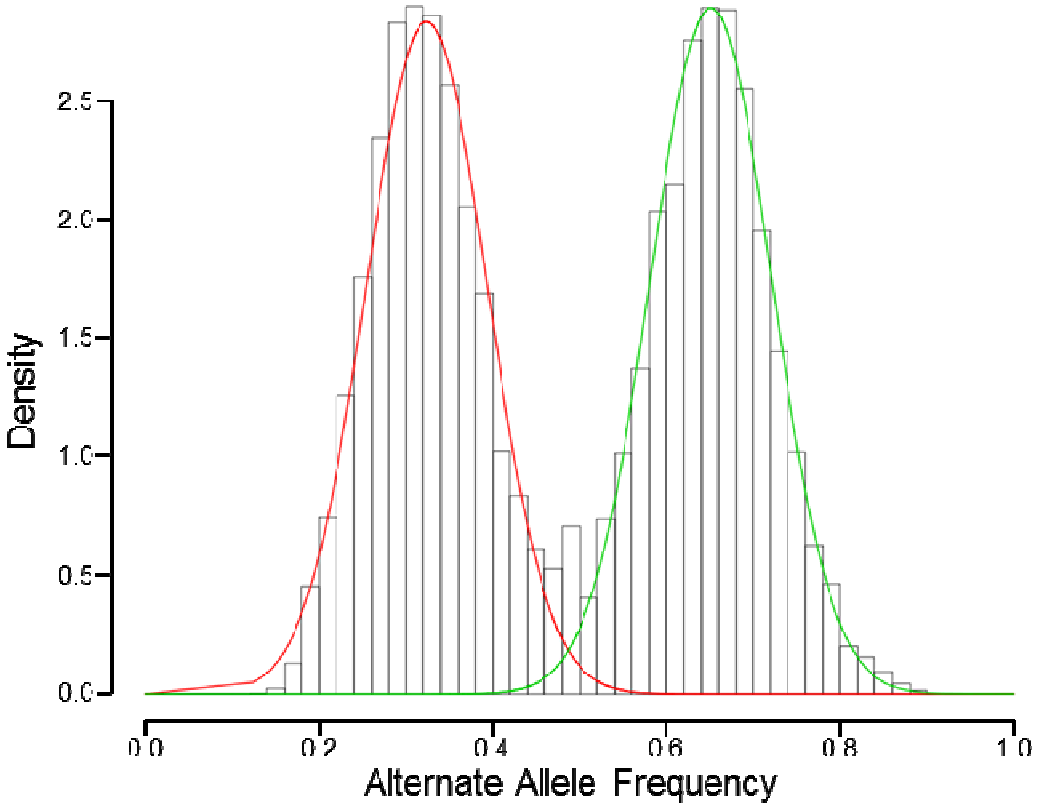


Figure C.3.14 | Gaussian mixture model of AAFs in non-LOH copy number 3 regions.

A histogram of alternate allele frequencies among shotgun reads is shown for all heterozygous variants present in regions of copy number 3 in which one haplotype is at copy number 2 and the other at copy number 1. A two-component Gaussian mixture model was fit to this distribution, and the centers of each component (red and green lines) were at 0.324 and 0.651, near the expected values of 1/3 and 2/3.

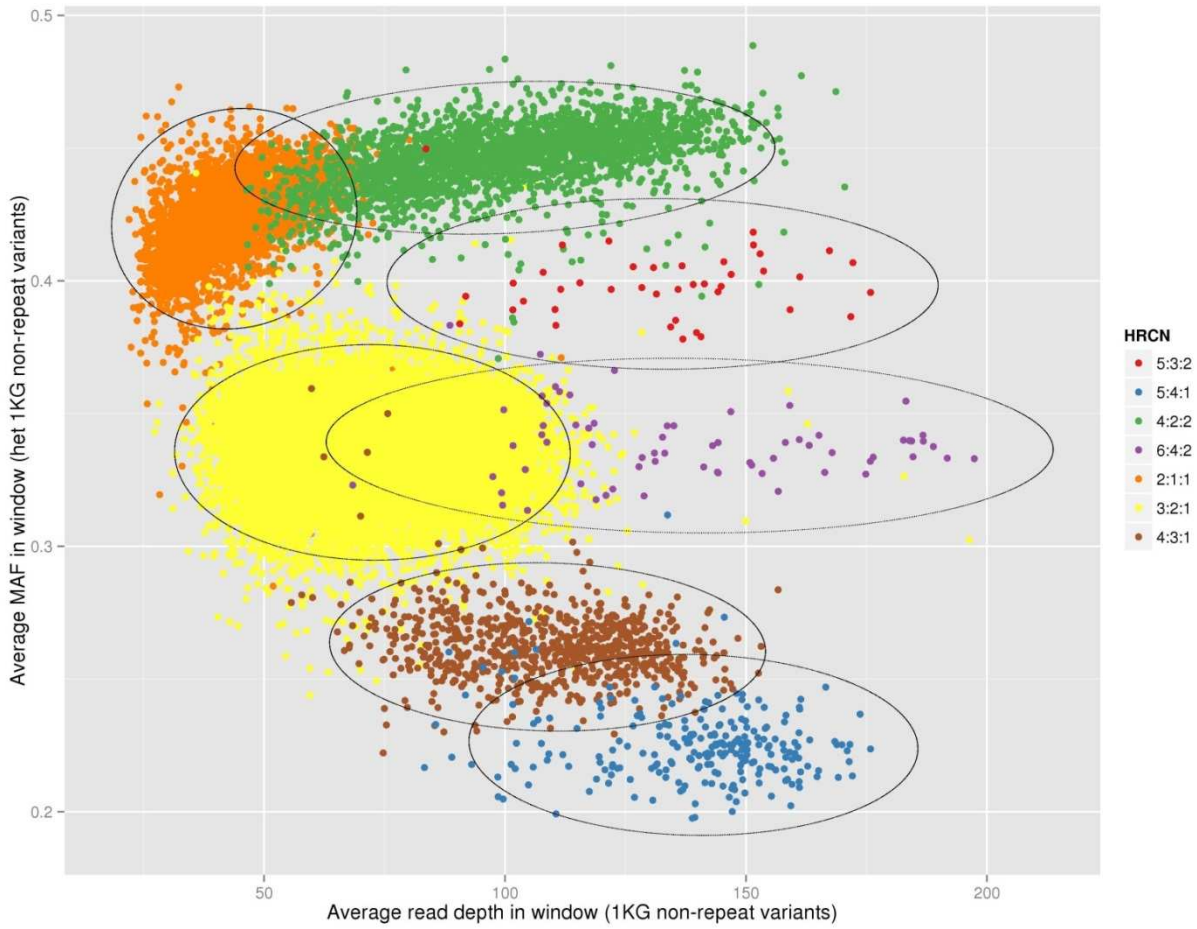


Figure C.3.15 | HeLa allele balance by read depth for HRCN regions.

For each low resolution SUNK window (~77 kb), the average minor allele frequency of all heterozygous variants was plotted against those sites' average read depth. Each point was shaded by the window's predicted HRCN (total copy number : haplotype A copy number : haplotype B copy number). Overlaid ellipses represent 95% confidence intervals for each HRCN grouping.

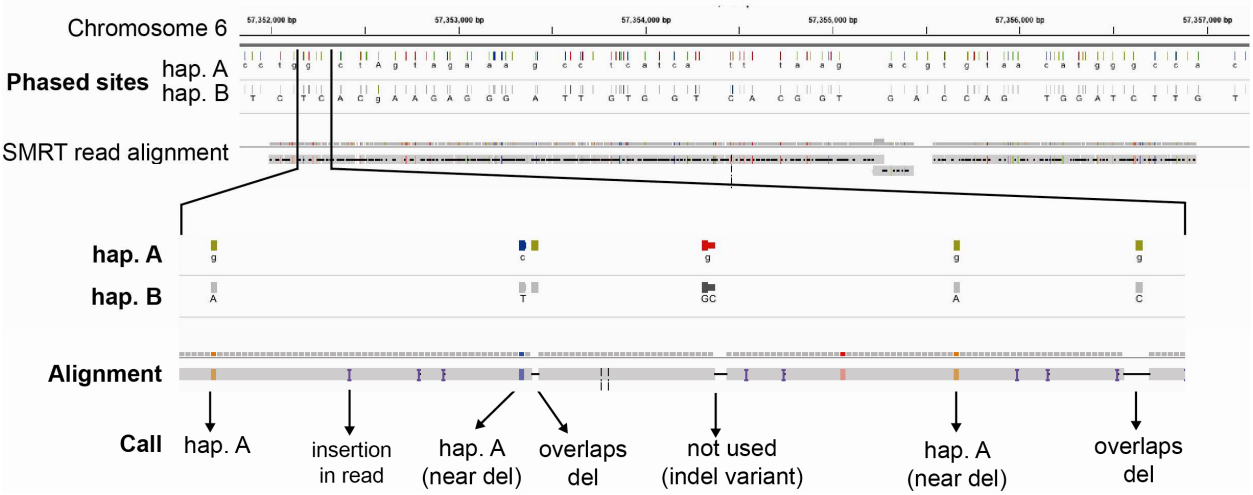


Figure C.3.16 | Long read haplotype validation.

Haplotype phase validation by single molecule long-read sequencing. An example alignment from one read spanning 4.96 kbp is shown. **a.** Upper track: phased variants in HeLa CCL-2 are shown for each inherited haplotype (A and B), with gray ticks indicating the reference allele and colors representing the alternate allele. Lower track: the aligned read spans 98 phased heterozygous sites, of which 19 sites are more than 10 bp from the nearest alignment indel. Of those, all 19 sites match the allele predicted on haplotype C. **b.** Detail showing aligned positions matching haplotype A or rejected due to overlapping or nearby indel errors.

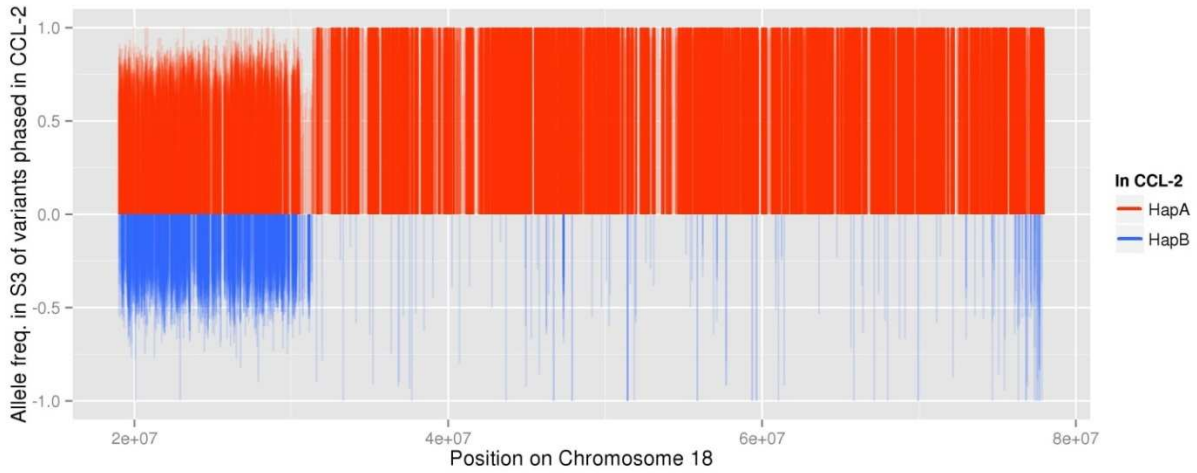


Figure C.3.17 | Allelic state across LOH event specific to HeLa S3.

Allele frequencies among HeLa S3 shotgun reads are shown for all heterozygous and phased variants from HeLa CCL-2, across 78.1 Mbp of chromosome 18. Allele frequencies in S3 are plotted on the y-axis, with points' direction and color indicating whether each CCL-2 allele is phased on haplotype A (red, upward) or haplotype B (blue, downward). In HeLa CCL-2, chromosome 18 is triploid without LOH, but in HeLa S3 it is observed to have a large (47.3 Mbp) distal region that is diploid with LOH. Nearly all (99.7%) of the variants with allele balance >0.9 within this region (in S3) correspond to haplotype A from HeLa CCL-2.

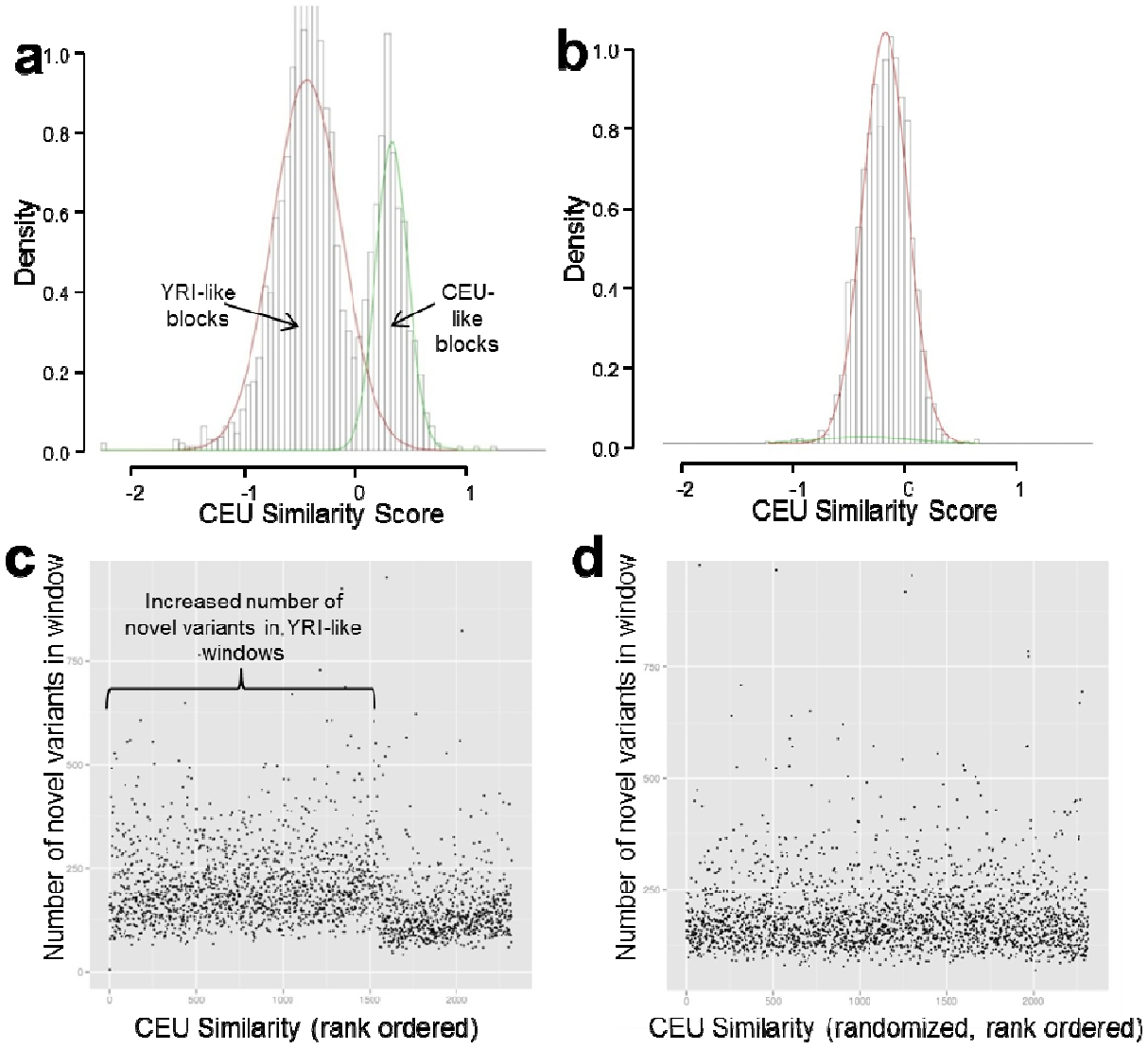


Figure C.3.18 | Population-based haplotype analysis.
a. Histogram of windowed scores based based upon phased sites' popluation allele frequencies in CEU vs YRI individuals (from the 1000 Genomes Project). Red and green lines indicate density from a two-component mixture model fit. **b**. Randomization test. Histogram of windowed scores, identical to **a**, except that the phase is randomized between each successive pair of 1000 Genomes variants. **c.** Counts of novel variants (non-1000 Genomes Project) for windows ranked as in **a.** (windows with more CEU-like alleles to the left, more YRI-like alleles to the right). More highly YRI-like haplotype blocks on average contain more novel variants. **d.** Randomization test. Counts of novel variants in each window, identical to **c**, except that the phase is randomized as in **b**.

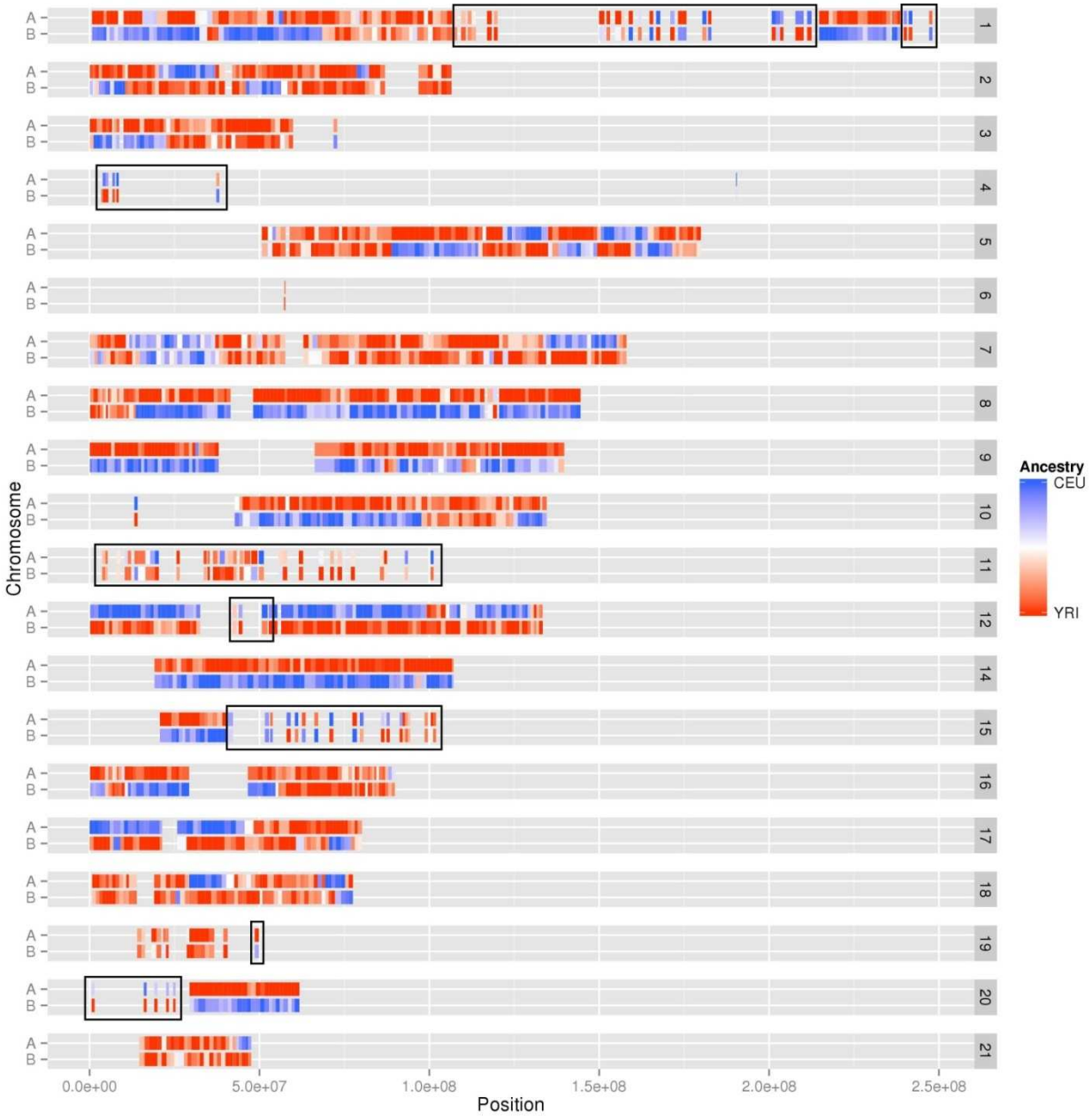
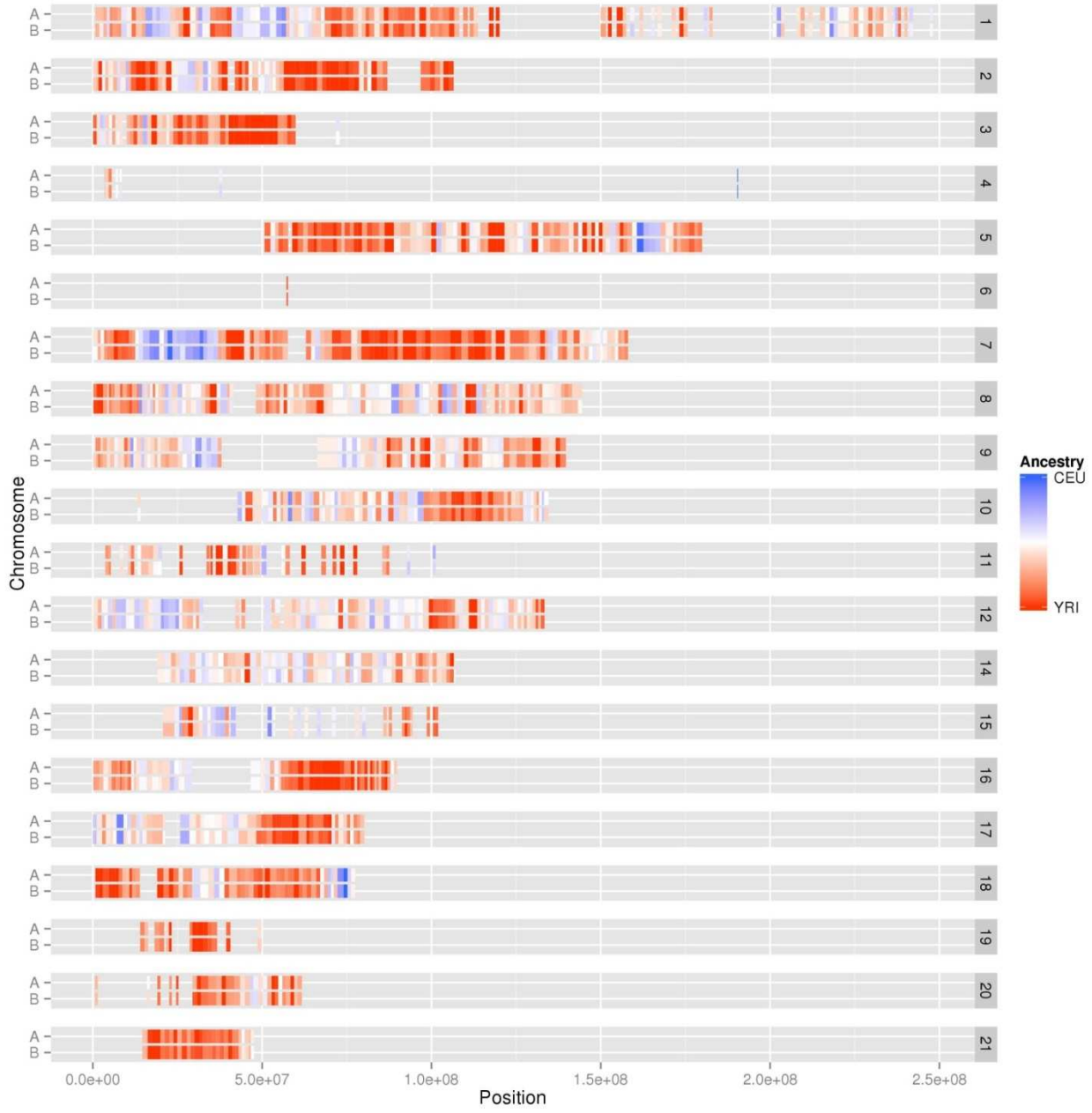


Figure C.3.19 | Haplotype-based local inference of genetic ancestry.
a. Predicted genetic ancestry is shown for haplotype windows, using scores of allele frequency to CEU or YRI populations and colored by the ancestry similarity (Blue = CEU, Red = YRI). Windows in LOH regions, in haplotype scaffolds with insufficient numbers of phased variants (fewer than 1,000 variants 1000 Genomes Project variants), are not shown. Regions of balanced copy number shown by black boxes were excluded because haplotype imbalance could not be used to create long scaffolds. **b.** Randomization test. Windows are painted as in **a**, except that the phase is randomized between each successive pair of 1000 Genomes variants.


A
B
1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
17
18
19
20
21
Chromosome
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
Position
Ancestry
CEU
YRI

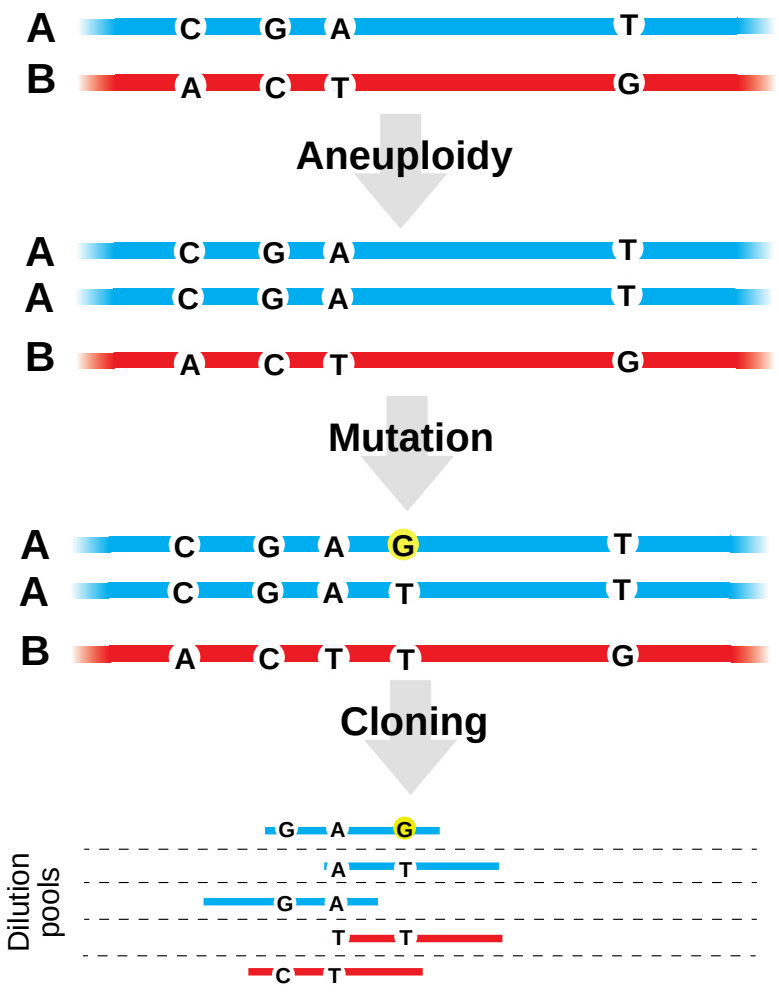


Figure C.3.20 | Post-aneuploidy mutation analysis.

Schematic of validation process for somatic, post-aneuploidy mutations by large insert clone pool sequencing. Mutations arising after duplication of a germline haplotype (blue) are confirmed by the presence of both the mutant allele (yellow, “G”) as well as the reference allele (T) in separate clones derived from the duplicated haplotype.

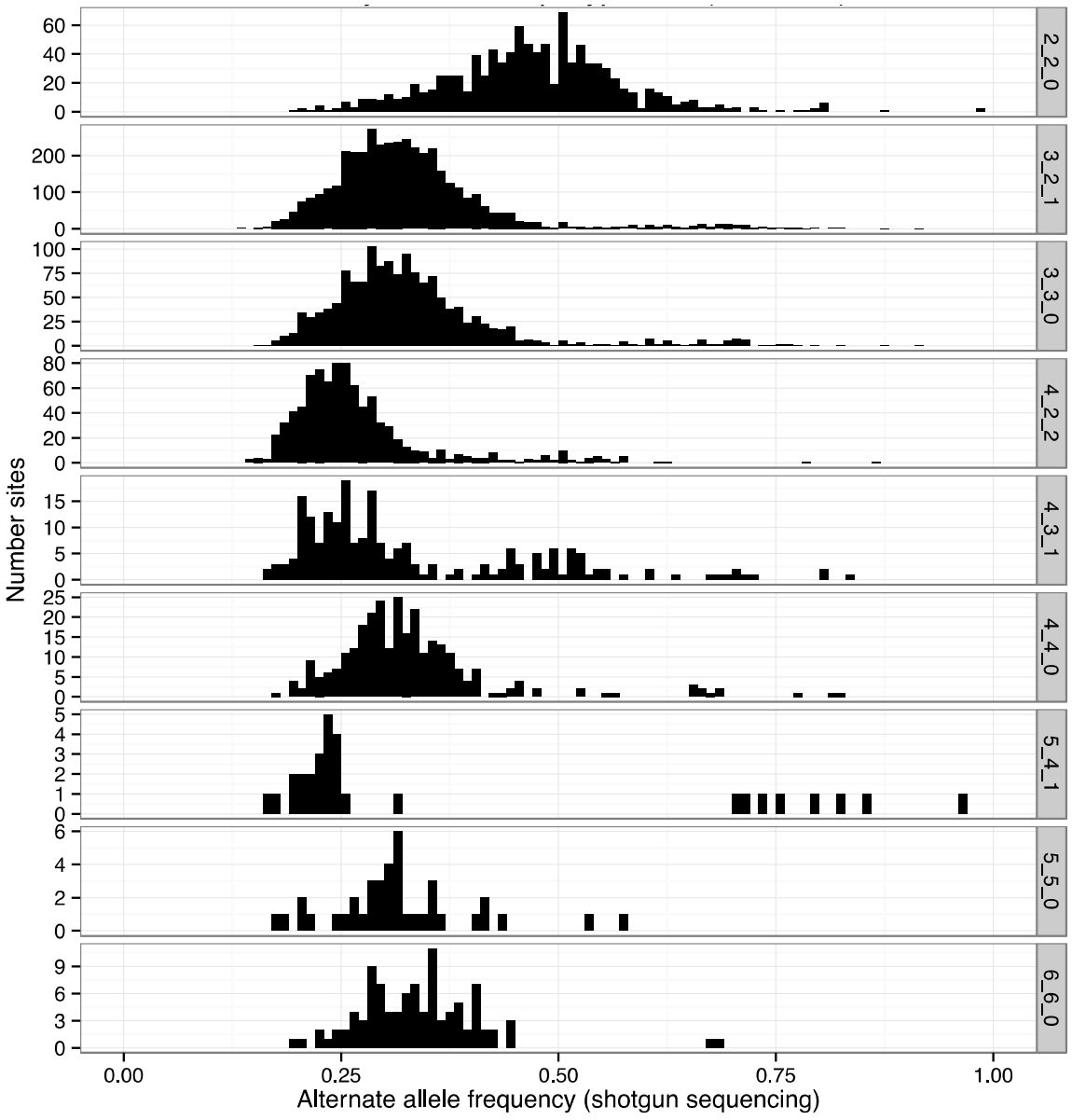


Figure C.3.21 | Somatic mutation allele frequencies.

Histograms of allele frequency within shotgun data of clone-validated somatic mutations, split by haplotype-resolved copy number (HRCN) state. Regions with HRCN of 5:3:2 and 6:4:2 were omitted because there were few sites (each <= 10).

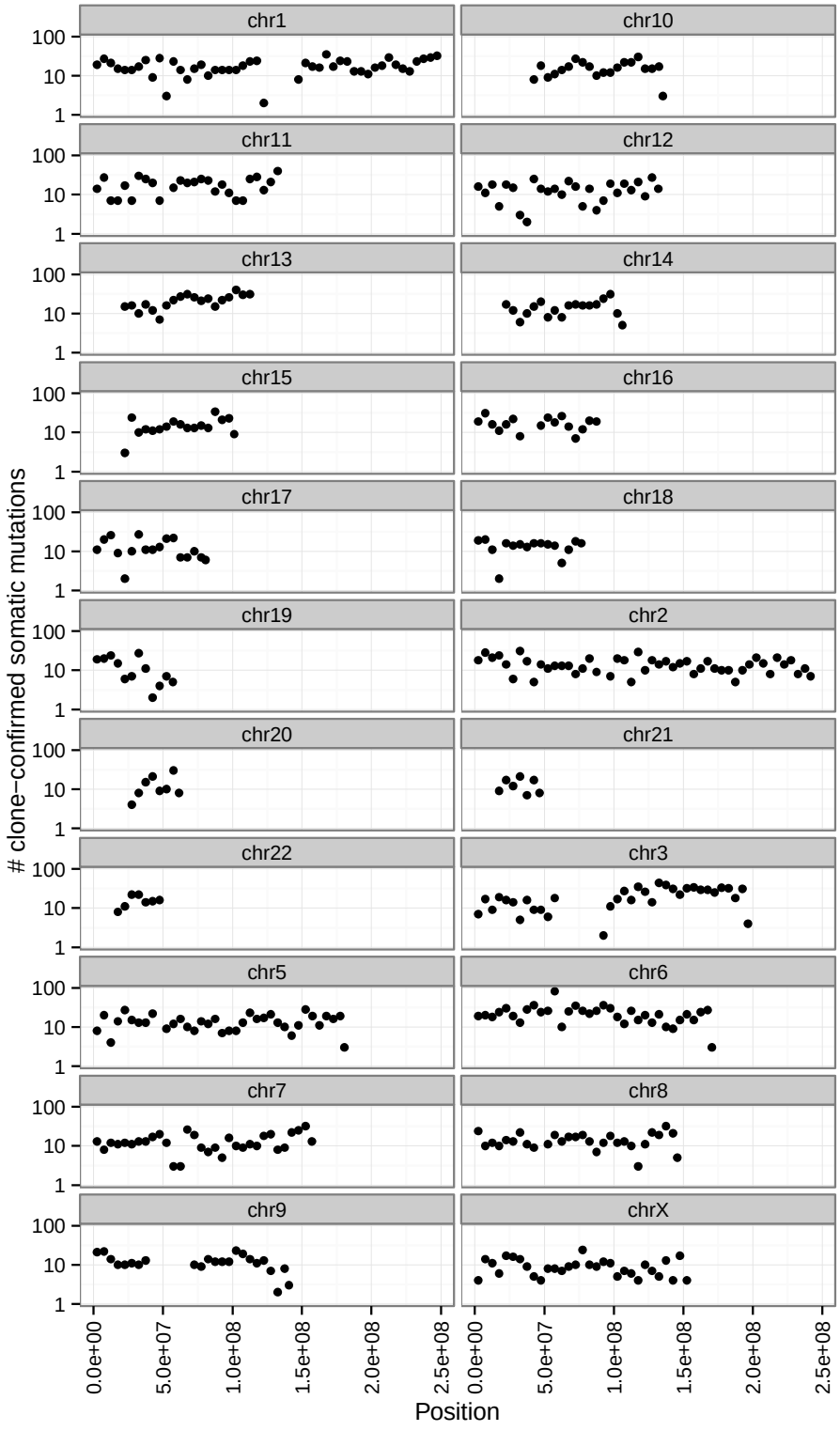


Figure C.3.22 | Somatic mutation counts.

Count of somatic mutations per 5 Mbp window along each chromosome.

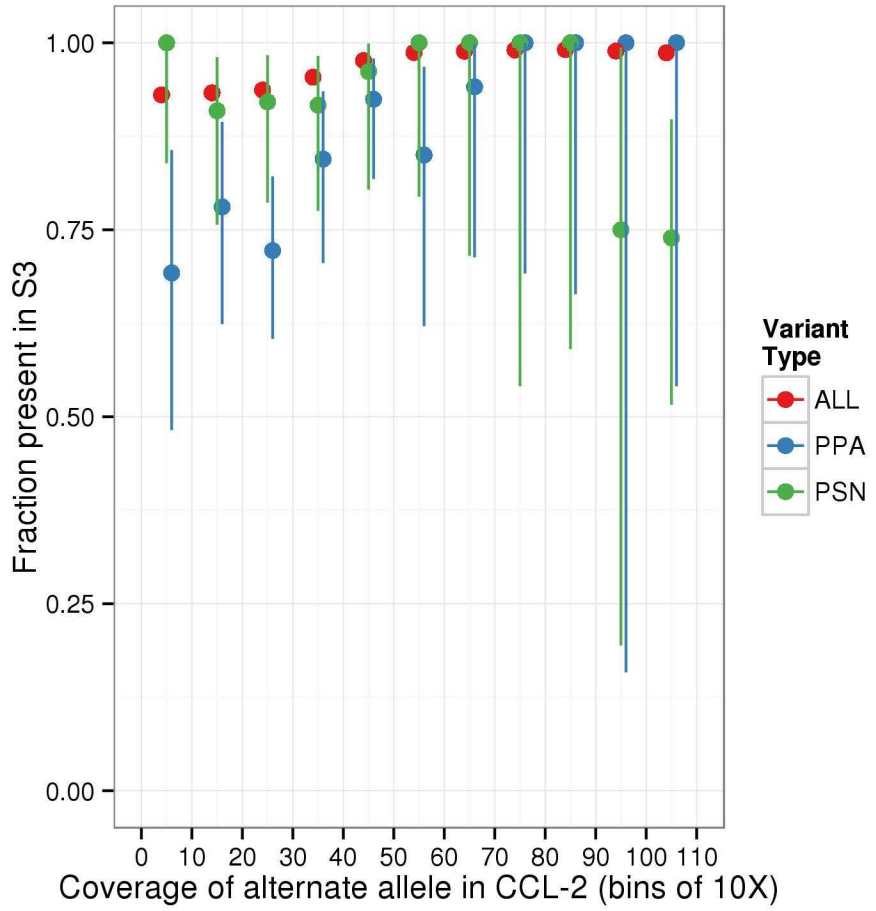


Figure C.3.23 | Private alleles shared between HeLa CCL-2 and S3.

The fraction of private SNVs (not found in the 1000 Genomes Project) from HeLa CCL-2 that are also observed in S3 is shown, binned by the number of reads supporting the alternate allele in CCL-2. The fraction of shared alleles is shown for different categories of sites: all private sites in CCL-2 (Red, "ALL"), private protein-altering variants in CCL-2 (Blue, "PPA") and private coding synonymous variants in CCL-2 (Green, "PSN"). Variant alleles supported by >100 reads in CCL-2 were grouped into the "100+" bin.

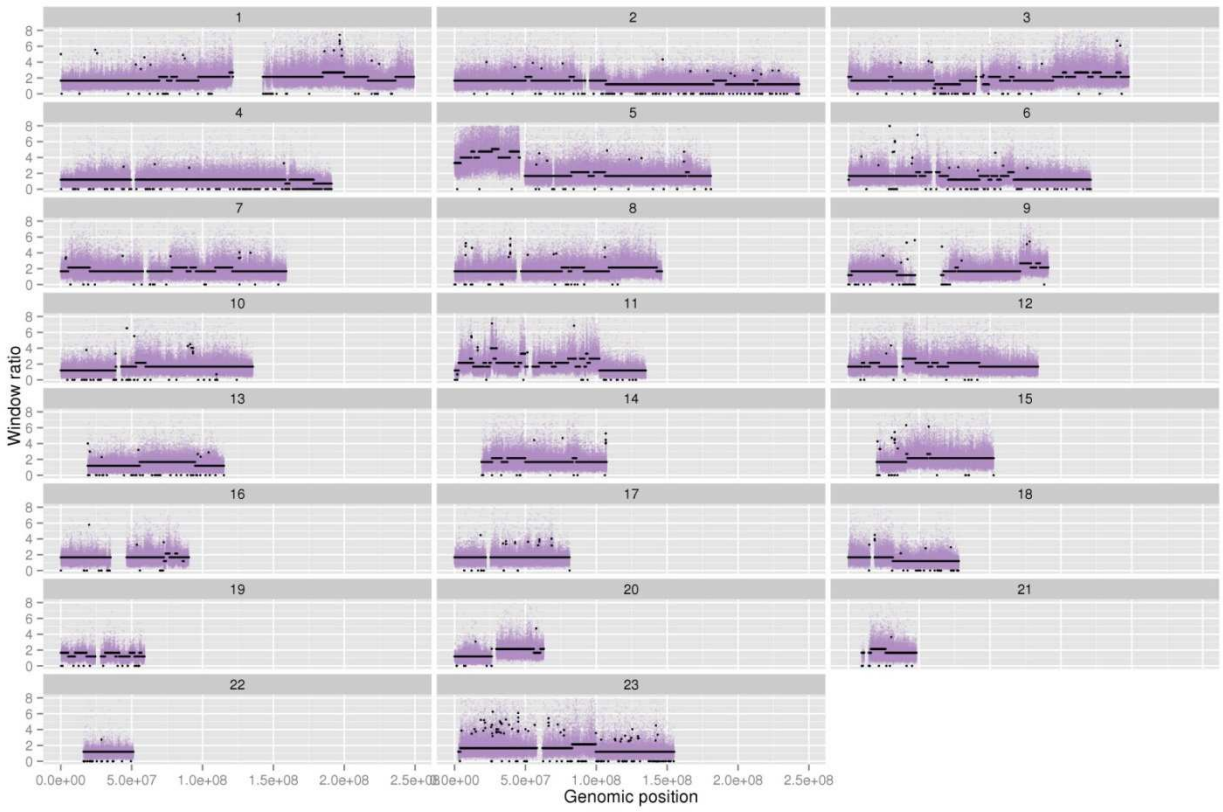


Figure C.3.24 | Hela S3 high resolution copy number calls.

Copy number ratios versus control genomes are plotted within high-resolution SUNK windows (green dots, each window size ~1.5 kb), with predicted copy number state overlaid (black dots).

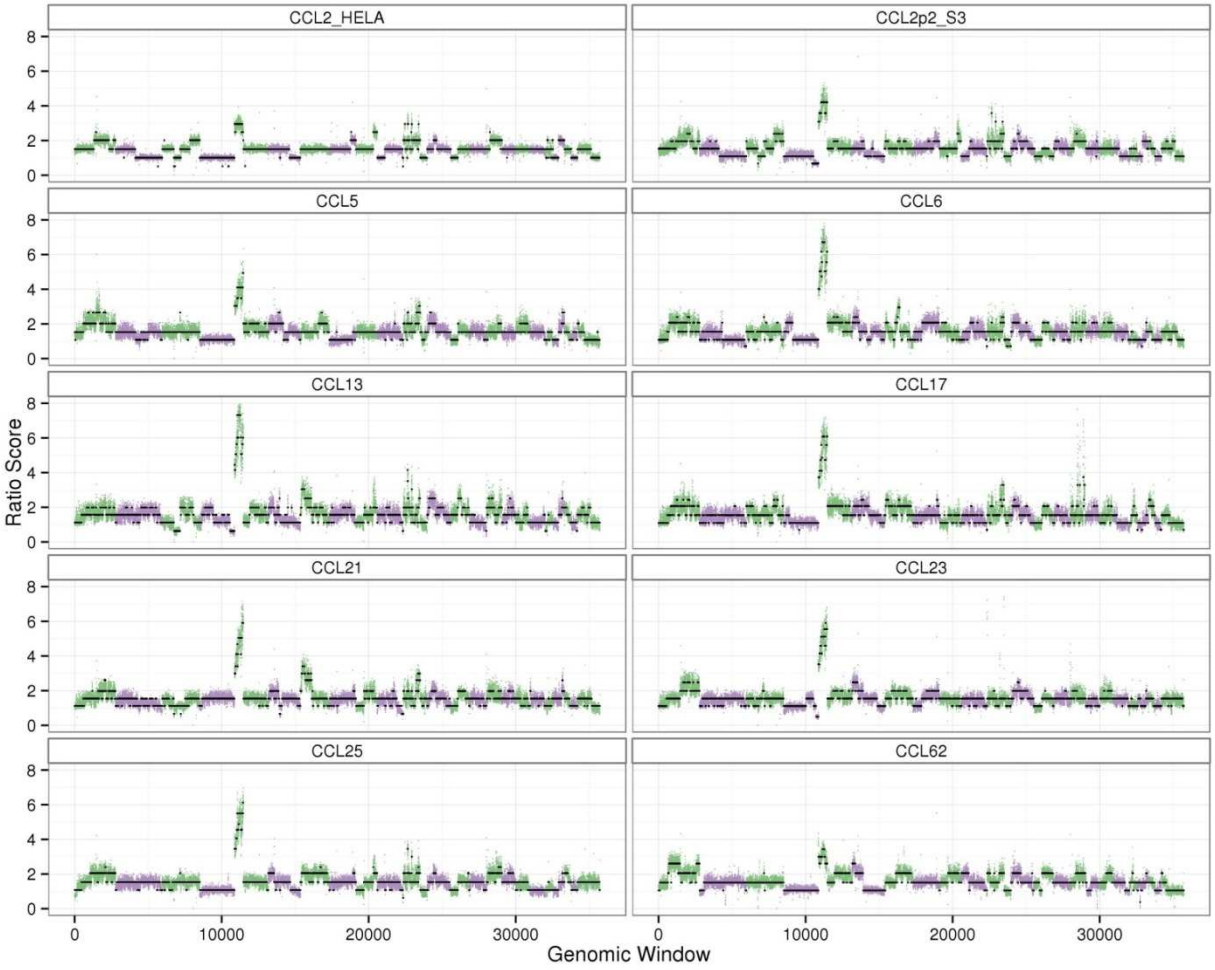


Figure C.3.25 | Copy number profiles for 10 HeLa strains.

Copy number ratios versus control genomes are plotted within high-resolution SUNK windows for HeLa CCL-2, HeLa S3, and eight additional HeLa strains (green and purple dots, alternating by chromosome, window contains 500 unique 30mers), with predicted copy number state overlaid (black dots).

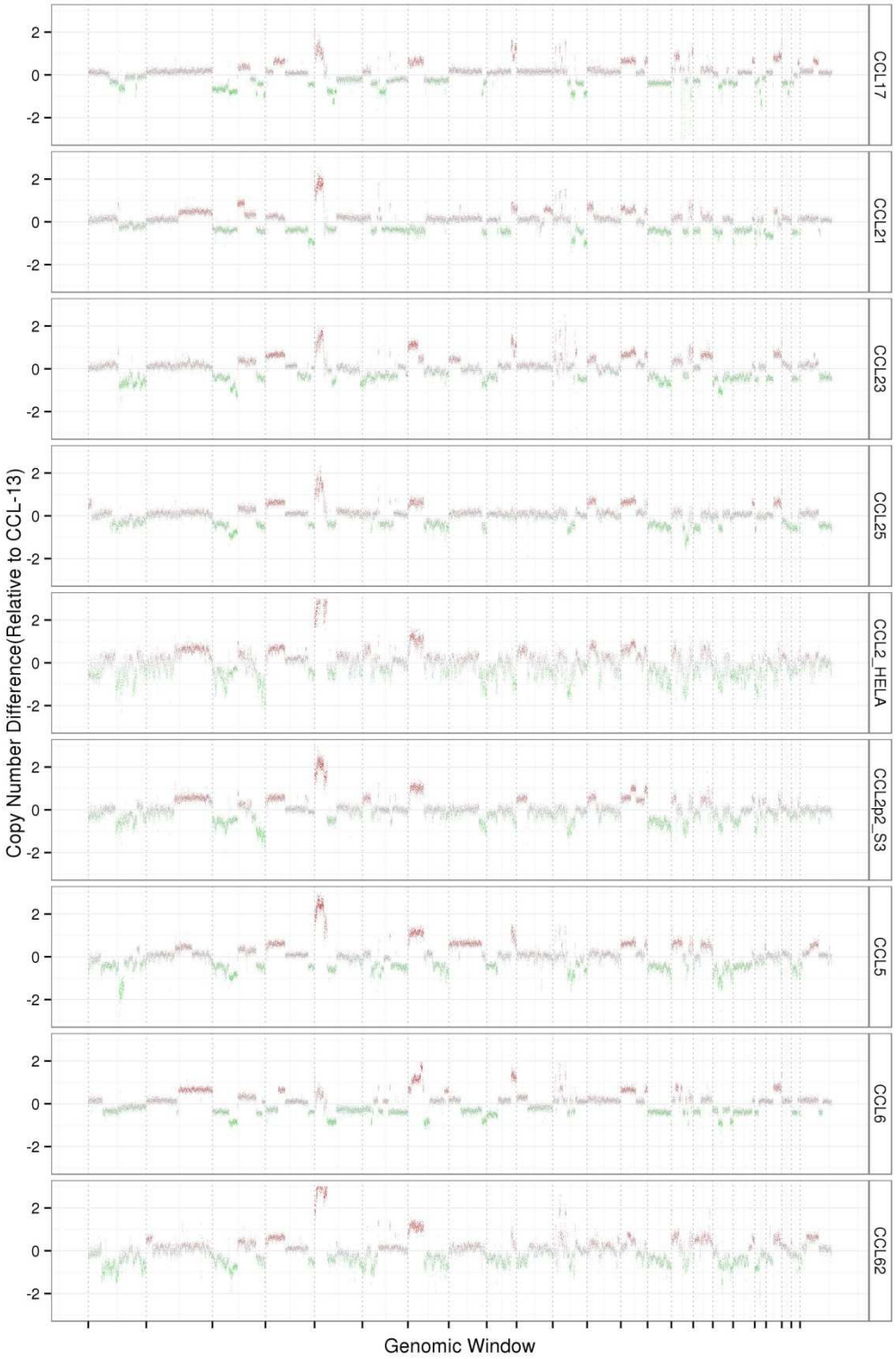


Figure C.3.26 | Comparison of read depth profiles in HeLa strains.
Copy number differences across low-resolution SUNK windows relative to HeLa CCL-13 were plotted for HeLa CCL-2, S3, and 7 additional strains. Note: Increased values indicate increased copy number in CCL-13 compared to alternate strain.

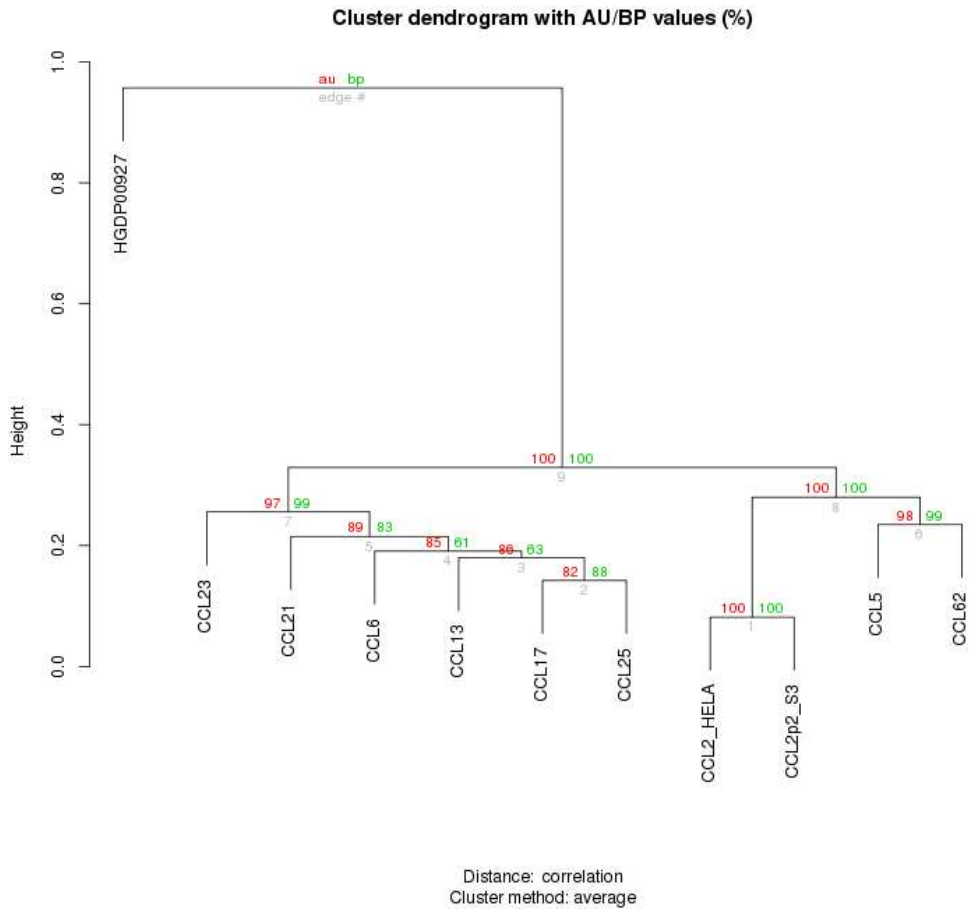


Figure C.3.27 | Clustergram of 10 HeLa strains based on copy number profile similiarity.

Copy number scores were averaged within large windows (~1 Mbp) for 10 HeLa strains as well as an outgroup control genome (HGDP00927). Scores were clustered in (R package 'pvclust') with 1000 bootstrap iterations. “au” values correspond to “Approximately Unbiased” scoring that is computed by multiscale bootstrap resampling while the “bp” value corresponds to “Bootstrap Probability”, or standard bootstrap scoring. Due to batch differences in library preparation, comparison with HeLa CCL-2, HeLa S3 and the HGDP outgroup is much less reliable. It is important to note that this dendrogram is not necessarily the actual phylogeny and simply represents the similarity between marker chromosome / copy number subsets for the individual strains.

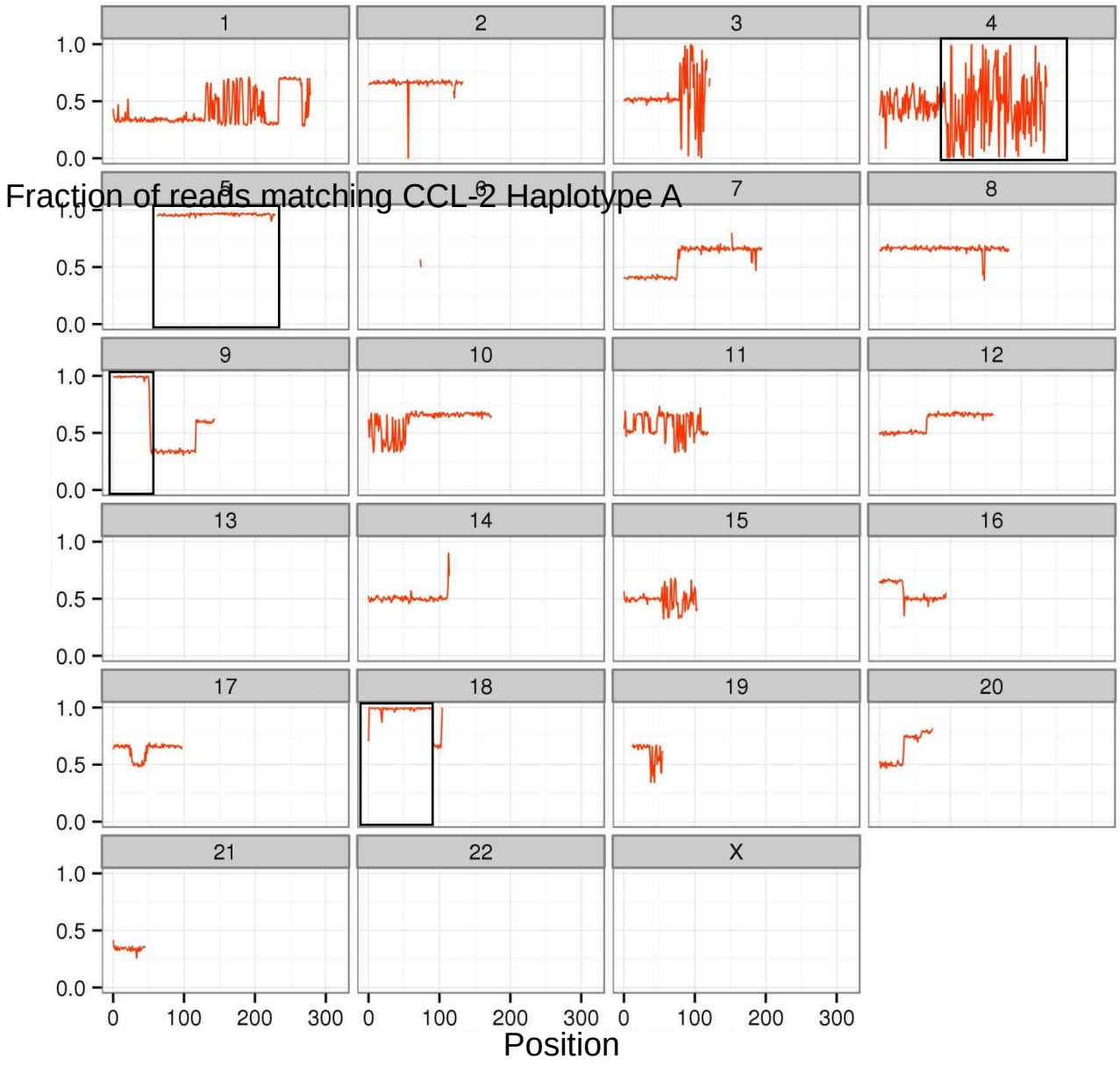


Figure C.3.28 | Regions of LOH in HeLa CCL-13 by comparison to CCL-2 haplotypes. Shown in windows (mean ~800 kbp) across each chromosome are the fraction of reads matching the allele phased to haplotype A in HeLa CCL-2. LOH in CCL-13 (but not CCL-2) manifests as long stretches where shotgun reads from CCL-13 (mean depth 4.0X) exclusively match CCL-2 haplotype A (y value = 1) or haplotype B (y value=0). A total of NNN Mbp of LOH regions were detected in CCL-13 (highlighted by shaded bars). Regions lacking haplotype scaffolds in CCL-2 (e.g., in LOH or in regions of balanced copy number in CCL-2) were omitted. Black boxes indicate predicted regions of LOH.

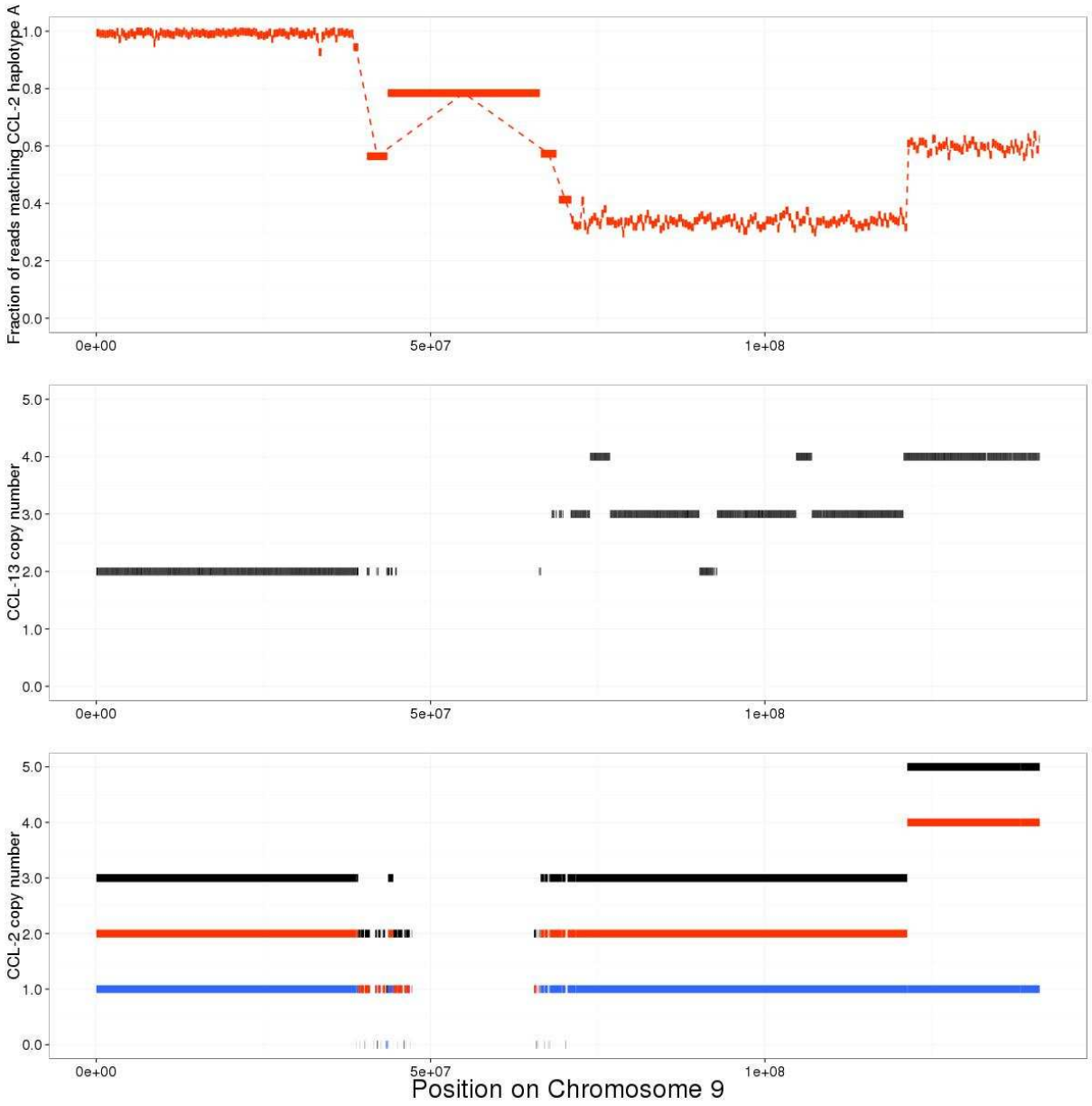


Figure C.3.29 | Copy-number loss and LOH on chromosome 9 in HeLa CCL-13.

a. LOH on chromosome 9 in HeLa-CCL13, as detected by a shift towards CCL-2 haplotype A alleles was accompanied by reduction of copy number to 2 in CCL-13 in the affected region shown in **b.** relative to copy in HeLa CCL-2 shown in **c.** (Black = total copy number, Red = haplotype A copy number, Blue = haplotype B copy number).

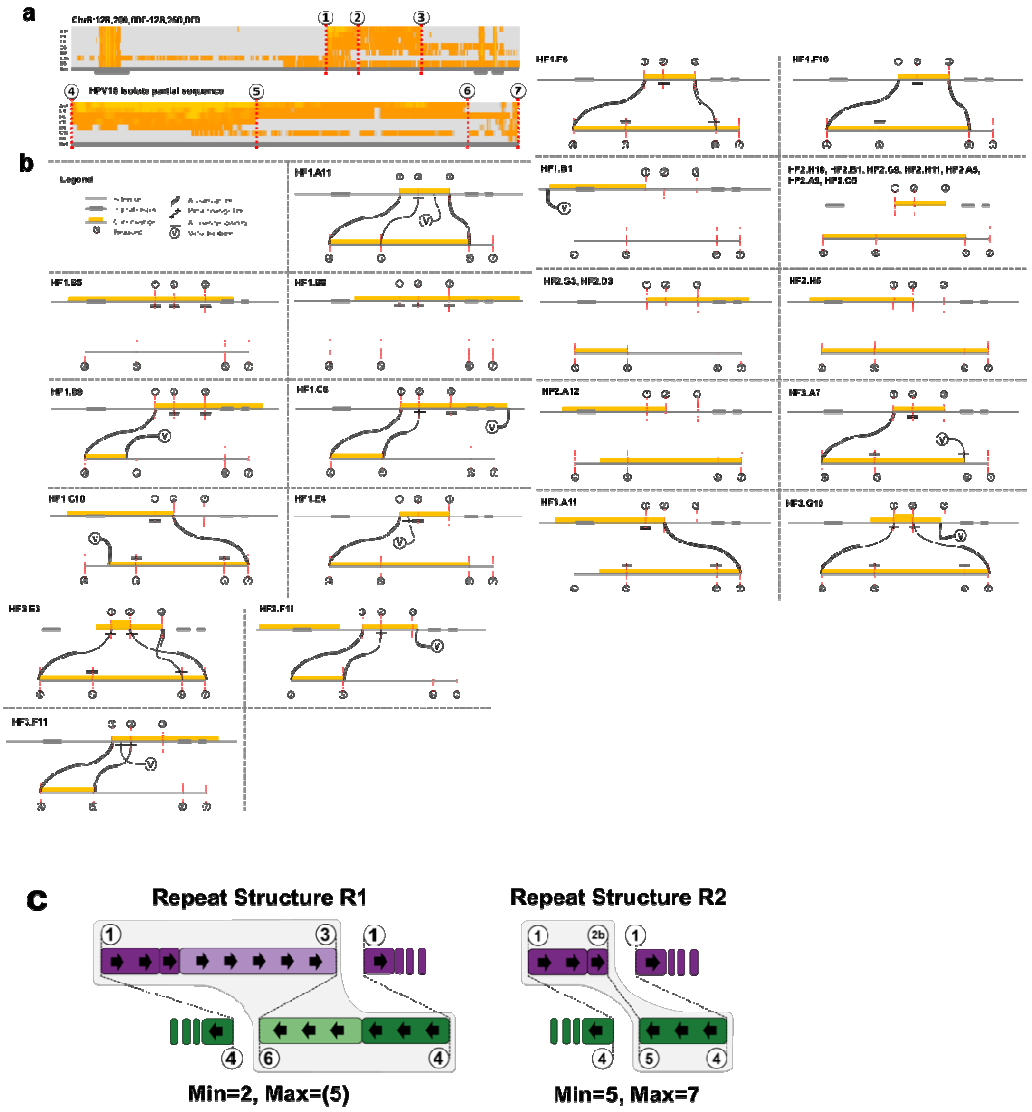
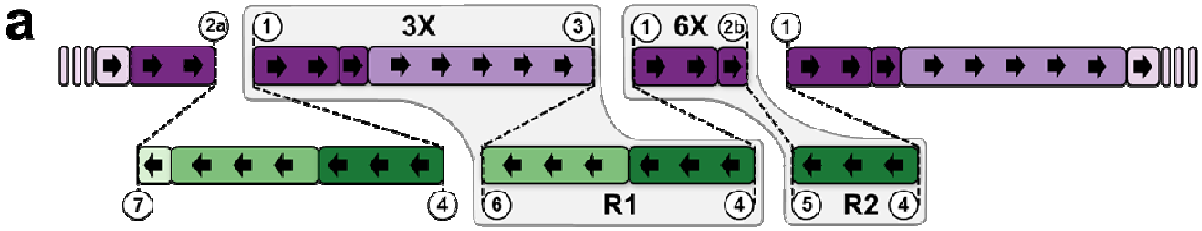
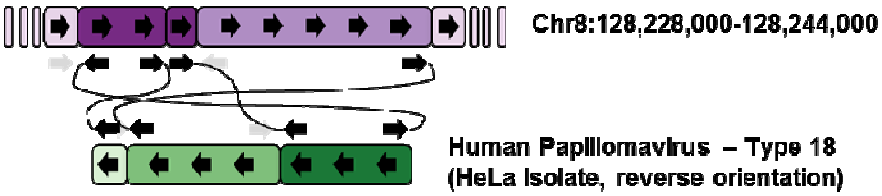


Figure C.3.30 | Structure of the HPV-18 integration locus.
a. Heat maps showing coverage from fosmid clone pools across the insertion site flanking region (upper) or the HPV-18 partial sequence (lower). Circled numbers correspond to breakpoints in which discordantly mapping read pairs were found that link the HPV-18 and chromosome 8 references. **b.** Diagrams of individual clones' coverage across the integration site and HPV-18. Yellow bars indicate coverage in the region, double black lines parallel to the reference indicate all read pairs are concordant across the breakpoint; single black links with a single line parallel to the reference indicate some read pairs are concordant across the breakpoint and others correspond to the link; double black links indicate that all read pairs support the link; links to a circled "V" indicate that read pairs are present at that location that link to the backbone vector sequence and therefore mark the end of a clone's coverage. **c.** Proposed structure of repeat units based on fosmid coverage profiles. Repeat R1 has a minimum of 2 copies, as breakpoint 2b-5 is never observed in fosmids containing breakpoint 3-6. The observed coverage profile over repeat 1 indicates a maximum of ~5 copies. Repeat R2 has a minimum of 5 tandem repeats, as the HPV region from breakpoint 5 to breakpoint 6 is never observed in clones entering the repeat from the centromeric end of the region, and a maximum of 7 due to never observing fosmids solely containing the chromosome 8 region from breakpoint 1 to 2b and HPV region from breakpoint 4 to breakpoint 5.





c

Breakpoint 2a-7 Assembled

GTT ATT ACA CAG CTA TCA GAG CAA GAG GCA GGT TAG TAA AAG CTG GTG GAC CTT AAA GTT TCT CTA CTT TTG CAA
GTG TAA AAA CTG GGG TAA AGA TAG AGT TTT GTT TTT CCT CGG TTT TGT ATG CAC TTT GTG CAA GGC CTT GTA GGG
CCA TTT GCA GTT CAA TAG CTT TAT GTG CTT TAC TTT TTG AAA TGT TAT AGG CTG GCA CCA CCT GGT GGT TTA ATG
TCT GTA TGC CAT GTT CCC TTG CTG CAA ACA ATA TTG CAT TTT CCC AAC GTA TTA GTT GCC AAT ACT GTA TTT GGC

Breakpoint 1-4 Assembled

GAA ACC TTA GGA ATA TCC TGC TTA TTG CCA CCA CCT GCA GGA ACC CTA AAA TAT GGA TTA CCA ACA GTT AAT AAC
CAG ACA AAA ACT TTA ATA ATA TTT GTC AAA TGC CAA ATC GGA GTC CAA AGC CAT TGT CCA TTT TAA GAA AAT CAT
CTG ACT TAA CAT CAC TAC TGC TTT TCA ACA GAG CAT CAT GCC CAT TTC ACA GAA GAG AAA ATT TGG CCT CAT ACT
CCT CAG TCT CCA TGT TTT AGC TTA GAT ATT GTT TCC TCC AGG CGG CCC TCC TTG ACC TTC CAA TTC TGG TTA AAT
TGC TCT TTT TCT GAG TTC TCA TTA CTT TAC TGA TTT TAT ATA TGT GTG TGT GTA TAT ATA TAT ATA CAC ACA CAC
ACA CAC ACA CAT ATA TAT ACA CAC ACA TAT ATA TAT ATG TTG TGC CTA GCA AGT GTA TGA CAC AAA ATA CCC ATA
AAT TGA ATG AAT GAA TGA ATT AAT AAA GAA ATG AAT AAC TTA CCC AAC CTG GTA AGT GGC AGG GCT GGC CAG GTC
AGT GCA ACT TCA AAG TCG ATG TTG TCA GTG AAT GCT CCA GAT GGA TTG CAG AGA AGA CC

Breakpoint 3-6 Assembled

TCT TCT ATG AGC TTC GTC AAG TCA TTT AAG CTT GGT ACC CGT CAG TTT CCT CAT CTG AAA ACT GAG AAA AGT TGT
TTC AAA TTG TCT AAG TCC ATT CCA GCT TGA TCA TAC TAG CAT CTT ATG TGC AGC TTC TTA AAG TCC AGC TCA CAC
CTC TGT CAA CTC CCT GTA TAA TAT GAC TTC CAA AAA AAC ACC TGT GGT TTG GTT ATA CAT ATA TAT GGA CAT ATA
TAT GTT ATA ACA TGG CCA CCT TAG TAT CTG TTA ACG GTT CCA ACC AAA AAT GAC TAG TGG AAT TCA CAA ATG ATA
TTA CTG CTC CTT GTA TAA AGT GTA TAA AAC TCA TTC CAA AAT ATG ATT TTC CTG TAT TTG CTG GTC CAC AAA ATA
CTA AAC

Breakpoint 2b-5 Assembled

AGG AAC AAA GGA ATC GAG GCA GGA AGG GAA GAA AAA ATG AGA AAA ACC ATA AGG CCA GGC GCG GTA GCT CAC GCC
AGT AAT CCT AAT ACT TTG GGA AGC TGA GGC GGG TGG GCG GAC CAC GAA GTC AGG AGT TCG AGA CCA GCC TGA CCA
ATA TGG CAA AAC CCC ATC TCT ACT AAA AAT CCC AAA AAA AAA AAA AAA AAA AAA AAA AAG TTA GCC GGG TAT GGT
GGC ACG TGC CTG TAA TCC CAG CTT CTG GGG AGG CTG AAG CAG GAG AAT TGC TTG AAA CCG GGA GGT GGA GGT TGC
AGT GAG CCG AGA TCA CAC CAC TGC ACT CCA GCC TGG GTG ACA GAG TGA GAC TCC GTC TCA AAA AAA AAA AAA AAA
AAA AAA AAA AAA AAA AGA AAA AGG AAA AAG AAA AAA AAG CAA CCA TGA GAC GAG CAA GAA GCT AAG TTT ACT ACA
ATT GTT AAA AGT ATT AAT GAA AAG TAT AGT ATG TGC TGC CCA ACC TAT TTC G

Figure C.3.31 | Assembly and sequencing of the HPV-18 integration site.
a. Proposed structure of the chromosome 8 locus containing the HPV-18 integration. **b.** Priming sites used to generate amplicons for breakpoint confirmation and assembly. Connecting black arrows indicate successful PCR amplicons and assembled breakpoints, gray arrows indicate additional primer sites that were tested which did not yield products. **c,** Assembled breakpoints performed via shotgun sequencing and assembly of gel-based size selected amplicons. Purple corresponds to human sequence and green to viral sequence, black nucleotides without an underline indicate sequence that share no homology with human or HPV-18 sequence, black nucleotides with double underline indicate sequence micro-homology with both human and HPV-18 sequence, underlined regions in color are the primer sequences used to generate amplicons.

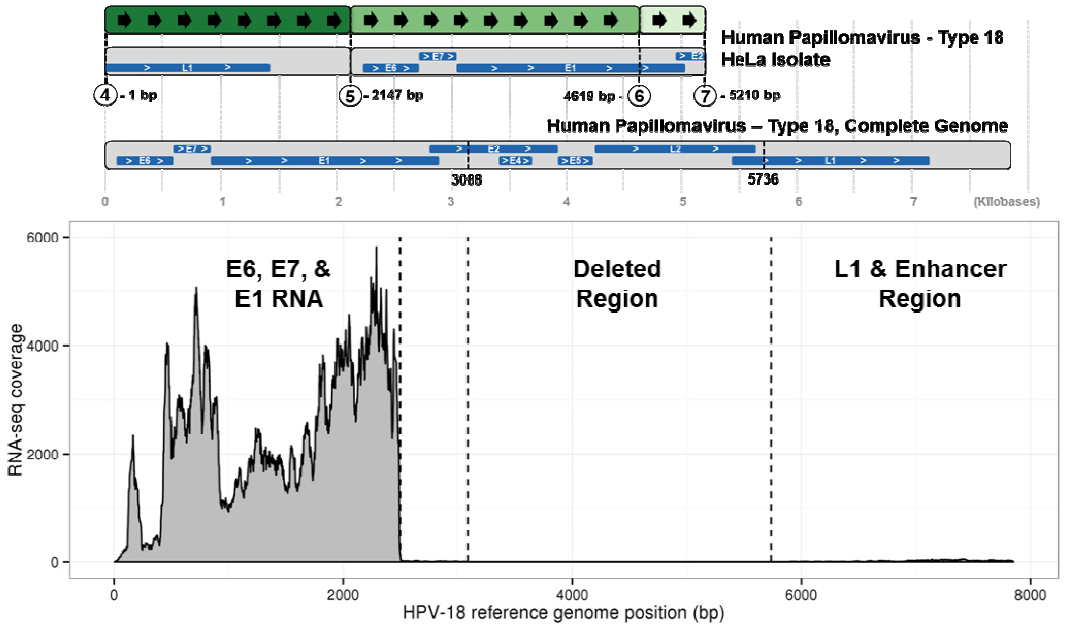


Figure C.3.32 | HPV-18 RNA-Seq coverage.
Area chart (bottom panel) represents RNA-Seq level of coverage that reaches nearly 6,000 fold. Above the chart is the diagram of the HPV-18 portion of the integration locus on chromosome 8q24.21 from **Figure 5.2** for reference.

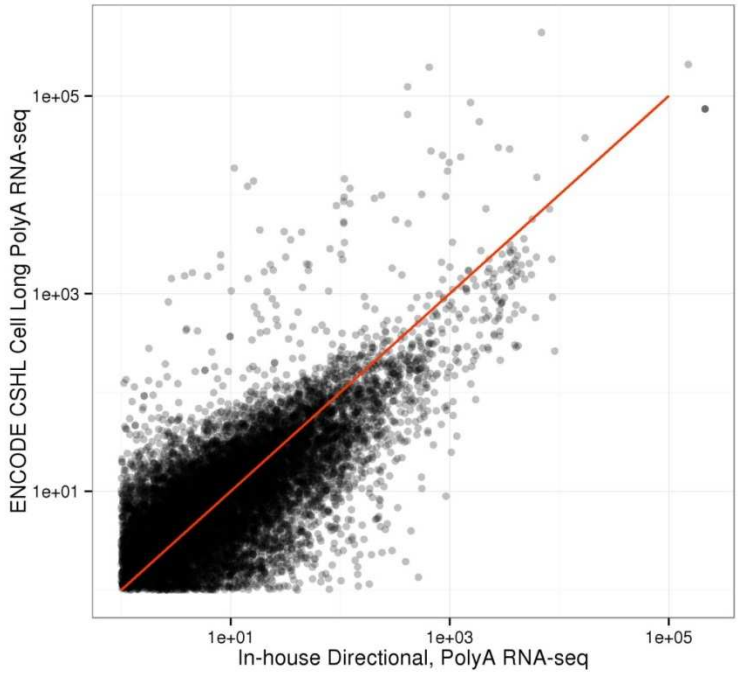


Figure C.3.33 | Correlation between RNA-Seq datasets.

HeLa S3 transcript abundances (reads per kilobase per million reads, RPKM) from ENCODE RNA-Seq (Cold Spring Harbor – Cell long PolyA) were plotted against those our own RNA–Seq data. Each point represents one RefGene-annotated transcript (for transcripts with >= 1 RPKM). Red line is y=x.

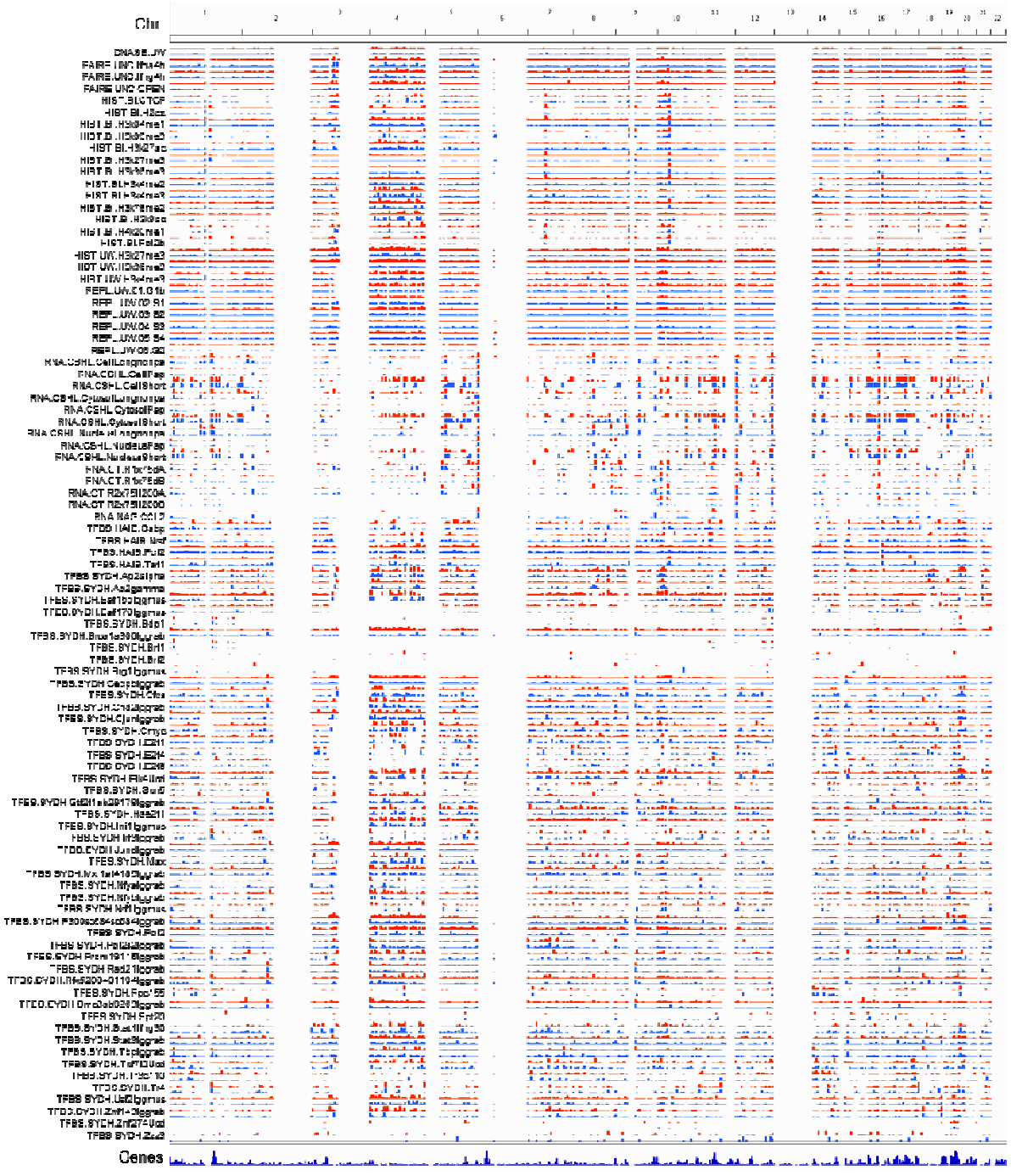


Figure C.3.34 | Phased ENCODE data tracks.

Per-haplotype enrichment scores for (A: red, B: blue) ENCODE HeLa S3 dataset. Height of each point indicates the degree of bias towards towards each haplotype after normalizing for its underlying copy number.

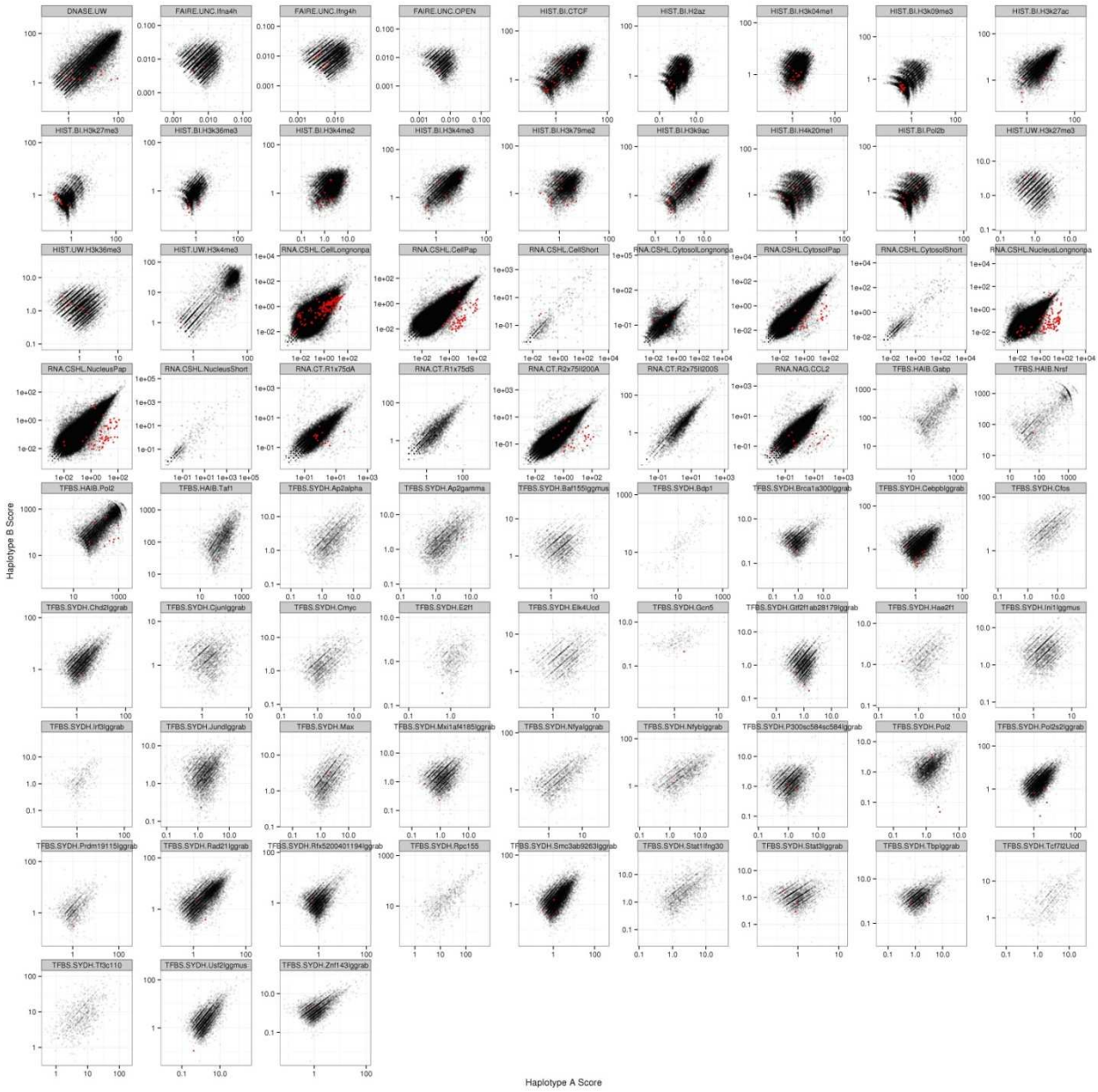


Figure C.3.35 | Correlations between haplotype-specific signals for ENCODE HeLa datasets.

Copy number normalized haplotype B-specific signals are plotted against haplotype A-specific signals for ENCODE HeLa S3 datasets. Each point represents the mean haplotype-specific scores for called peaks (ChIP-seq and DNase-seq) or annotated transcripts (RNA-Seq). Peaks residing near the HPV integration site on chromosome 8q21.24 are represented by red points.

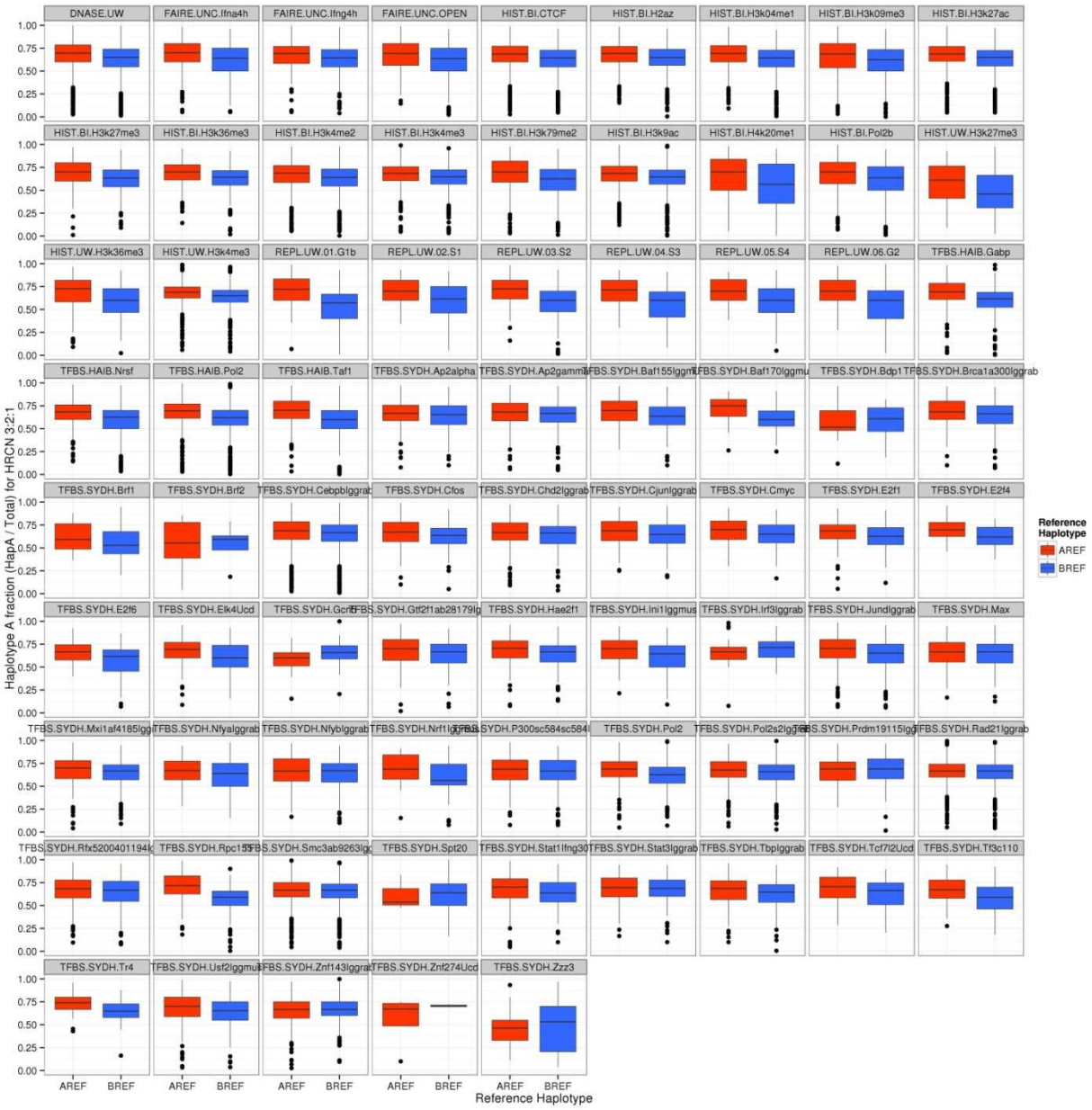


Figure C.3.36 | Reference bias in ENCODE peaks.

Average degree of reference biases in ENCODE peaks within HRCN 3:2:1 regions are shown as box-and-whisker plots. Red bars represent the haplotype A fractional contribution when the haplotype A allele is the reference base. Blue bars represent haplotype A fractional contribution where the haplotype B allele is the reference base.

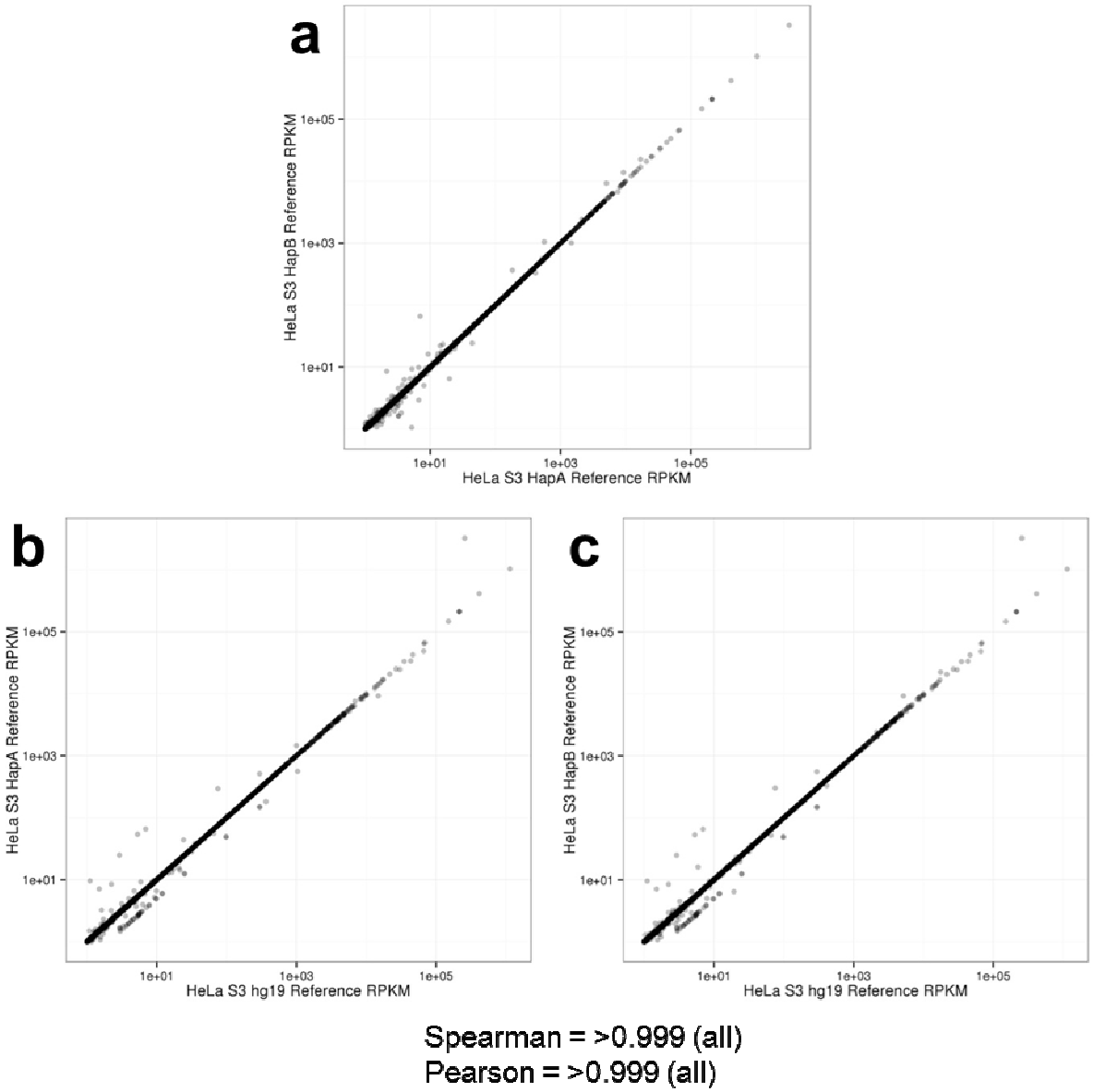


Figure C.3.37 | Minimal impact of reference bias upon transcript quantitation.

HeLa S3 RNA-Seq reads (this study) were aligned using TopHat[212] to the reference genome ("hg19"), as well as to HeLa haplotype-specific reference genomes ("HeLa Haplotype A" and "HeLa Haplotype B"). Transcript abundances were estimated against RefGene annotations using Cufflinks[213] then compared for all transcripts with an RPKM score >=1. **a.** Comparison between HeLa Haplotype A reference (x-axis) and HeLa Haplotype B reference (y-axis). **b.** Comparison between hg19 and HeLa Haplotype A reference (y-axis). **c.** Comparison between hg19 and HeLa Haplotype B reference (y-axis).

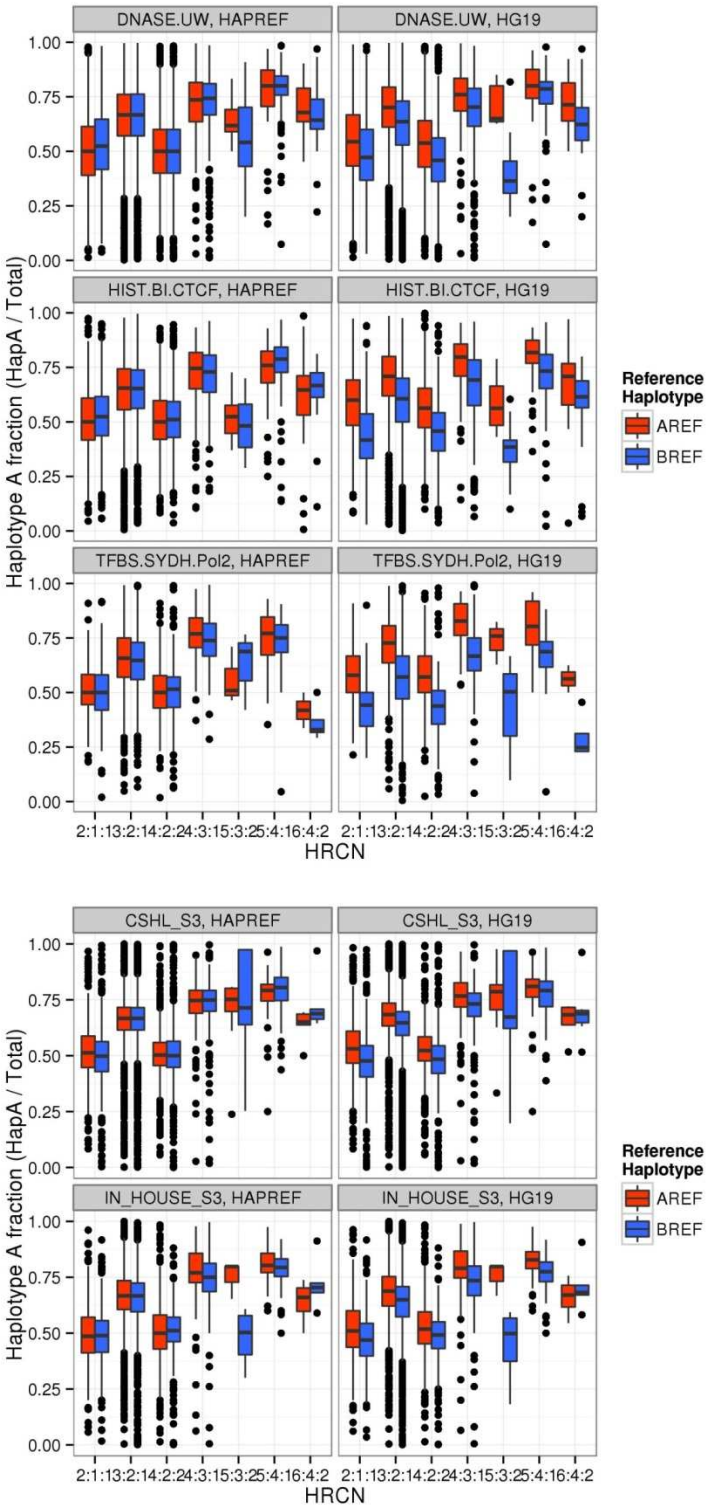


Figure C.3.38 | Reference bias removal.

Reference haplotype imbalance for different HRCN classifications for in HeLa when aligning to a HeLa haplotype-resolved reference (left) or hg19 (right). **a.** Reference bias in ChIP-seq peaks. **b.** Reference bias in RNA-Seq. Red bars represent the haplotype A fractional contribution where the haplotype A allele is the reference base. Blue bars represent haplotype A fractional contribution where the haplotype B allele is the reference base. The use of a haplotype-resolved HeLa reference greatly reduced the reference associated bias.

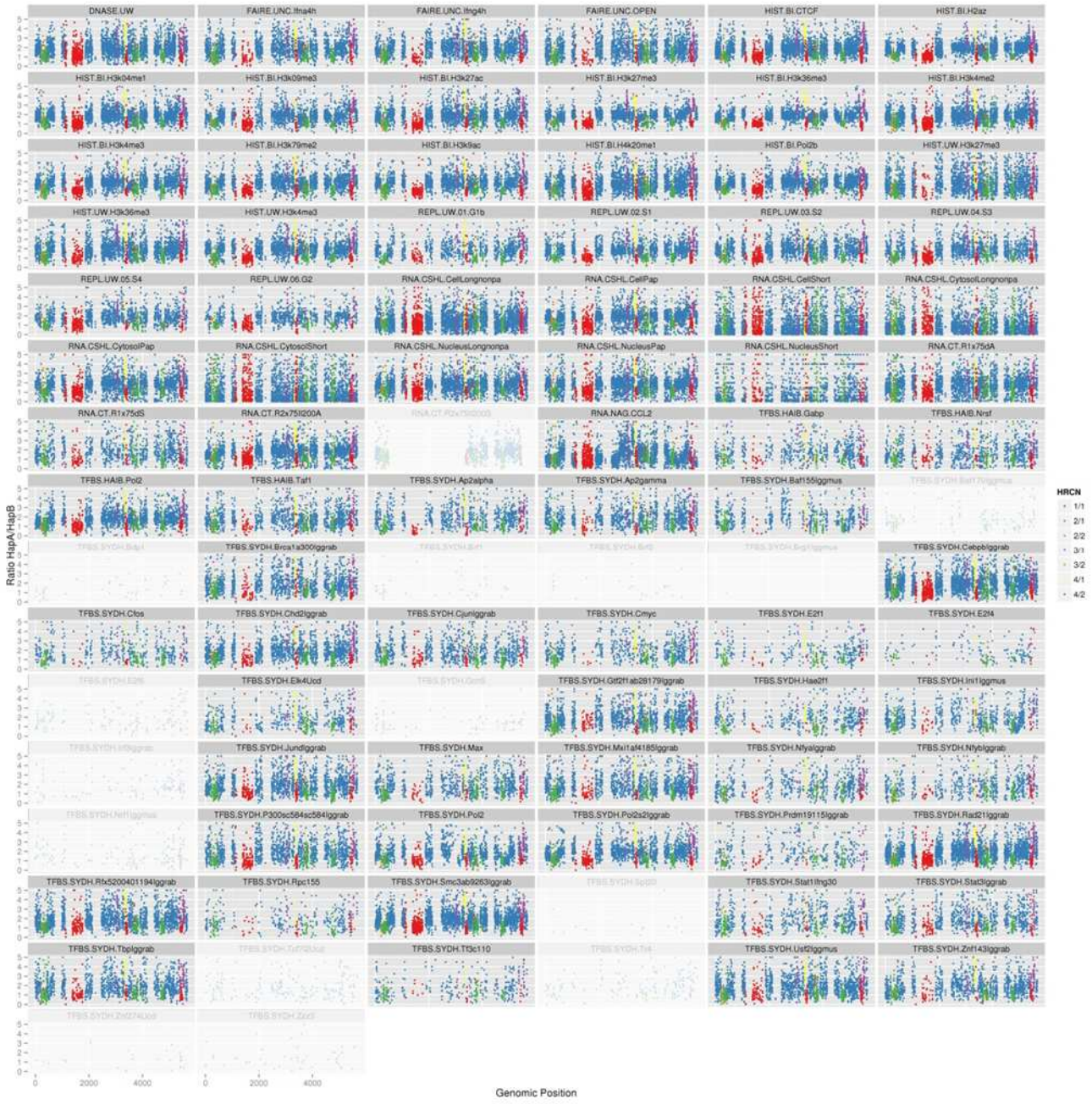


Figure C.3.39 | Haplotype contributions of phased ENCODE data (windows). Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows. Each window is color coded by the haplotype A to haplotype B ratio. Dimmed panels indicate data sets with very insufficient numbers of peaks for windowed analysis.

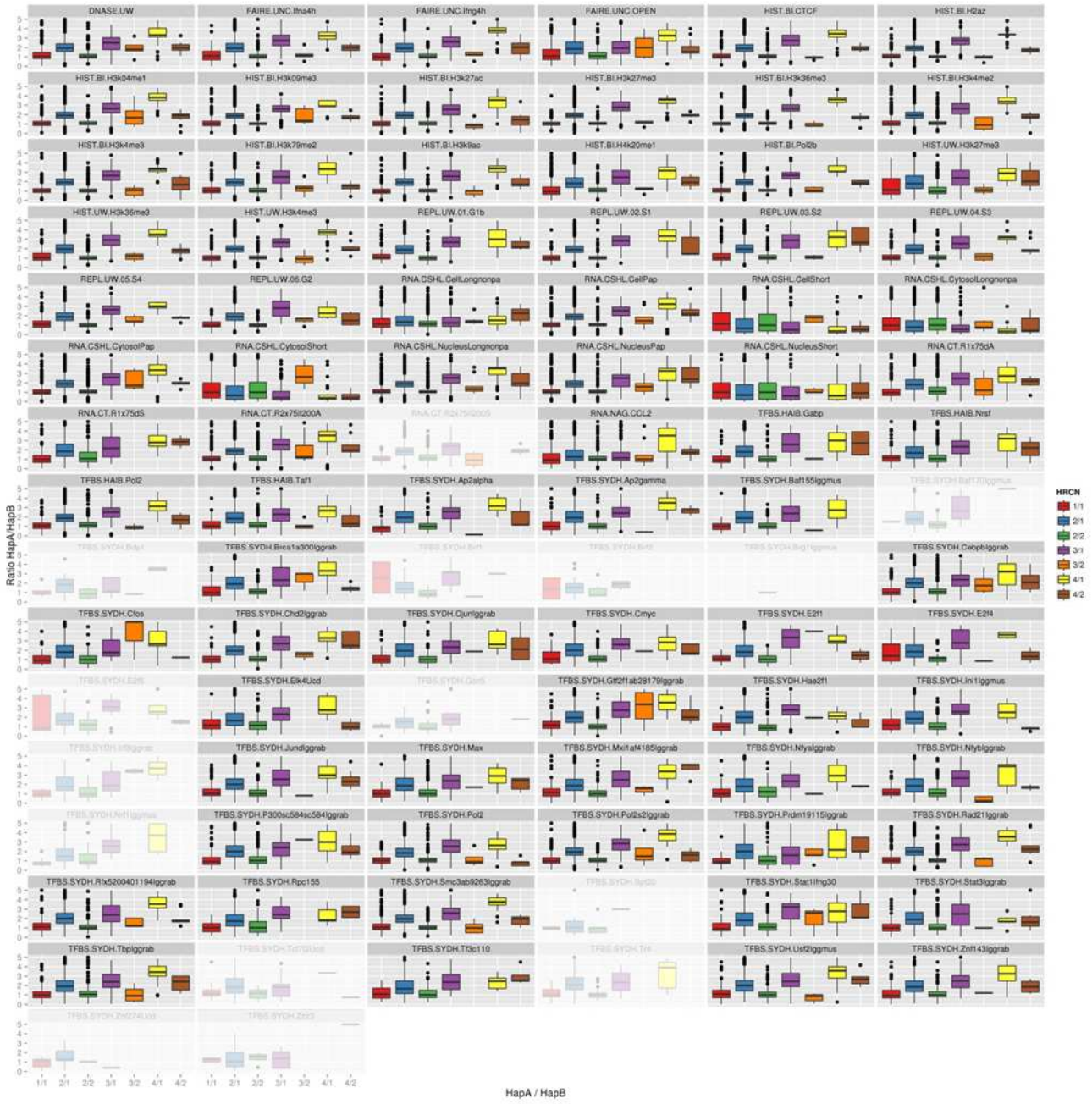


Figure C.3.40 | Haplotype contributions of phased ENCODE data (box plots).
Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows shown as box-and-whisker plots. Shaded out panels indicate data sets with very low peak counts and thus can not be reliably analyzed.

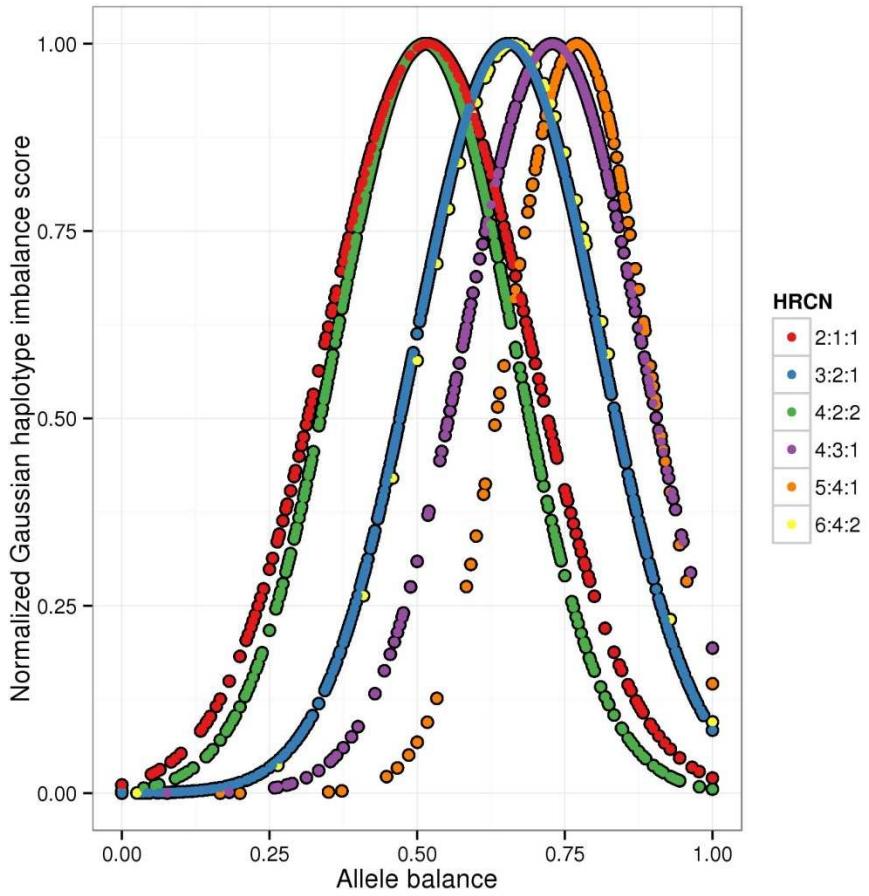


Figure C.3.41 | Normalized haplotype imbalance scores by copy number.

Normalized haplotype imbalance scores were calculated and split by the underlying HRCN (total CN : hapA CN : hapB CN). The majority of the HeLa genome has a higher haplotype A copy number (as per naming conventions) and therefore expected allele balances of haplotype A over total are shifted closer to 1 (except in haplotype-balanced regions, ie 2:1:1 and 4:2:2). This results in a reduced ability to call outliers of excessive haplotype A contribution due to the reduced range of allele balance from the null hypothesis to 1 (eg. for HRCN 3:2:1, the range for haplotype B to be considered an excessive contributor is 0.33<B<=1 whereas the range for haplotype A is 0.66<A<=1).

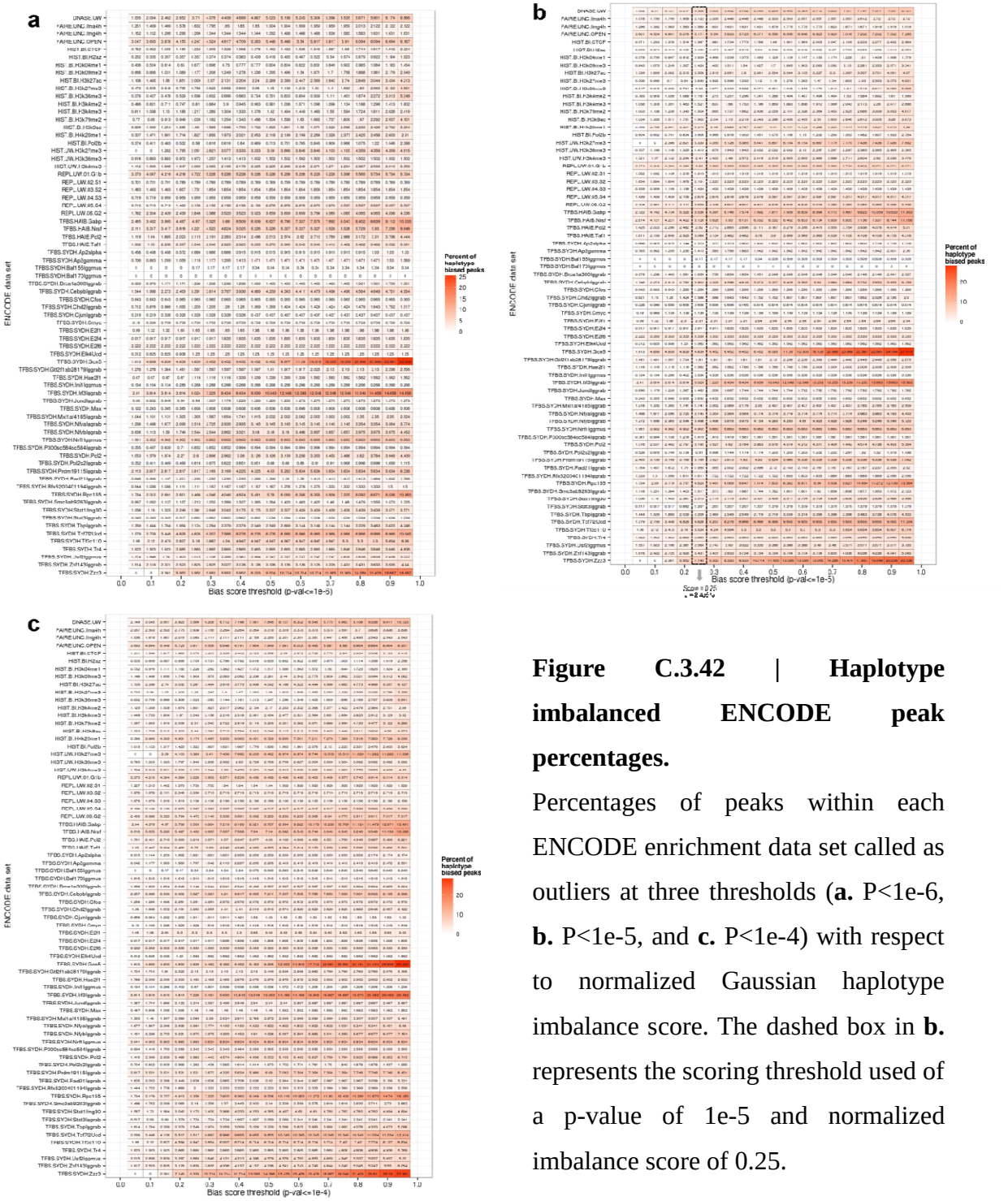


Figure C.3.42 | Haplotype imbalanced ENCODE peak percentages.

Percentages of peaks within each ENCODE enrichment data set called as outliers at three thresholds (**a.** P<1e-6, **b.** P<1e-5, and **c.** P<1e-4) with respect to normalized Gaussian haplotype imbalance score. The dashed box in **b.** represents the scoring threshold used of a p-value of 1e-5 and normalized imbalance score of 0.25.

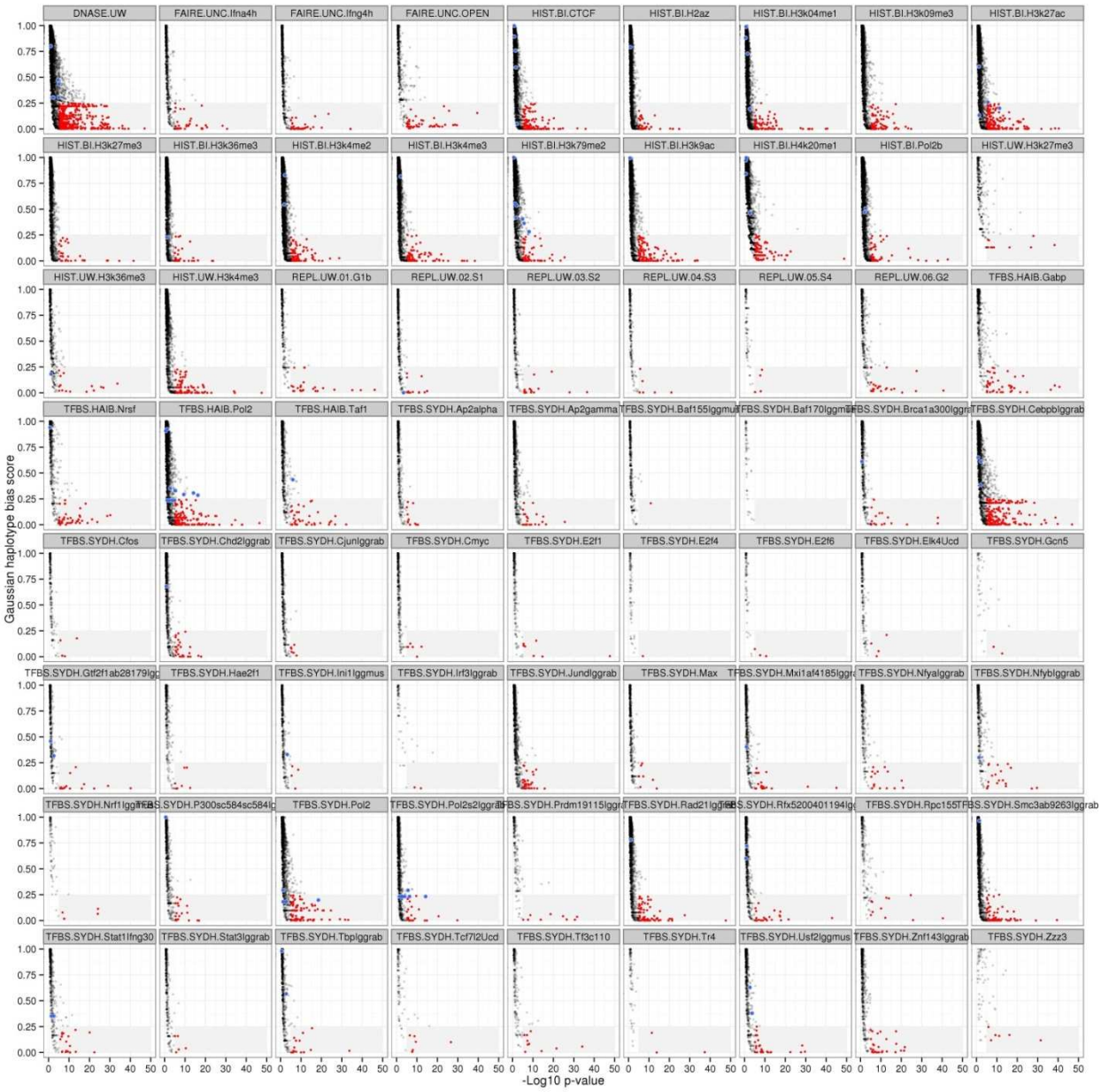


Figure C.3.43 | ENCODE peak haplotype imbalance scoring.

For each peak with an ENCODE data track, the normalized haplotype imbalance score is plotted against the –log10 p-value (the degree of significance against the null hypothesis of haplotype-balanced signal). Gray boxes with red points represent peaks called as outliers at a P<1e-5 and normalized haplotype imbalance score of <=0.25. Blue dots represent peaks near the HPV-18 / MYC locus.

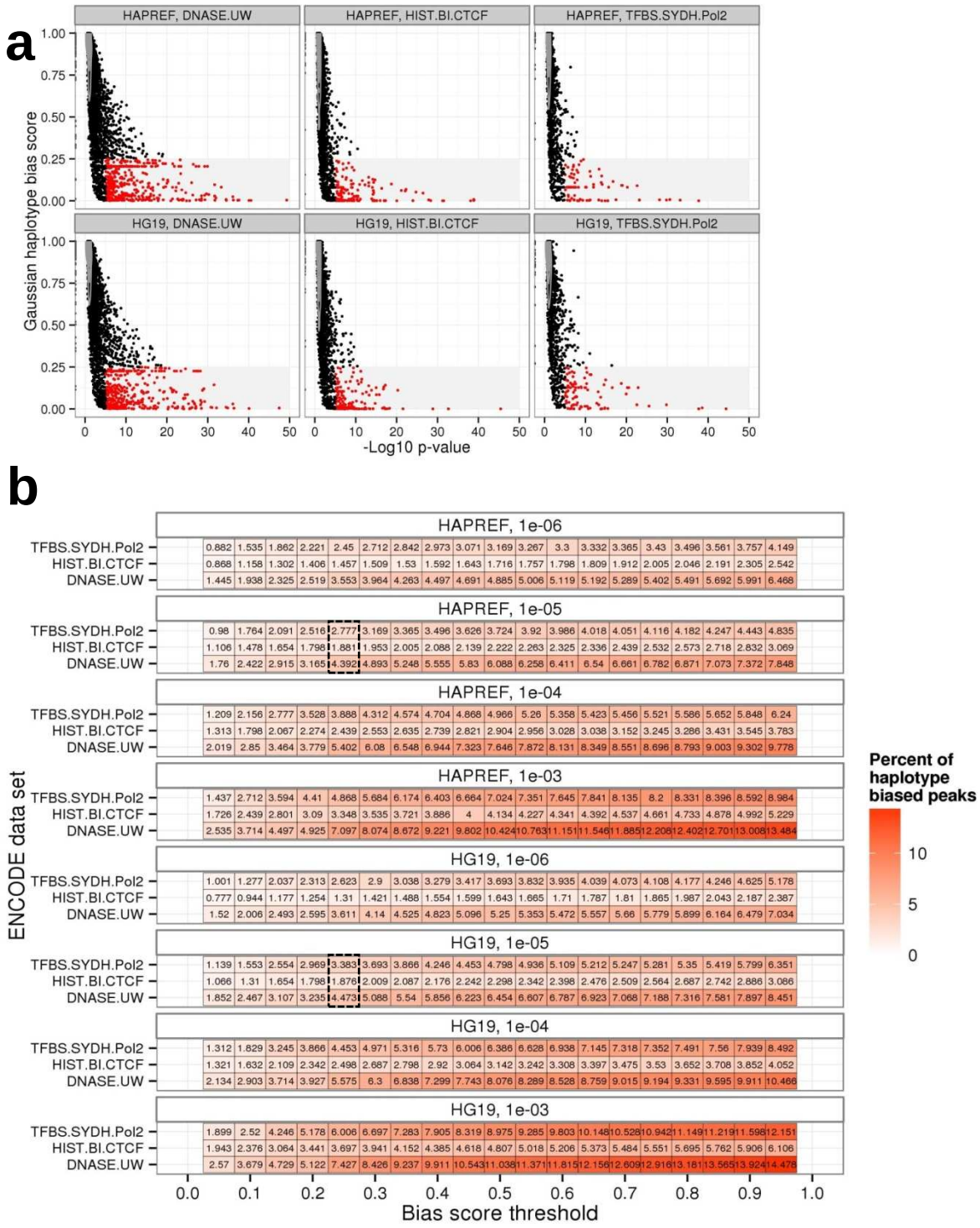


Figure C.3.44 | ENCODE peak reference bias effects on outlier calling.
The use of a HeLa-specific haplotype-resolved reference eliminates the reference bias, but does not substantially change the set of peaks called as outliers. **a.** Haplotype imbalance scores when aligning to a haplotype-resolved HeLa reference (top) or hg19 (bottom). **b.** Percentage of peaks called as outliers with P<1e-5 and an imbalance score cutoff of 0.25. Using HeLa haplotype-specific reference sequences changes the set of outliers called by only 0.606% 0.005%, and 0.081% for Pol2 ChIP-seq, CTCF ChIP-seq, and DNaseHS-seq, respectively.

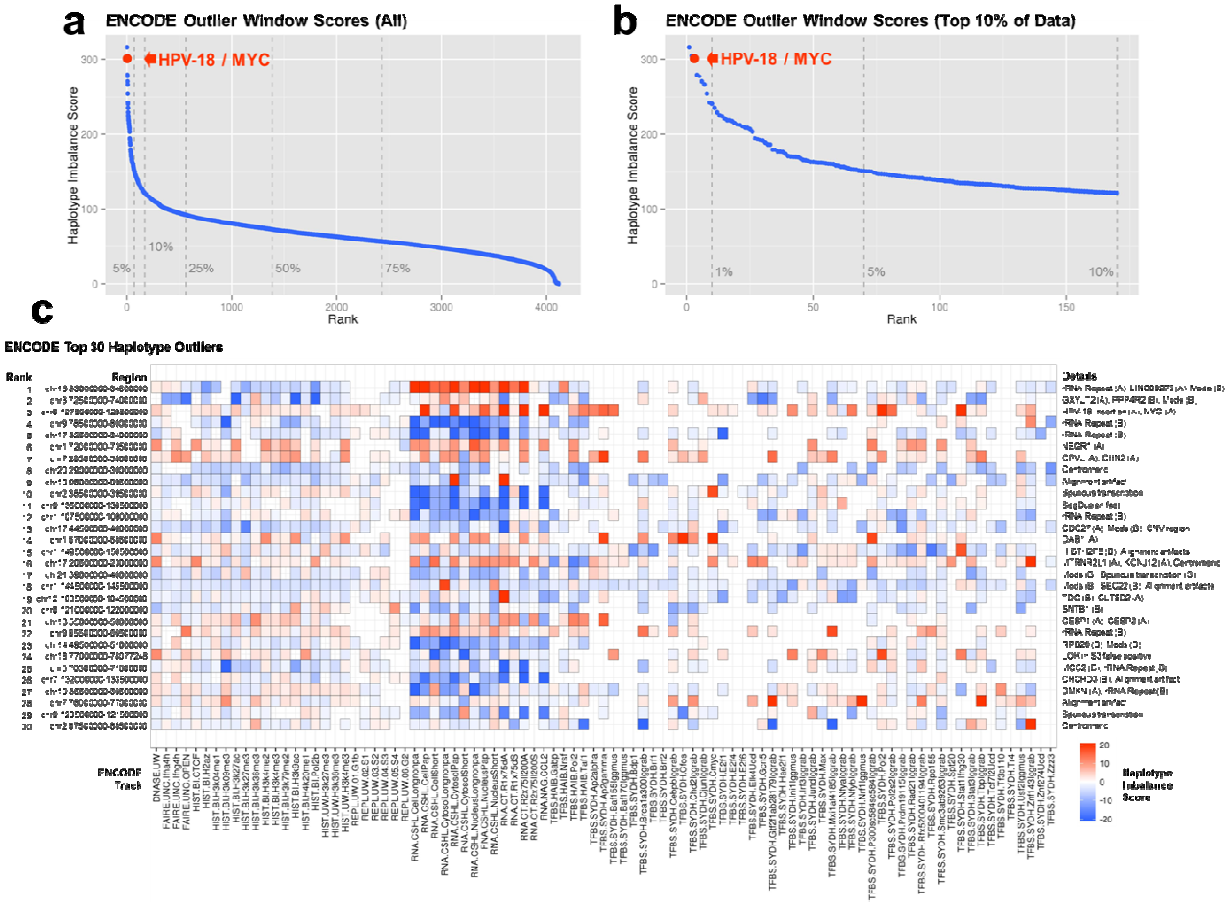


Figure C.3.45 | Phased ENCODE outlier analysis.

a. Ranked sliding window scores for ENCODE outliers. The window containing the HPV-18 integration and MYC loci is highlighted in red. **b.** Closer investigation of the top 50 scores from **a**. **c.** Top 30 outlying windows after being condensed for overlapping regions. Columns represent haplotype imbalance scores for haplotype A (red) and haplotype B (blue) for the ENCODE tracks across each region.

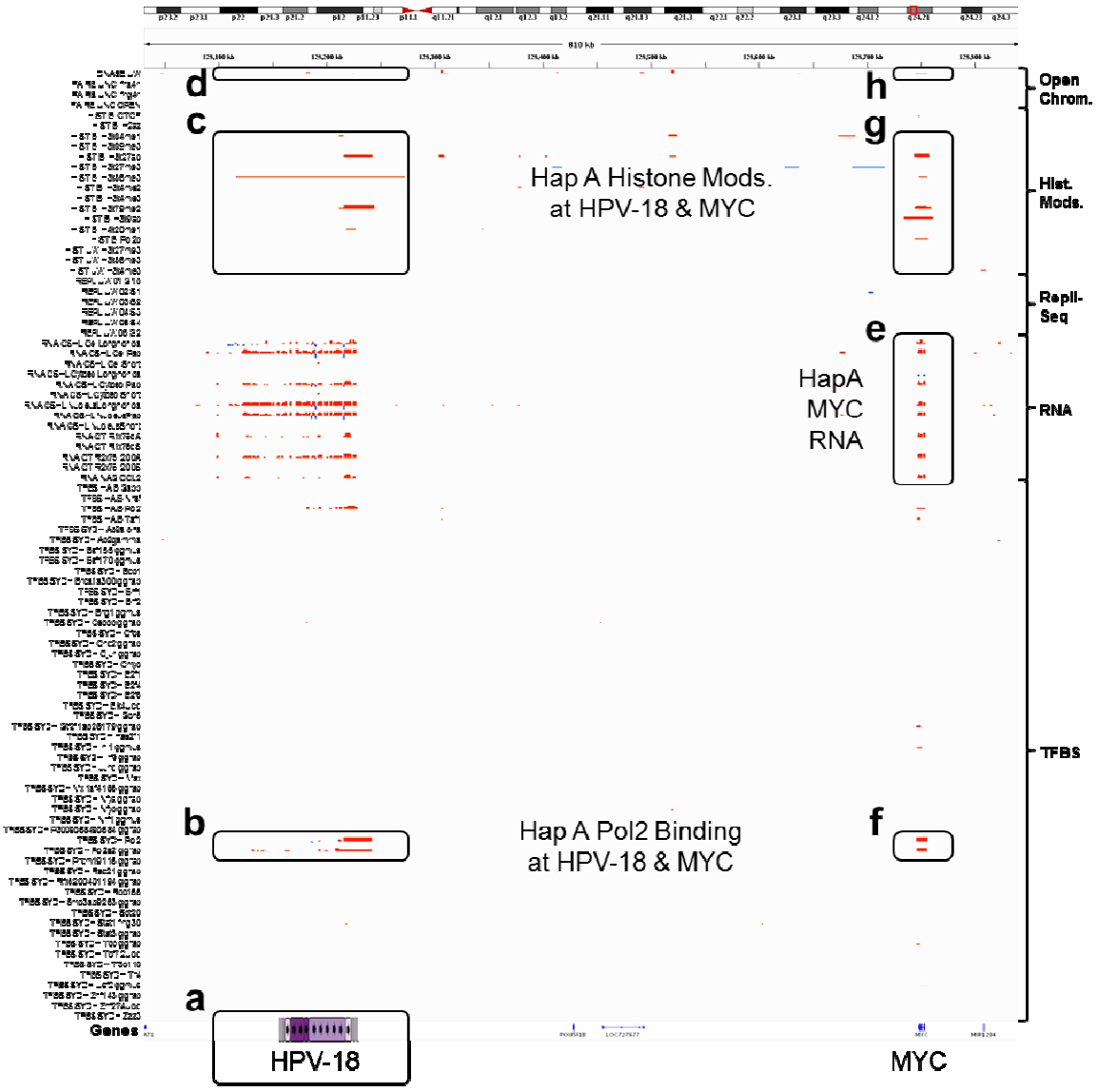


Figure C.3.46 | ENCODE haplotype imbalances for HPV-18 and MYC.

Red peaks indicate haplotype A imbalance, blue peaks represent haplotype B for copy number normalized haplotype imbalance scores. Letters represent regions corresponding to items in the schematic presented in **Figure 3.4a**.

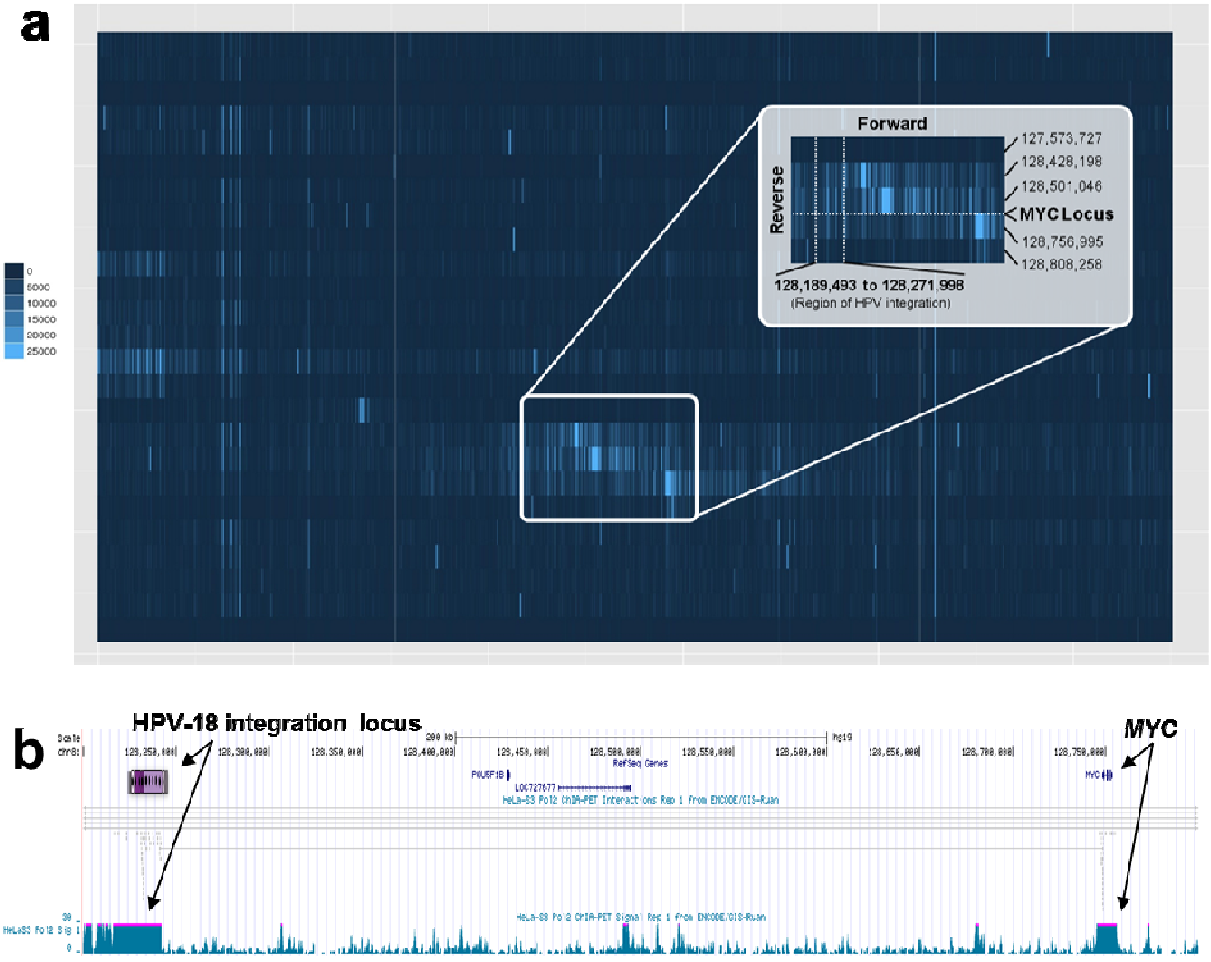


Figure C.3.47 | Long range with *MYC* from 5C and ChIA-PET data.

a. ENCODE 5C chromatin interaction data (available only for the GM12878 cell line) demonstrates long-range interactions between *MYC* and distal upstream sites. The highlighted region includes the site of HPV-18 integration (into the HeLa but not GM12878 genome). **b.** Spanning reads from ENCODE ChIA-PET data in HeLa S3 cells indicate long range integration between the HPV-18 interaction and site and *MYC* locus. Teal profile represents Pol2 signal and contains peaks at the HPV-18 and *MYC* loci.

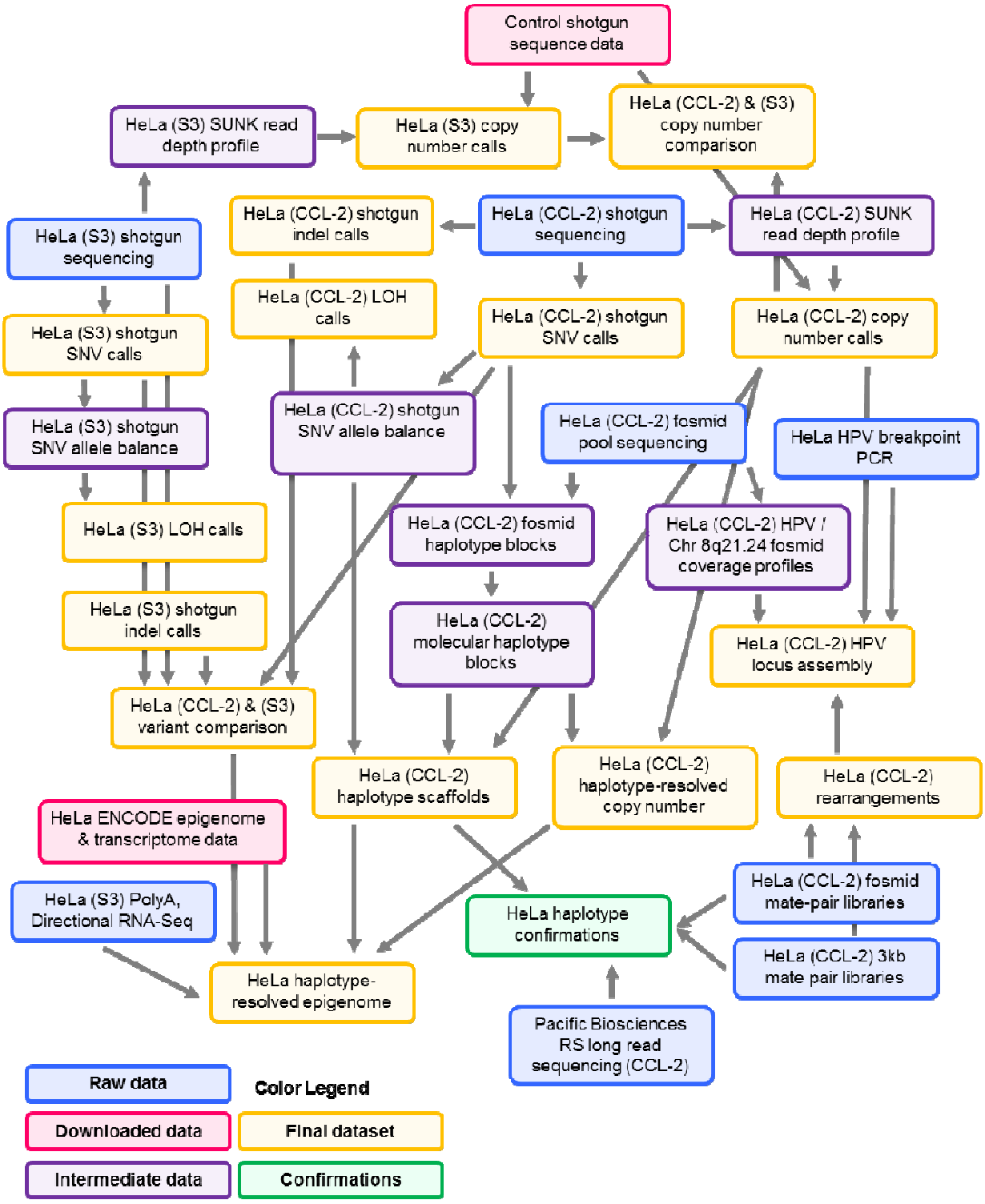


Figure C.3.48 | Datasets and analyses for HeLa CCL-2 and HeLa S3.

GLOSSARY

This Glossary is provided as a reference to make my dissertation more accessible to readers who may not be familiar with the terminology of genomics. Some of the definitions here are adapted from the NCBI genomics glossary at http://www.ncbi.nlm.nih.gov/projects/genome/glossary.shtml, the NHGRI genomics glossary at http://www.genome.gov/glossary/, and Wikipedia.

Alignment: Alignment is the process of matching two or more nucleic acid sequences to find similarities. There are many contexts and purposes for alignment, but one typical application is in resequencing of individuals from a species with a known reference genome (*e.g.*, humans) for the purpose of identifying differences between the individual an the reference.

Assembly: see ***Genome assembly***

Base: one of the chemical nucleotides incorporated into DNA or RNA: adenine (**A**), cytosine (**C**), guanine (**G**), thymine (**T**), or uracil (**U**).

Base pair (bp): A base pair is two chemical bases hydrogen-bonded to one another forming a rung of the DNA double helix. A double-stranded genome sequence consists of a string of base pairs. Sequence lengths are typically stated in either bases or base pairs, which are equivalent, although the latter technically only applies to double-stranded sequence. A kilobase (**kb**), megabase (**Mb**), and gigabase (**Gb**) are respectively 10^3, 10^6, and 10^9 bases of length.

Chromosome: A chromosome is a large molecule of DNA found in a cell, consisting of a primary DNA sequence with many proteins bound onto it, including chromatin and DNA binding proteins. In the human genome, there are 22 pairs of chromosomes as well as the X and Y sex chromosomes, and each chromosome has an average of 130 Mbp of DNA sequence length.

Contig: Short for "contiguous sequence." When two sequences (*e.g.*, reads) overlap at their ends, the sequences can be collapsed into a single, non-redundant sequence. Contigs can be combined into scaffolds, and contigs and scaffolds comprise assemblies.

Contiguity: Long-range information about the relative positions of sequences in a genome. In genome assembly, it is comparatively easy to assemble reads into contigs representing non-repetitive genomic regions, but creating longer contigs is more difficult, as is determining the relative position of these contigs. Sequencing-based methods that provide contiguity information include mate-pair sequencing, fosmid sequencing, and Hi-C.

Copy number variation (CNV): Large-scale structural variants in DNA that vary from individual to individual within a species. These include insertions, deletions, duplications, and complex multi-site variants that range from kilobases to megabases in size.

Coverage: A measure of the quantity of shotgun sequencing. A genome assembly project is said to have 1-fold or 1x coverage of reads if the total length of shotgun reads is equal to the genome size. The amount of coverage necessary to properly assemble a genome depends on the type and especially on the length of the reads. With Sanger sequencing, genome assemblies were typically created using 7x-8x read coverage; with Illumina sequencing, coverages in the range of 40x-60x are suggested. This increase in required coverage somewhat offsets the vast drop in the per-base cost of short-read sequencing.

Draft assembly: This term generally refers to an assembly that is not yet finished but is of generally high quality. A draft assembly consists of contigs and scaffolds with an unknown relative order and orientation. These sequences are a useful substrate for genome assembly and annotation, and may be used as reference genomes, but do not fully describe the genome of the organism.

Deletion: see ***Indel***

De novo assembly: A *de novo* assembly in a genome assembly in which no information is known about the genome prior to the sequencing and assembly. This is contrasted with building an assembly off of existing linkage or physical maps of the genome, or assembling by comparing with the genome of a closely related species (assisted assembly). *De novo* is Latin for "from the beginning".

Diploid: A cell or organism that has paired chromosomes, one from each parent, is diploid. In all organisms that reproduce sexually, including humans, the majority of somatic (body) cells are diploid. Diploidy can give rise to heterozygosity.

DNA: DNA (**d**eoxyribo**n**ucleic **a**cid) is the chemical name for the molecule that carries genetic instructions in all living things. The DNA molecule consists of two strands that wind around one another to form a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases. The two strands are held together by bonds between the bases; adenine bonds with thymine, and cytosine bonds with guanine. The sequence of the bases along the backbones serves as instructions for assembling RNA and protein molecules.

Finishing: Genome finishing is the process of converting a genome assembly into a complete sequence that accurately represents biological reality. Because the finishing process is time- and labor-intensive, most genome assemblies are left unfinished to some degree. The human genome is still not fully finished.

Fosmid: A cloning system based on the *E. coli* F factor. These clones have an average insert size of 40 Kb, with a very small standard deviation. Fosmid end sequencing is a sequencing strategy that is often used to generate mid-range genomic contiguity.

Gap: A region of the genome for which no sequence is currently available. Gaps may occur both within scaffolds (in which case they are represented as a set of Ns and may have a roughly known size) and between scaffolds.

Gene: A gene is the molecular unit of heredity of a living organism. On a molecular level, it refers to some stretches of DNA that code for a polypeptide or for an RNA chain that has a function in the organism. The exact definition of a gene depends on functional analysis and is imperfectly delineated, as is the question of what constitutes "different" genes.

Genome: The genome is the set of genetic material of an organism. In all organisms other than viruses, the genome is encoded in DNA and consists of one or more chromosomes. Every species has a genome, which is understood to be a rough representation of the "average" genome of a member of that species. In reality, every organism has a different species, and in fact every cell in every organism may have a slightly unique genome due to somatic mutations.

Genome assembly: Genome assembly is the process of taking a large number of short DNA sequences (reads) such as are produced by sequencing machines, and putting them back together to create a representation of the original chromosomes from which the DNA originated. The term "assembly" also refers to the product of this process, *i.e.*, "a genome assembly".

Genomics: A word coined in 1986, genomics refers generally to the study of genomes. Genomics may be thought of as a more high-throughput, 21st-century version of genetics, in which the power of contemporary computing and sequencing technologies are brought to bear.

Germline mutation: A mutation in a germline cell. The germline cells are the sex cells (eggs and sperm) that are used by sexually reproducing organisms to pass on genes from generation to generation. Germline mutations are contrasted with somatic mutations, which are not passed on to offspring. If germline mutations are established within a population, they become variants.

Haplotype: A contraction of "haploid genotype". A haplotype is a set of DNA variants that co-occur on a chromosome and thus tend to be inherited together.

Haplotyping: Also known as haplotype phasing or haplotype resolution, this is the process of determining an individual's haplotype from genotype by determining which heterozygous variants are present on the same haplotype as one another. There are many methods for haplotyping, some of which are molecular and some of which are statistical.

Heterozygous: A diploid individual is said to be heterozygous at a particular genomic position if he/she has two versions of that position in his/her two copies of the same chromosome.

Homozygous: Not heterozygous. A diploid individual is homozygous at a genomic position if the two alleles are the same.

Human Genome Project (HGP): The Human Genome Project was an international project that mapped and sequenced the entire human genome, begun in 1990 and completed in April 2003. It made possible the 21st-century practice of human genetics and is widely cited as one of the greatest successes in scientific history.

Illumina: A company based in San Diego, California that develops several integrated systems for biological applications, notably sequencing machines. Since its 2007 acquisition of Solexa, Illumina's sequencing machines have succeeded in producing increasingly massive numbers of short reads and have been a major player in the field of next-generation sequencing.

Indel: Short for "insertion or deletion", an indel is a type of variation that consists of an insertion or deletion of a small number of bases. Indels are typically defined as <1 kbp; larger changes are classified as CNVs.

Inversion: An inversion is a chromosomal rearrangement in which a segment of a chromosome is reversed end to end. An inversion occurs when a single chromosome undergoes breakage and rearrangement within itself.

Library: A sequencing library is a collection of DNA fragments that are derived from the same original genomic source, and designed to be sequenced together on a sequencing machine.

Mate-pair sequencing: A method of providing medium-range contiguity in genome assemblies. A fragment of roughly known length is isolated, and both its ends are sequencing, producing a pair of "paired reads" separated by a gap of approximately known length.

Metagenome: A metagenome is a set of genomes that are present in an environmental sample, such as a microbial community. Metagenomics is the study of genetic material recovered directly from environmental samples.

Metagenome assembly: An assembly made from shotgun sequencing and *de novo* assembly of a metagenome sample. A metagenome assembly consists of contigs with no species information. Assigning species information to contigs is called metagenomic deconvolution

Metagenomic deconvolution: The process of separating a metagenome into individual genomes, by determining which contigs in the metagenome assembly are from the same species. Metagenomic deconvolution is one of the great challenges in metagenomics. There are many methods for metagenomic deconvolution, some of which are molecular and some of which are statistical.

Microbial community: A community of microbes that live together in an environment. When it occurs inside or on the human body, often referred to as a microbiome. Microbial communities can be studied via metagenomics.

Microsatellite: Microsatellite sequences, also known as short tandem repeats (STRs) or simple sequence repeats (SSRs) are repetitive DNA sequences, usually several base pairs in length. Microsatellite sequences are typically composed of non-coding DNA.

Mobile genetic element: A type of DNA that can move around within the genome, such as transposons, retrotransposons, plasmids, and bacteriophage elements.

Mutation: A change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses. Mutations may be germline or somatic.

Next-generation sequencing (NGS): A term for sequencing technologies developed since the end of the Human Genome Project in 2003, in contrast to the older method of Sanger sequencing. The high demand for low-cost sequencing has driven the development of NGS technologies that parallelize the sequencing process, producing thousands or millions or even billions of sequences concurrently.

Nucleotide: see ***Base***

Pair: see ***Mate-pair sequencing***

Pan-genome: The set of all genes that are present in any strain of a bacterial species. Some bacteria, such as *E. coli*, have so much inter-strain variation in gene content that their pan-genomes are many times larger than any one strain's genome.

Quality score: A base quality score from a sequencing machine that indicates how confident the machine is in the base call. Quality scores are typically reported as phred scores. Alignment algorithms may also output mapping quality scores, indicating how confident they are in a mapping.

Read: A sequence of DNA that is output ("read") from a sequencing machine. Reads are the input to all assembly and alignment algorithms. Depending on the sequencing technology, reads may be of any length; they may have associated quality scores for their bases; and they may be paired.

Read pair: see ***Mate-pair sequencing***

Reference genome: A genome assembly that is deemed sufficiently high-quality to be used as a reference in further studies of that species. Reference genomes are used as targets in alignment. There has been a marked decline in quality of reference genomes since the days of the Human Genome Project.

Repeat: Repeats are a general term for all DNA sequences that occur more than once in a genome, including copy number variants, microsatellites, mobile genetic elements, and segmental duplications. Repeats are often polymorphic within a population. Repeats cause difficulty in genome assembly and alignment algorithms because a read falling within a repeat may not be unambiguously mappable to that instance of the repeat.

Sanger sequencing: A method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication. Sanger sequencing was developed by Frederick Sanger and colleagues in 1977 and became the most widely used sequencing method for nearly three decades, until the advent of next-generation sequencing (NGS) technologies.

Scaffold: A scaffold or supercontig is a set of two or more contigs that cannot be combined directly into a single contig, but are known to be close to each other with a gap between them. This commonly occurs using contiguity information such as can be obtained from paired plasmid ends. The process of combining contigs into scaffolds is called scaffolding.

Segmental duplication: A region of genomic DNA that may be found at more than one site in the genome. Segmental duplications are typically defined as sequences of length >1 kbp sharing >90% sequence identity.

Sequencing: Sequencing is the process of determining the precise order of nucleotides within a DNA or RNA molecule. It includes any method or technology that is used to determine the order of the four bases in a strand of nucleic acid. The first widely used sequencing method was Sanger sequencing.

Short read: A read that is short, typically defined as <200 bp. Sequencing companies such as Illumina have created technologies to produce staggering numbers of short reads. However, short reads are difficult to use in assembly algorithms because they lack contiguity information and cannot bridge repeats.

Shotgun sequencing: A sequencing method by which an entire genome is cut into chunks of discrete sizes and used to prepare genomic DNA libraries for sequencing and assembly. The term comes from the scattershot but high-volume nature of shotgun shooting. This is distinguished from targeted sequencing, in which certain regions of the genome are isolated and/or enriched prior to sequencing.

Single-nucleotide variant (SNV): A single base difference found when comparing the same DNA sequence from two different genomes. When the two genomes are from different individuals in the same population, this is known as a single-nucleotide polymorphism, or SNP.

Somatic mutation: A mutation in a somatic cell, *i.e.*, any cell other than eggs and sperm and their progenitors. Somatic mutations may cause cancer and other changes to the cellular phenotype. Somatic mutations are contrasted with germline mutations, which are passed on to offspring.

Structural variation: The variation in structure of an organism's chromosome. It consists of many kinds of variation in the genome of one species, and usually includes microscopic and submicroscopic types, such as deletions, duplications, copy number variants, insertions, inversions and translocations. A structure variation is typically defined as affecting a sequence length about 1Kb to 3Mb. Like other mutations, structural variants may be germline or somatic.

Translocation: A translocation is a structural variant caused by rearrangement of parts between nonhomologous chromosomes or different parts of the same chromosome. Translocations can be balanced (in an even exchange of material with no genetic information extra or missing) or unbalanced (where the exchange of chromosome material is unequal resulting in extra or missing genes).

Variant: A variant is a region of the genome that varies between individuals. It initially occurs as a germline mutation but then becomes part of a population's gene pool as it is passed down. Types of variants include single-nucleotide variants (SNVs), indels, and many types of structural variation.

REFERENCES

1 Winkler, H. *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. (Verlag von Gustav Fischer, 1920).

2 Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *The Journal of Experimental Medicine* **79**, 137-158 (1944).

3 Hershey, A. D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology* **36**, 39-56 (1952).

4 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

5 Gamow, G. Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature* **173**, 318-318 (1954).

6 Olson, M. V. The Human Genome Project: a player's perspective. *J Mol Biol* **319**, 931-942 (2002).

7 Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82-88 (1972).

8 Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500-507 (1976).

9 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).

10 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564 (1977).

11 Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687-695 (1977).

12 Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage λ DNA. *J Mol Biol* **162**, 729-773 (1982).

13 França, L. T., Carrilho, E. & Kist, T. B. A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics* **35**, 169-200 (2002).

14 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

15 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

16 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-547 (1996).

17 Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058-1073 (1996).

18 Sinsheimer, R. L. The Santa Cruz Workshop–May 1985. *Genomics* **5**, 954-956 (1989).

19 DeLisi, C. Meetings that changed the world: Santa Fe 1986: Human genome baby-steps. *Nature* **455**, 876-877 (2008).

20 Kuska, B. Beer, Bethesda, and biology: how "genomics" came into being. *Journal of the National Cancer Institute* **90**, 93 (1998).

21 "Great 15-Year Project To Decipher Genes Stirs Opposition" Angier, N. in *The New York Times* (New York, NY, 1990).

22 Watson, J. D. The human genome project: past, present, and future. *Science* **248**, 44-49 (1990).

23 Venter, J. C., Smith, H. O. & Hood, L. A new strategy for genome sequencing. *Nature* **381**, 364-366 (1996).

24 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
25 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
26 Gauging the Economic Impact of the Human Genome Project. *Human Gene Therapy* **22**, 777-779 (2011).
27 NHGRI press release, "International Consortium Completes Human Genome Project". (Bethesda, MD, 2003).
28 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187-197 (2011).
29 Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**, 1146-1153 (2008).
30 Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724-731 (2005).
31 Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**, 2601-2610 (1979).
32 Staden, R. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* **8**, 3673-3694 (1980).
33 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).
34 Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**, 9748-9753 (2001).
35 Tautz, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* **17**, 6463-6471 (1989).
36 Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626-1632 (2004).
37 Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics : TIG* **17**, 661-669 (2001).
38 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46 (2012).
39 Church, G. M. & Gilbert, W. Genomic sequencing. *Proc Natl Acad Sci U S A* **81**, 1991-1995 (1984).
40 She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930 (2004).
41 Edwards, A. & Caskey, C. T. Closure strategies for random DNA sequencing. *Methods (Orlando)* **3**, 41-47 (1991).
42 Roach, J. C., Boysen, C., Wang, K. & Hood, L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**, 345-353 (1995).
43 Kim, U. J., Shizuya, H., de Jong, P. J., Birren, B. & Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20**, 1083-1085 (1992).
44 Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89**, 8794-8797 (1992).
45 Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319-337 (1987).
46 Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274**, 540-546 (1996).
47 Green, P. Against a whole-genome shotgun. *Genome Res* **7**, 410-417 (1997).
48 Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res* **7**, 401-409 (1997).
49 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).
50 Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res* (2014).
51 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, doi:10.1038/nature13907 (2014).
52 Marziali, A. & Akeson, M. New DNA sequencing methods. *Annual Review of Biomedical Engineering* **3**, 195-223 (2001).

53 Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 335-344 (2004).
54 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145 (2008).
55 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).
56 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732 (2005).
57 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
58 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).
59 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).
60 Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352 (2011).
61 Kircher, M. & Kelso, J. High-throughput DNA sequencing–concepts and limitations. *BioEssays* **32**, 524-536 (2010).
62 Hayden, E. C. Technology: The $1,000 genome. *Nature* **507**, 294-295 (2014).
63 Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat Biotechnol* **30**, 1084-1094 (2012).
64 Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327 (2010).
65 Compeau, P. E., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**, 987-991 (2011).
66 MacCallum, I. *et al.* ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol* **10**, R103 (2009).
67 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**, 1513-1518 (2011).
68 Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nature Methods* **8**, 61-65 (2011).
69 Li, R. *et al.* Somatic point mutations occurring early in development: a monozygotic twin study. *Journal of Medical Genetics* **51**, 28-34 (2014).
70 Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. *Genome Res* **20**, 291-300 (2010).
71 Evans, W. E. & McLeod, H. L. Pharmacogenomics–drug disposition, drug targets, and side effects. *The New England Journal of Medicine* **348**, 538-549 (2003).
72 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
73 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
74 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
75 Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-451 (2011).
76 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
77 Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6**, 99-103 (2009).
78 Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-376 (2011).
79 International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
80 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
81 Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210-223 (2009).

82 Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**, 59-63 (2011).
83 Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3** (2002).
84 Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* **5**, R245-249 (1998).
85 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
86 Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-1428 (2012).
87 Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res* **40** (2012).
88 Warnecke, F. & Hugenholtz, P. Building on basic metagenomics with complementary technologies. *Genome Biol* **8**, 231 (2007).
89 Iverson, V. *et al.* Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* **335**, 587-590 (2012).
90 Carr, R., Shen-Orr, S. S. & Borenstein, E. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLoS Comput Biol* **9** (2013).
91 Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-538 (2013).
92 McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4**, 63-72 (2007).
93 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563-569 (2013).
94 Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24**, 688-696 (2014).
95 Rodrigue, S. *et al.* Whole genome amplification and *de novo* assembly of single bacterial cells. *PLoS One* **4**, e6864 (2009).
96 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
97 Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125 (2013).
98 Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289-293 (2009).
99 Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* **4**, 1339-1346 (2014).
100 Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207-211 (2013).
101 Skloot, R. *The immortal life of Henrietta Lacks*. (Crown Publishers, 2010).
102 Hudson, K. L. & Collins, F. S. Biospecimen policy: Family matters. *Nature* **500**, 141-142 (2013).
103 Duitama, J., Huebsch, T., McEwen, G. K., Suk, E. K. & Hoehe, M. R. in *First ACM International Conference on Bioinformatics and Computational Biology* 160-169 (ACM, 2010).
104 Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272 (2010).
105 Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
106 Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49-54 (2012).
107 Schwartz, D. C. *et al.* Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110-114 (1993).
108 Zhang, Q. *et al.* The genome of *Prunus mume*. *Nature Communications* **3**, 1318 (2012).
109 Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat Biotechnol* **31**, 135-141 (2013).

110 Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* **30**, 771-776 (2012).
111 Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376 (2008).
112 Genome 10k Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity* **100**, 659-674 (2009).
113 Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363-367 (2010).
114 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA* **95**, 14863-14868 (1998).
115 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
116 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-U1040 (2011).
117 Mackay, T. F. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173-178 (2012).
118 Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458-472 (2012).
119 Landry, J. J. *et al.* The genomic and transcriptomic landscape of a HeLa cell line. *G3* **3**, 1213-1224 (2013).
120 Simonis, M. *et al.* High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nature Methods* **6**, 837-842 (2009).
121 Macville, M. *et al.* Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer Research* **59**, 141-150 (1999).
122 Moissiard, G. *et al.* MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* **336**, 1448-1451 (2012).
123 Fraley, C. & Raftery, A. E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* **41**, 578-588 (1998).
124 Jung, J., Park, H., Du, D. Z. & Drake, B. L. A decision criterion for the optimal number of clusters in hierarchical clustering. *J Global Optim* **25**, 91-111 (2003).
125 Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat Biotechnol* **31**, 1143-1147 (2013).
126 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**, 1111-1118 (2013).
127 Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci U S A* (2014).
128 Xin, G., Glawe, D. & Doty, S. L. Characterization of three endophytic, indole-3-acetic acid-producing yeasts occurring in *Populus* trees. *Mycol Res* **113**, 973-980 (2009).
129 Hug, L. A. *et al.* Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**, 22 (2013).
130 Renouf, V., Claisse, O. & Lonvaud-Funel, A. Inventory and monitoring of wine microbial consortia. *Appl Microbiol Biot* **75**, 149-164 (2007).
131 Qin, J. J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-U70 (2010).
132 Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
133 Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *P Natl Acad Sci USA* **105**, 3805-3810 (2008).
134 David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-563 (2014).
135 Saeed, I., Tang, S. L. & Halgamuge, S. K. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* **40** (2012).
136 Mitra, S. *et al.* Analysis of the intestinal microbiota using SOLiD 16S rRNA gene sequencing and SOLiD shotgun sequencing. *BMC Genomics* **14** (2013).
137 Narasingarao, P. *et al. De novo* metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J* **6**, 81-93 (2012).

138 Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437 (2013).
139 Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10** (2009).
140 Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**, 111-120 (2013).
141 Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403 (2013).
142 Umbarger, M. A. *et al.* The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. *Mol Cell* **44**, 252-264 (2011).
143 Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science* **342**, 731-734 (2013).
144 Jarvis, R. A. & Patrick, E. A. Clustering Using a Similarity Measure Based on Shared near Neighbors. *IEEE T Comput* **C-22**, 1025-1034 (1973).
145 Baker, C. R., Tuch, B. B. & Johnson, A. D. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *P Natl Acad Sci USA* **108**, 7493-7498 (2011).
146 Moustafa, A. *et al.* Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One* **5**, e9688 (2010).
147 Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ PrePrints* (2014).
148 Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
149 Gey, G. O., Coffman, W. D. & Kubicek, M. T. Tissue Culture Studies of the Proliferative Capacity of Cervical Carcinoma and Normal Epithelium. *Cancer Research* **12**, 264-265 (1952).
150 Gartler, S. M. Apparent HeLa Cell Contamination of Human Heteroploid Cell Lines. *Nature* **217**, 750-751 (1968).
151 Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **7**, 548 (2011).
152 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
153 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).
154 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
155 Exome Variant Server. http://evs.gs.washington.edu/EVS/ (NHLBI GO Exome Sequencing Project (ESP), January 2012).
156 Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**, 81-94 (2008).
157 The Cancer Genome Project. http://www.sanger.ac.uk/genetics/CGP/ (Wellcome Trust Sanger Institute, January 2013).
158 Goodwin, E. C. *et al.* Rapid induction of senescence in human cervical carcinoma cells. *Proc Natl Acad Sci U S A* **97**, 10978-10983 (2000).
159 Rosty, C. *et al.* Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. *Molecular Cancer* **4**, 15 (2005).
160 Talora, C., Sgroi, D. C., Crum, C. P. & Dotto, G. P. Specific down-modulation of Notch1 signaling in cervical cancer cells is required for sustained HPV-E6/E7 expression and late steps of malignant transformation. *Genes & Development* **16**, 2252-2263 (2002).
161 White, E. A. *et al.* Comprehensive analysis of host cellular interactions with human papillomavirus E6 proteins identifies new E6 binding partners and reflects viral diversity. *Journal of Virology* **86**, 13174-13186 (2012).
162 Corver, W. E. *et al.* Genome-wide allelic state analysis on flow-sorted tumor fractions provides an accurate measure of chromosomal aberrations. *Cancer Research* **68**, 10333-10340 (2008).
163 Wingo, S. N. *et al.* Somatic LKB1 mutations promote cervical cancer progression. *PLoS One* **4**, e5137 (2009).

164 Wistuba, II *et al.* Deletions of chromosome 3p are frequent and early events in the pathogenesis of uterine cervical carcinoma. *Cancer Research* **57**, 3154-3158 (1997).
165 Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**, 51-57 (2011).
166 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
167 Puck, T. T. & Marcus, P. I. A Rapid Method for Viable Cell Titration and Clone Production with Hela Cells in Tissue Culture: The Use of X-Irradiated Cells to Supply Conditioning Factors. *Proc Natl Acad Sci U S A* **41**, 432-437 (1955).
168 Nelson-Rees, W. A., Daniels, D. W. & Flandermeyer, R. R. Cross-contamination of cells in culture. *Science* **212**, 446-452 (1981).
169 Wentzensen, N., Vinokurova, S. & von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Research* **64**, 3878-3884 (2004).
170 Lazo, P. A., DiPaolo, J. A. & Popescu, N. C. Amplification of the integrated viral transforming genes of human papillomavirus 18 and its 5'-flanking cellular sequence located near the *myc* protooncogene in HeLa cells. *Cancer Research* **49**, 4305-4310 (1989).
171 Bouallaga, I., Massicard, S., Yaniv, M. & Thierry, F. An enhanceosome containing the Jun B/Fra-2 heterodimer and the HMG-I(Y) architectural protein controls HPV 18 transcription. *EMBO Reports* **1**, 422-427 (2000).
172 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98 (2012).
173 Peter, M. *et al. MYC* activation associated with the integration of HPV DNA at the *MYC* locus in genital tumors. *Oncogene* **25**, 5985-5993 (2006).
174 Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with *MYC*. *Proc Natl Acad Sci U S A* **107**, 9742-9746 (2010).
175 NHGRI. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcosts/
176 Ajay, S. S., Parker, S. C., Abaan, H. O., Fajardo, K. V. & Margulies, E. H. Accurate and comprehensive sequencing of personal genomes. *Genome Res* **21**, 1498-1505 (2011).
177 PacBio Blog, February 12, 2014. "Data Release: ~54x Long-Read Coverage for PacBio-only *De Novo* Human Genome Assembly". http://blog.pacificbiosciences.com/2014/02/data-release-54x-long-read-coverage-for.html
178 Hayden, E. C. Data from pocket-sized genome sequencer unveiled. *Nature* (2014).
179 Deamer, D. W. & Akeson, M. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends in Biotechnology* **18**, 147-151 (2000).
180 Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotechnol* **30**, 295-296 (2012).
181 Manrao, E. A. *et al.* Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol* **30**, 349-353 (2012).
182 Timmerman, L, March 3, 2014. "PacBio, the Post-Hype Sleeper of Genomics". http://www.xconomy.com/national/2014/03/03/pacbio-the-post-hype-sleeper-of-genomics/
183 Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* **14**, R51 (2013).
184 Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
185 Coupland, P., Chandra, T., Quail, M., Reik, W. & Swerdlow, H. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. *BioTechniques* **53**, 365-372 (2012).
186 Miyamoto, M. *et al.* Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* **15**, 699 (2014).
187 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
188 Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* (2014).

189 Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* **44**, 631-635 (2012).
190 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).
191 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160 (2011).
192 Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nature Methods* **10**, 723-729 (2013).
193 Richter, M. & Rossello-Mora, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126-19131 (2009).
194 Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13**, 358-368 (2003).
195 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955 (2005).
196 Rasko, D. A. *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology* **190**, 6881-6893 (2008).
197 Salipante, S. J. *et al.* Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res* (2014).
198 Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol* **6**, 229 (2005).
199 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
200 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403-410 (1990).
201 van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments : JoVE* (2010).
202 Belloch, C., Barrio, E., Garcia, M. D. & Querol, A. Inter- and intraspecific chromosome pattern variation in the yeast genus Kluyveromyces. *Yeast* **14**, 1341-1354 (1998).
203 Jeffries, T. W. *et al.* Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* **25**, 319-326 (2007).
204 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7** (2011).
205 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).
206 Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**, 1154-1162 (2012).
207 Huang, D.W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44-57 (2009).
208 Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods* **7**, 576-577 (2010).
209 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646 (2010).
210 Talkowski, M. E. *et al.* Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *American Journal of Human Genetics* **88**, 469-481 (2011).
211 Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol* **11**, R119 (2010).
212 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
213 Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-2329 (2011).
214 Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res* **22**, 939-946 (2012).