Comprehensive, precision genomics

Andrew C. Adey

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

Doctor of Philosophy

University of Washington

2014

Reading Committee:
Jay Shendure, MD, PhD., Chair
Evan Eichler, PhD.
R. David Hawkins, PhD.

Program Authorized to Offer Degree:
Molecular and Cellular Biology

University of Washington

Abstract

Comprehensive, precision genomics
Andrew C. Adey

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

The past decade has observed a significant drop in the cost-per-base of DNA sequencing. Driven by a new era of 'next-generation' sequencing (NGS), there has been an explosion of new technologies that utilize DNA sequencing, not just for primary sequence but a wide variety of biological assays. Despite the versatility of NGS, there are a number of drawbacks, including high sample input requirements and short read lengths. Because of the latter, the majority of genome studies cannot resolve haplotype or structural variation which requires long-range information and can play an important role in studying evolution, disease, and is crucial in the *de novo* assembly of genomes. In this dissertation I describe and apply methods to overcome these obstacles. First, I describe a method for the construction of DNA sequencing libraries that utilized a hyperactive transposase to fragment DNA and append universal sequencing primers in a single enzymatic step. This approach reduced the turnaround time from sample to sequencing-ready libraries, and significantly reduced the sample input requirements due to fewer enzymatic steps. I then describe a modified version of the method that allowed for a greater than 100 fold decrease in input requirements for the construction of libraries for the detection of DNA methylation. Next, I discuss a method that utilized the inherent properties of Tn5 transposase to provide long-range sequence information that served as the input for a novel *de novo* genome assembly algorithm. I applied this method to human, mouse, and fly assemblies to produce output scaffolds with contiguity improvements of up to 75 fold with high accuracy. Last, I describe the application of long-range sequence information to haplotype-resolve the genome and epigenome of the aneuploid HeLa cancer cell line. I investigated the global effects of copy number and haplotype on transcript abundance and epigenetic landscape and identified a number of outliers, including haplotype-specific expression of the proto-oncogene *MYC*. I reveal the mechanism responsible for this activation as the complex integration of the HPV-18 viral genome that includes an epithelial-specific enhancer at high copy number 500 kilobasepairs upstream of *MYC* locus.

ACKNOWLEDGEMENTS

Throughout the course of my graduate studies I have had the fortune of working with and being mentored by a number of outstanding individuals. All that I have accomplished these past several years would not have been possible without their guidance and support and I want to take the opportunity to acknowledge their contributions.

First and foremost I want to thank my advisor Jay Shendure. I started graduate school wide-eyed and excited and am finishing just as enthusiastic and optimistic as I was those years ago – something that is not all that common when pursuing a PhD, and something that is in no small part attributed to his mentorship. Jay has been incredibly supportive of even my most far-fetched of ideas yet kept me balanced out by advising me to also pursue some other, safer projects. I could not ask for a better advisor or lab environment that I had the privilege to work in.

I also want to thank my committee who has given me invaluable advice. Evan Eichler, Debbie Nickerson, David Hawkins, and Katie Peichel have been there for me every step of the way and kept me on track to be as successful as possible throughout my studies. I appreciate their honesty and genuine care for my well-being and future as a scientist.

I am incredibly grateful to my fellow Shendure Lab members. The open, collaborative environment has provided a safe, intellectually stimulating setting to do research. I first want to thank Jacob Kitzman who has been an excellent colleague and partner on a number of projects. I also want to thank Brian O'Roak for all the stimulating scientific conversations – somehow we still have not co-authored a paper, but his helpful suggestions have certainly influenced my projects. I want to thank Jerrod Schwartz for some really exciting technology projects that really pushed the limits of what was possible. Matthew Snyder and Martin Kircher for advice on a number of computational and statistical methods; Joshua Burton, Akash Kumar, Joe Hiatt, Steve Salipante, Emily Turner, and Riza Daza, for a number of enjoyable collaborations; Charlie Lee for getting the impossible done on a regular basis – his patience and willingness to accommodate are above and beyond; David Roach for introducing me to my wife as well as some great science talk; and finally Rupali Patwardhan, Ruolan Qiu, Greg Cooper, Jason Klein, Evan

Boyle, Beth Martin, Bethany Stackhouse, Alex Lewis and all other Shendure Lab members – past and present – who have all been extremely helpful throughout these last four years. It has been an incredible place to do exciting work with good friends.

I have also been incredibly fortunate with the environment and collaborators outside the Shendure Lab. I want to thank the Genome Sciences department for providing a constructive environment that welcomed me fully and truly helped me become a better scientist. The Molecular and Cellular Biology program for providing an interdisciplinary setting that has been an invaluable network that has been a valuable resource for several projects. I also want to thank several of my incoming class: Roie Levy for countless conversations on a number of projects and life in general as well as Sam Lancaster, Antonio Abeyta, Hugo Arellano-Santoyo, and Sandrine Boissel for being a great group of colleagues. I also want to thank my collaborators at Epicentre and Illumina who have allowed access to some exciting technologies. My collaborators in the Physics department: Jens Gundlach and his lab members Andrew Laszlo and Ian Derrington for really exciting nanopore work.

I would also like to thank my mentors during my undergraduate studies. Andy Ellington taught me how to carry out well-controlled science, which has served me well over the years. I especially want to thank him for taking a chance on me, believing in me, and letting me start my career in research in his lab. I would also like to thank Scott Hunicke-Smith for allowing me to explore some really fun projects while working as a research scientist.

Last, but far from least, I want to thank my family. My parents have given me so much throughout my life and have always supported me in whatever direction I wanted to go in. They encouraged my curiosity from the very beginning and I cannot thank them enough. I also want to thank my brother, grandparents and other members of my family as well as the family of my wife for their support and love. I finally want to thank my wife Jamie Souhrada. Her devotion to educating young minds is a constant inspiration. Her support and love keeps me going and can put me in the best of moods, even after the most difficult of days. Without her I would not be where I am today.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Opening comments

Life as we know it is built off of a single, universal molecule – DNA – which both acts as a blueprint for the organism as well as provides a means to self-replicate and evolve. In spite of the magnitude of the role DNA plays in biology, it is a relatively simple structure with the majority of the information coded in the form of four distinct bases, the primary order of which dictates the majority of function of the genome, or the entire complement of DNA native to the organism. As such, the ability to determine the sequence of these bases and interpret the function as well as the impact of variation on function is a research area of primary importance in biology that has evolved into the fields of genetics and genomics. Since the discovery of DNA and subsequent confirmation of its role[1] and structure[2], a number of technologies have been developed to sequence DNA. The first wave of these methods[3] culminated in the sequencing of the human genome just over a decade ago[4,5], a project that cost $2.7 billion US dollars (NHGRI). Today, the majority of variation in a human genome can be ascertained for just over $1,000 by the use of "next-generation sequencing" (NGS) technologies; however these new methods rely heavily on the original genome reference as a source of comparison. Beyond primary sequence variation, NGS has been adapted to interrogate another layer of information encoded within cells, often referred to as the epigenome, which can take the form of direct modifications to DNA bases such as methylation, or the alterations to the proteins that package DNA such as histone variants, among many others. Despite our ability to rapidly and cheaply interrogate genomes and epigenomes, there is a long way to go before we can claim a complete understanding of the functional role of these components in biology.

1.2 Organization

In this dissertation I attempt to address major challenges facing the fields of genomics and epigenomics. I first describe a history of DNA sequencing technologies including the development of next-generation sequencing platforms and how they have not only been applied to the interrogation of genomes, but as a means of investigating a host of epigenetic characteristics present within the various tissues of an organism. Despite these advantages, there are a number of shortcomings, two of which I directly address

by this body of work. (I) Conventional, *in vitro* DNA sequencing library construction is a long, laborious process that requires large amounts of input material thus making large-scale experiments and the sequencing of samples with limited quantity excessively difficult and (II) next-generation sequencing platforms are predominantly short read technologies that fail to discern structural variation, haplotypes, or to accurately assemble genomes *de novo*. In Chapter 3 I describe a technology that addresses the first limitation by condensing the bulk of library construction to a single, 5-minute, enzymatic reaction that allows for rapid, low-input library construction. In Chapter 4, I extend this method to interrogate genome-wide DNA methylation marks with greatly reduced sample input requirements when compared to standard methods. In Chapter 5, I describe a method that adapts the previously described library construction approach to serve as a means to acquire long-range sequence information which I then apply using a novel algorithm to *de novo* genome assembly. In Chapter 6, I describe the haplotype resolved genome and epigenome of the aneuploid HeLa cancer cell line. In this chapter I highlight how long-range information acquired by a fosmid-based method was used to identify and reconstruct the architecture of a complex viral integration which *cis* activates a proto-oncogene 500 kilbases away. Lastly, in Chapter 7, I summarize the work presented in this dissertation with a focus on the successes as well as the challenges that still remain. I then describe the future direction of this work and detail a research plan that builds on the successes of previous work to provide a tractable way for addressing the problem of epigenetic heterogeneity within a sample.

Chapter 2 DNA sequencing methods for genome and epigenome interrogation

In this chapter, I first briefly outline the history of DNA sequencing technology and how it drove the field of genomics leading up to the advent of massively parallel DNA sequencing, often referred to as Next-Generation DNA sequencing (NGS). I then discuss how this technology came about and the initial impact it had on the fields of genetics and genomics. In the next section I describe how NGS has been adapted to serve as a proxy to answer a host of biological questions. Despite these advantages, NGS has a number of limitations that prevent the ascertainment of a number of important properties of a genome. I discuss these limitations and then describe the aims of my research to overcome them.

2.1 History of DNA sequencing

Just a year after DNA was confirmed as the conveyor of genetic information by Hershey and Chase in 1952 by the use of bacteriophages[1], James Watson and Francis Crick published their findings on the double helix structure of this molecule that serves as the blueprint of life[2]. Twenty years later the first nucleotide sequences were produced for the RNA coding the coat protein for bacteriophage MS2 in 1972 in the laboratory of Walter Fiers[6]. Shortly after in 1975 Frederick Sanger and Alan Coulson produced an early method of DNA sequencing referred to as "plus-minus" sequencing that involved primer extension[7] – work that clearly influenced Sanger's subsequent development of chain terminator sequencing that was published just two years later[3] and a mere 23 days after a competing method – chemical sequencing[8].

On February 1, 1977 Allan Maxam and Walter Gilbert published their method of DNA sequencing which involved the use of chemical compounds that cleave DNA at either A+G, G, C, or C+T followed by phosphorylation with a radioactive phosphate and gel electrophoresis to read out fragment sizes and allow inference of the underlying DNA sequence[8]. While this approach was a significant improvement, it harbored a number of insurmountable drawbacks, including scalability. Maxam-Gilbert sequencing really only allowed the interrogation of short, pure populations of DNA sequences and additionally required the

use of a number of hazardous chemicals and very difficult procedures that limit the throughput and automation potential of the method.

Later that same month Frederick Sanger published his work on chain terminator sequencing[3], now referred to as Sanger sequencing, that slowly rose to dominance over Maxam-Gilbert sequencing as the primary sequencing technology for the next 30 years and is still heavily used today for its ease, accuracy, and long read length. Briefly, this technology relied on primer extension using a polymerase; however instead of a pure population of dNTPs one of the nucleotides was a mix of predominantly dNTP, but with a small fraction of dideoxy NTPs, or ddNTPs, that terminated the ability for the polymerase to extend. In each of the four reactions (each containing its respective ddNTP) a subset of the population of molecules would cease to extend at any position where that respective base resides in the complement to the template sequence. Subsequent gel electrophoresis then allowed a direct readout of the sequence based on the sizes of terminated chains in each of the four reactions. To initially demonstrate the power of this method, Sanger used his technology to sequence the first complete genome, that of the bacteriophage PhiX in 1977[9]. A number of substantial improvements were developed over the next decade that included a shift from slab gels to capillary electrophoresis[10] as well as a shift from radioactive labeling to fluorescent detection where each of the four ddNTP was instead terminated with a separate fluorescent moiety to allow a single terminator reaction for each sequence which made automation more easily achieved[11]. Sanger sequencing was also the first platform for which sequence read base calling and accuracy measurements were developed to provide a universal classifier of sequence quality[12,13].

2.2 Next-generation DNA sequencing

The broad adoption of Sanger sequencing and the development of increased throughput machines allowed for monumental achievements in the field of genomics; however it also highlighted the fundamental limitations of a technology that requires an individual enzymatic reaction and electrophoresis workflow for each individual read. This constraint motivated the field to develop approaches that are much more parallelizable that generally fell into either “sequencing by synthesis” (SBS) or “sequencing by ligation” (SBL) classifications. Early SBS approaches included pyrosequencing, originally described in

Ronaghi *et. al.* (1996)[14] which utilized the iterative addition of pure populations of individual dNTPs labeled with a pyrophosphate. If primer extension occurred with the respective base, it could be detected via luciferase activity. In spite of this advancement in sequence detection – the problem of parallelization remained. To enable multiple sequencing reactions on the same device, each template must be amplified in a restricted space to enable enough signal for each template to be sequenced as well as for the ability to distinguish between different templates. This problem was initially solved by the implementation of polymerase colonies, or "polonies"[15] which later gave rise to the first SBL platform that was able to sequence the entire genome of an evolved *E. coli* strain at approximately 1/9 of the cost of Sanger sequencing which utilized emulsion PCR as a means of clonal amplification[16]. Shortly after, a commercialization of this technology, the SOLiD sequencing platform, was released with only minor modifications from its original form described in Shendure *et. al.* (2005)[16]. The same year a second emulsion PCR based sequencer was released by 454 Life Sciences that combined emulsion PCR methods with pyrosequencing[17]. The third player in the first wave of next generation sequencing platforms used an alternative method that generated clonal amplified templates by surface-bound primers (a process now called "bridge-PCR") in conjunction with fluorescent, reversible ddNTP nucleotides formed the basis of the Solexa sequencing technology[18], now owned by Illumina Inc.

Since the release of the first three NGS platforms, a number of improvements and variations have come about. These improvements have largely been focused on raw throughput by improving the number of reads as well as the length of sequence that can be obtained per read, with current machines able to produce billions of reads in a single run of the instrument. A subsequent SBS technology that involves semiconductor sequencing on templates generated via emulsion PCR[19] has since displaced the 454 and SOLiD platforms.

Motivated by a push for longer reads and the reduction of bias associated with library amplification, two new single-molecule technologies have entered the scene – both of which have enormous potential to provide a more complete view of a genome. The first of these two utilizes single-molecule real-time sequencing (SMRT) by the incorporation of fluorescent nucleotides by a tethered polymerase positioned

inside of a zero-mode waveguide that allows for fluorescence of only nucleotides undergoing incorporation[20]. The long read lengths produced by SMRT sequencing and the absence of biased PCR amplification in library construction have allowed the reconstruction of extremely difficult regions of genomes[21], as well as the assembly of repeat expansions, such as in patients with fragile X[22]. Additionally, SMRT sequencing has demonstrated the detection of modified bases in DNA, particularly in the form of methyl-adenine and hydroxymethyl-cytosine residues, but with poor detection of methyl-cytosine[23]. In spite of these advantages, the cost-per-base is still far greater than that of short-read SBS platforms, thus making it a less viable option for routine whole-genome sequencing. The second of these single-molecule technologies has produced some proof-of-principle results on the detection of current level changes as nucleotides pass through a nanopore[24]. This method also has the advantages of long read lengths that in an ideal form have no theoretical limit up to the length of the input template and also performs sequencing on native un-amplified DNA. Furthermore, nanopore sequencing has also been shown to detect modifications in DNA in the form of both methyl-cytosine as well as hydroxymethyl-cytosine [25]. The promise of nanopore sequencing is huge, though it is still a technology in its infancy and will likely take some time before it becomes a viable means of sequencing.

2.3 Beyond DNA

Shortly after the advent of NGS and its ability to obtain many millions of sequence reads from an individual experiment, a host of methods were ported over from their microarray form. One of the first of such transitions was RNA-Seq in order to characterize the transcription profile of an organism[26] and was a direct translation from its expression array counterpart. Shortly, other microarray assays followed, including DNA copy number profiling[27], chromatin modifications[28] and DNA methylation[29] to name a few, all of which benefitted from the direct, quantitative nature of DNA sequencing and the ability to interrogate the sample at increasing levels of depth. Since, there has been an explosion of novel methods that utilize the advantages of NGS platforms as proxy to investigate a number of aspects of the genome and epigenome from the functional dissection of regulatory elements[30,31], to the timing of DNA replication[32], or even the three dimensional structure of chromatin[33].

2.3.1 Transcriptome profiling

Perhaps the most natural extension of next generation DNA sequencing platforms is the application to profile the transcriptome, or all RNA transcripts, present in an organism. Up until this point, the most comprehensive approaches to transcriptome profiling involved the use of microarrays. However, expression array technology relied on the use of synthetic probes that hybridize to known, labeled, transcripts[34], thus limiting the scope of RNA transcripts analyzed to those explicitly synthesized on the array. Furthermore, the read-out of expression arrays is the amount of fluorescence observed at each probe, or feature, and thus not a direct measurement of the number of hybridized transcripts present. The implementation of NGS technologies to transcriptomics did not suffer from these limitations and thus allowed the first complete, quantitative view of all RNA transcripts present in an organism[26].

2.3.2 Epigenetic interrogation

Around the same time that NGS methods were being developed for RNA-Seq, other microarray-based epigenetic assays were being adapted to take advantage of the benefits of NGS. One of these was the use of chromatin immunoprecipitation of crosslinked DNA followed by microarray interrogation. These "ChIP-chip" array methods were first utilized to identify the binding sites of cohesins on budding yeast chromosomes[35] and subsequently expanded for interrogating a vast number of transcription factor binding sites and histone modifications in the genome[36]. The application of NGS to ChIP methods allowed for a substantial increase in the resolution of binding sites as well as more quantitative measurements of peak intensities at these sites[37].

A second microarray-based method of epigenetic interrogation was the use of arrays to identify sites of DNA methylation[38]. This approach relied on bisulfite conversion of DNA which converts all non-methylated cytosine residues to uracils[39], a method originally used in conjunction with Sanger sequencing, followed by microarray hybridization. The most significant disadvantage of microarray-based methylation profiling methods was the ability to only interrogate a known set of target sites, a limitation that was eliminated by the implementation of NGS. Since the transition to a high-throughput sequencing

platform, several methods of methylation profiling have risen to high usage that fall into two categories: bisulfite-based, which includes reduced representation bisulfite sequencing[40] and whole genome bisulfite sequencing[41,29]; and enrichment based, which includes meDIP-Seq[42], and other DNA immunoprecipitation strategies[43].

Beyond the more widely utilized microarray methods, a torrent of unique and interesting approaches has been developed to interrogate a host of epigenetic characteristics. These have been extensively utilized by the ENCODE project – an effort to identify and characterize all functional elements in the genome[44]. These methods include all of the former-microarray methods along with assays like DNase hypersensitivity (DHS) sequencing to identify "open" chromatin[45,46], the profiling of replication timing in the genome (Repli-Seq)[32] which allows the isolation of DNA with newly-incorporated bases in a population of synchronized cells, RIP-Seq, which involves the immunoprecipitation of cross-linked ribosomes to identify actively translated transcripts[47], as well as a number of methods to investigate the three-dimensional configuration of the genome within the nucleus by a method known as Hi-C[33]. Outside of ENCODE, a number of novel methods are regularly being released. Some of the more interesting methods involve profiling double-stranded-break sites in the genome by "direct *in situ* breaks labeling, enrichment on streptavidin and next-generation sequencing" or "BLESS"[48], as well as the ability to identify sister-chromatid exchange events by identifying the transmitted strand of each chromosome to daughter cells in a method called "Strand-Seq"[49].

2.3.3 The broad use of NGS as a proxy

The quantitative nature of NGS can be taken advantage of in a number of alternative ways beyond that of the characterization of genomic and epigenomic properties. Many of these strategies involve the concept of DNA barcode counting – or read-counting based methods that can be used to investigate functional or evolutionary effects. For instance, in yeast competition experiments molecular barcodes can be used to "tag" strains such as in the yeast deletion collection[50]. These barcodes were formerly isolated and hybridized to a microarray; however NGS allows for a much more quantitative measurement of the

proportion of each barcode within the population, and thus a more accurate representation of the competition experiment[51].

Another novel implementation of NGS is in the form of "functional dissection". This approach utilizes microarray-derived oligonucleotides that contain all possible mutations of a functional element linked to a DNA barcode. The resulting performance of that functional element, such as a promoter[30] or enhancer[31] can then be assessed by counting the barcodes that are copied via transcription driven by the respective, mutated, functional element. This has allowed an unprecedented level of insight into the functional impact of variants residing within noncoding sequence.

2.4 Limitations of NGS

Despite the versatility, throughput, and low costs associated with NGS, the fact remains that the most cost-effective platforms deliver short read and require a substantial amount of input material and laborious methods to produce a library that is ready to be sequenced.

2.4.1 Library construction throughput bottleneck

Initial methods of NGS library construction relied on a series of enzymatic steps with intervening clean ups. Briefly, this approach involves either mechanical (sonication or nebulization) or enzymatic (endonuclease digestion) fragmentation of sample DNA. This is then followed by end end-repair step that results in blunt-end DNA and then A-tailing by appending a single adenosine residue at the 3' end of each strand. Next, universal adaptor sequences are ligated and adaptor-dimer molecules are separated out via gel-based size selection prior to final PCR amplification. This size selection step is the most limiting on the throughput potential of NGS library construction, with automated systems only being able to process several at a time. The sum of all enzymatic and cleanup steps is approximately 8-10 hours and is difficult to process in batches greater than 8 without the aid of expensive equipment that is not accessible by most research labs. These laborious, time-intensive protocols thus imposed a severe bottleneck on the number of samples that can enter the NGS workflow making experimental designs that involve large

sample counts, a characteristic particularly common for the yeast or other microbial research communities, not feasible.

2.4.2 Sample input requirements and sample heterogeneity

As described in the 2.4.1, the standard protocol for NGS library construction involves a number of enzymatic and clean up steps, each with an associated efficiency, and therefore loss of material. This results in extremely low overall conversion rates from genomic DNA to sequence-ready molecules and therefore input requirements are excessively high. Most protocols for NGS library construction require on the order of 1-3 micrograms of genomic DNA input – an amount that many studies cannot achieve, particularly in the fields of cancer research and developmental biology. Such applications require the analysis of highly-pure populations of cells to limit heterogeneity with respect to underlying DNA in the context of somatic mutations[52], or epigenetic modifications that often differ greatly between cell types[53,54]. This ultimately forces a tradeoff between reducing the amount of input material which results in reduction in final library quality, thus limiting the amount of information obtainable from the sample or sacrificing sample purity to meet sample input requirements, thus resulting in an average signal over the heterogeneous population being analyzed.

2.4.3 Read length limitations: Structural variation and haplotype resolution

The crux of the most cost-effective NGS platforms is that they are short read technologies that fail to resolve structural variation within genomes. While methods have been developed to accurately estimate the underlying copy number of regions of the genome and even discriminate between duplicated regions by use of discriminating variants[27,55], the breakpoints of these regions are often still elusive. Furthermore, the detection of balanced rearrangements such as inversions or translocations is not possible using read depth methods.

Another challenge lies in the fact that many eukaryotic genomes, humans included, are diploid, with two homologous copies of each chromosome, each copy with its own distinct set of variation making up a haplotype. This variation is primarily in the form of single nucleotide variants (SNVs) or short insertions or deletions (indels) that occur at a frequency such that haplotype-discriminating variants are often multiple

kilobases apart and beyond the range of NGS read lengths. This renders the long-range phasing of variants impossible by shotgun sequencing methods with the ability to phase variants only if they reside within 500 bp or so of one another and do not span a repetitive element which are frequent in many eukaryotic genomes.

2.4.4 Read length limitations: *De novo* genome assembly

The same limitation that prevents haplotype-resolution and structural variant identification also plagues the *de novo* assembly of genomes. Nearly all genomes contain repetitive elements create branches in the assembly process that cannot be resolved without sequence information that spans beyond the size of the repeat element. This is particularly challenging when these repeat elements are both frequent, and long as in the case of the human genome[4] with 8.3% of sequence being made up of LTR retrotransposons (up to 5 kbp in length), and 20.4% long interspersed repeats, or LINEs (~6 kbp); additionally 13.1% of the genome can be categorized as short interspersed elements, or SINEs (generally under 500 bp including ~300 bp Alu elements), though many of these can be properly assembled due to their short length[56]. These repeats, and the inability to resolve them using short reads, results in assemblies that are either missing these elements altogether or collapsing a large portion of sequence[57], as well as creating gaps in the assembly that produce separate contigs as opposed to a single contiguous stretch of sequence[58].

2.5 Research aims

In this body of work I aim to address the primary limitations of NGS by developing novel methods of library construction that substantially reduce the preparation time in order to alleviate the throughput bottleneck and cut the sample input requirements by orders-of-magnitude. Furthermore, I aim to pursue methods that interrogate long-range sequence information to work around the read-length limitations that both define and fundamentally hinder the successful utilization of NGS technologies.

2.5.1 Low-input & high-throughput library construction methods for genome and epigenome interrogation

My first aim is to address both the throughput and input requirement limitations that plague NGS workflows and experimental designs for both DNA sequencing and DNA methylation analysis. To accomplish this aim, I develop and extensively characterize a method of library construction that utilizes a hyperactive derivative of the Tn5 transposase. Described in more detail in Chapter 3 and in Adey *et. al.* (2010)[59], this approach condenses the majority of steps in library construction into one single 5 minute reaction that both fragments DNA and appends adaptors in a single step. The second portion of this aim involves the adaptation of transposase-based library construction such that it can be used for the interrogation of DNA methylation in the form of 5-methylcytosine. This method, described in Chapter 4 as well as in Adey *et. al.* (2012)[60], allows for a substantial decrease in the required DNA input mass as well as a reduction in the time required for library construction.

2.5.2 Experimental and computational methods for long-range sequence contiguity

Second, I aim to develop methods that take advantage of the low cost and high throughput of NGS, yet capture the long-range information that gets lost when implementing standard NGS methods. In order to accomplish this, I utilize both molecular methods as well as the development of novel algorithms to acquire sequence information over long distances. I demonstrate that not only can this information be obtained, but can be utilized to resolve complex structural rearrangements, perform haplotype-resolution, and aid in the *de novo* assembly of genomes. These approaches utilize both fosmid-based methods detailed in Kitzman *et. al.* (2011)[61] as well as a novel approach described in Amini *et. al.* (2014, under review)[62], Adey *et. al.* (2014, in review)[63], and in Chapter 5 of this dissertation.

2.5.3 Application of long-range methods to cancer genomics and epigenomics

My third aim of this work is to demonstrate the power that can be achieved when utilizing methods that provide long-range contiguity. I demonstrate this by applying fosmid-based haplotype-resolution technology to produce the first ever haplotype-resolved cancer genome – that of the aneuploid cancer cell

line “HeLa”, which I describe in detail in Chapter 6 and in Adey *et. al.* (2013)[64]. Briefly, I utilize long-range sequence information to decipher a complex viral genome integration and rearrangement in the genome of the cell line. Beyond DNA analysis, I investigate a wealth of epigenetic information that had been previously generated by reanalyzing these data sets in the context of aneuploidy and haplotype. This final aim is an example of what can be achieved when implementing not only long-range technologies, but by integrating multiple data types and reanalyzing that data in the true context of copy number and haplotype in the genome.

Chapter 3 Transposase-mediated construction of DNA sequencing libraries

This chapter is based on the following published paper:

Andrew Adey, **Hilary G Morrison**, **Asan**, **Xu Xun**, Jacob O Kitzman, Emily H Turner, Bethany Stackhouse, Alexandra P MacKenzie, Nicholas C Caruccio, Xiuqing Zhang, and Jay Shendure. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biology*. 2010; 11(12): R119.

Bold face indicates equal contributors.

Hilary Morrison, Asan, Xu Xun, Nicholas Caruccio, Xiuqing Zhang, Jay Shendure, and I conceived the study and planned experiments. I analyzed all data, conducted comparison experiments with the Illumina platform, and developed the PCR-free protocol, the exome capture protocol, and the low-input protocol. Hilary Morrison conducted comparison experiments and accompanying analyses with the 454 platform. Asan, Xu Xun, and Xiuqing Zhang performed YH1 and *Drosophila* genome sequencing, and SNP validation. Jacob Kitzman developed the direct-from-colony library construction protocol and analyzed the resulting data. Alexandra MacKenzie performed exome capture experiments. Emily Turner, Bethany Stackhouse and I developed and tested the 96-plex barcoding protocol. Jay Shendure and I wrote the manuscript. All authors reviewed drafts, contributed comments, and approved the final manuscript.

3.1 Summary

This chapter characterizes and extends a highly efficient method for constructing shotgun fragment libraries in which transposase catalyzes *in vitro* DNA fragmentation and adaptor incorporation simultaneously. This method is then applied to sequencing a human genome and reveals that coverage biases are comparable to those of conventional protocols. The capabilities of the method are also extended by developing protocols for sub-nanogram library construction, exome capture from 50 ng of input DNA, PCR-free and colony PCR library construction, and 96-plex sample indexing.

3.2 Introduction

Massively parallel DNA sequencing methods are rapidly achieving broad adoption by the life sciences research community[65,66]. As the productivity of these platforms continues to grow with hardware and software optimizations, the bottleneck experienced by researchers is increasingly at the front end (the construction of sequencing libraries) and at the back end (data analysis and interpretation) rather than in the sequencing itself.

The input material for commonly used platforms, such as the Illumina Genome Analyzer[18], the Roche (454) Genome Sequencer[17], the Life Technologies SOLiD platform[67], as well as for 'real-time' third-generation sequencers such as Pacific Biosciences[20], consists of complex libraries of genome- or transcriptome-derived DNA fragments flanked by platform-specific adaptors. The standard method for constructing such libraries is entirely *in vitro* and typically includes fragmentation of DNA (mechanical or enzymatic), end-polishing, ligation of adaptor sequences, gel-based size-selection, and PCR amplification (Figure 3.1a). This core protocol may be preceded by additional steps depending on the specific application, such as cDNA synthesis for RNA-seq libraries[26].

Although generally effective, several aspects of the standard method are throughput-limiting or otherwise suboptimal. These include: (1) Labor: there are several labor-intensive enzymatic manipulations with obligate clean-up steps. (2) Time: the protocol requires 6-10 hours from beginning to end, often including an overnight incubation. (3) Automation: although 96-plex, semi-automated processing has been

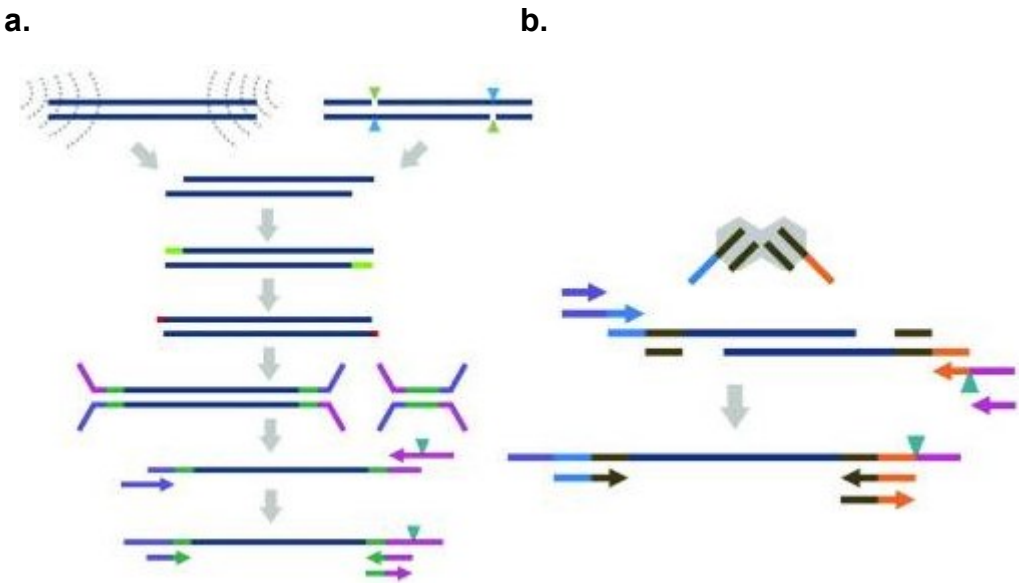


Figure 3.1. Methods for constructing *in vitro* fragment libraries.
a. In the conventional protocol, mechanical or endonuclease fragmentation is followed by end-polishing, A-tailing, adaptor ligation and PCR. **b.** With transposase-mediated adaptor insertion, fragmentation and adaptor insertion occur in a single 5-min *in vitro* step, followed by PCR. For both methods, a primer-embedded sample-specific barcode can be incorporated during PCR amplification (black triangle). Dark blue: Genomic DNA. Light green: End repaired sequence. Red: A-tail. Magenta/dark green and purple/dark green: Adaptors. Mid blue/brown/orange: Transposase adaptors. Cyan/light green triangles: Endonuclease fragmentation. Grey curved dotted lines: Sonication. Grey hexagon: Transposase.

achieved by large-scale genome centers[68], many researchers lack access to the requisite robotic liquid handling systems and/or instruments for parallelized mechanical fragmentation. (4) Sample indexing: incorporation of barcoded adaptors, which enable concurrent analysis of multiple samples and post-sequencing deconvolution, still requires most steps to be carried out on individual samples prior to pooling[69]. (5) High input requirements: standard protocols for shotgun DNA sequencing suggest 1-10 µg DNA as input material per library. This is often not possible, for example in cancer genomics where sample material can be limited. (6) Coverage bias: biases in sequence coverage correlated with G+C content can arise from steps secondary to library construction, including gel purification[70] and PCR amplification[71]. Amplification-free versions of these protocols may reduce G+C biases and eliminate PCR duplicates[71,72], while potentially increasing input requirements.

3.3 Method overview

In the alternative approach that we characterize and extend here, a hyperactive derivative of the Tn5 transposase is used to catalyze *in vitro* integration of synthetic oligonucleotides into target DNA at a high density ('Nextera', Epicentre, Madison, WI, USA). Wild-type Tn5 transposon DNA is flanked by two inverted IS50 elements, each containing two 19 bp sequences required for function (outside end and inside end). A 19 bp hyperactive derivative (mosaic end, ME) is sufficient for transposition provided that the intervening DNA is long enough to allow the two ends to come in close proximity in order to form a complex with a Tn5 transposase homodimer. The relatively low activity of the wild-type Tn5 transposase was cumulatively increased through several classes of mutation[73]. In a classical *in vitro* transposition reaction, hyperactive Tn5 transposomes (hyperactive transposase mutant bound to ME-flanked DNA) bind target DNA and catalyze the insertion of ME-flanked DNA into the target DNA with high frequency[74]. When free synthetic ME adaptors are used instead (isolated from one another, in contrast to ME-flanked DNA in which two ME sequences are linked by the intervening DNA), transposase activity results in fragmentation and end-joining of the synthetic ME adaptor to the 5' end of target DNA. To generate fragment libraries compatible with massively parallel DNA sequencing, limited-cycle PCR is used to append platform-specific primers (Figure 3.1b).

Significant potential advantages of transposase-catalyzed adaptor insertion as a library preparation method, relative to conventional library preparation, include, firstly, many fewer steps, as the fragmentation, polishing, and ligation steps are replaced by a single 5-minute reaction and optional 10-minute pre-PCR clean-up (3.2). Libraries requiring particularly constrained insert size distributions (such as for de novo assembly) may optionally be subjected to chip- or gel-based size selection, increasing preparation time by 1 hour or 3-4 hours, respectively. The second advantage is greatly reduced input requirements while maintaining library complexity. This is expected to be possible because of a more efficient conversion of input DNA into sequencing-compatible material. However, these potential advantages are balanced by the competing concern that transposase-mediated fragmentation will introduce significant sequence-dependent biases relative to conventional library construction.

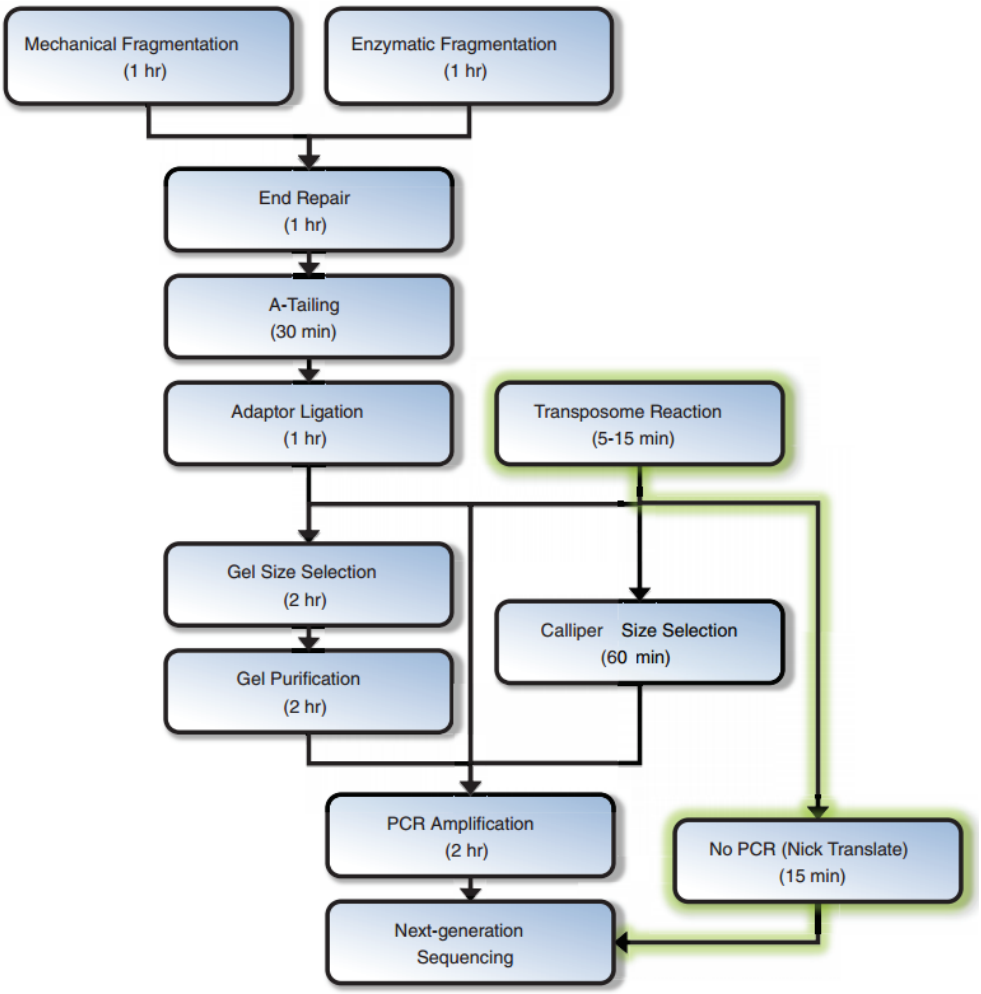


Figure 3.2. Schematic of steps associated with different library preparation methods. Transposase-catalyzed adaptor insertion significantly reduces the number of steps and time associated with library construction (green path).

Here, we report the results of an extensive comparison of transposase-catalyzed fragmentation with standard library construction protocols. We also describe the development of several derivative protocols for transposase-catalyzed fragmentation that significantly extend its capabilities. To evaluate performance with respect to key parameters including sequence-dependent biases, we compared methods across several organisms and sequencing platforms, including whole genome sequencing of a cell line derived from a previously sequenced human, YH1[75], on a single flow-cell with the Illumina HiSeq platform. New protocols reported here that extend the utility of this method include: (1) a 96-plex sample indexing scheme, validated on 96 bacterial genomes; (2) capture and sequencing of the complete coding exon content (exome) from 50 ng of input human genomic DNA; (3) a protocol for the construction and

sequencing of shotgun libraries from as little as 10 pg of starting material; (4) a PCR-free version of the method that mitigates associated G+C biases and decreases the total time for library preparation time to less than 30 minutes; and (5) a method analogous to 'colony PCR' for single-step preparation of genomic sequencing libraries directly from bacterial colonies.

3.4 Results

3.4.1 Comparison of standard versus transposase-based protocols

We performed a side-by-side comparison of three protocols: (1) standard library construction with mechanical fragmentation; (2) standard library construction with time-dependent endonuclease-based fragmentation ('dsDNA fragmentase', NEB); and (3) transposase-catalyzed adaptor insertion ('Nextera', Epicentre). To evaluate performance on the Illumina platform, sequencing libraries and technical replicates were prepared from two genomic DNA samples (Homo sapiens NA18507, Escherichia coli CC118) with each of the three methods. Paired-end, 36 bp reads were generated on an Illumina Genome Analyzer IIx (GAIIx). Reads were mapped using BWA[76] to the E. coli genome (K12) or human genome (GRCh36) as appropriate. To evaluate performance on the Roche (454) platform, sequencing libraries were constructed from two bacteriophage DNAs (CRW10 and PA1) with each of the three methods. Libraries were sequenced on a Roche (454) Genome Sequencer FLX, followed by de novo assembly (gsAssembler) and read mapping (gsMapper) to the appropriate reference genome. A summary of samples processed and sequence data generated on both platforms is provided in Supplementary Table A.2.1.

Sites of mechanical fragmentation, endonuclease fragmentation, and transposase-catalyzed adaptor insertion were characterized by calculating nucleotide composition in the vicinity of the mapping position of the first base of each sequence read (the fragmentation site; Supplementary Fig. A.3.1). This revealed a slight but highly correlated bias for mechanical and endonuclease fragmentation, which suggests that most bias for these two methods is introduced after these protocols converge (for example with A-tailing or adaptor ligation), and that both mechanical fragmentation (here, either acoustic sonication or nebulization) and endonuclease fragmentation (with dsDNA fragmentase) have very low intrinsic biases.

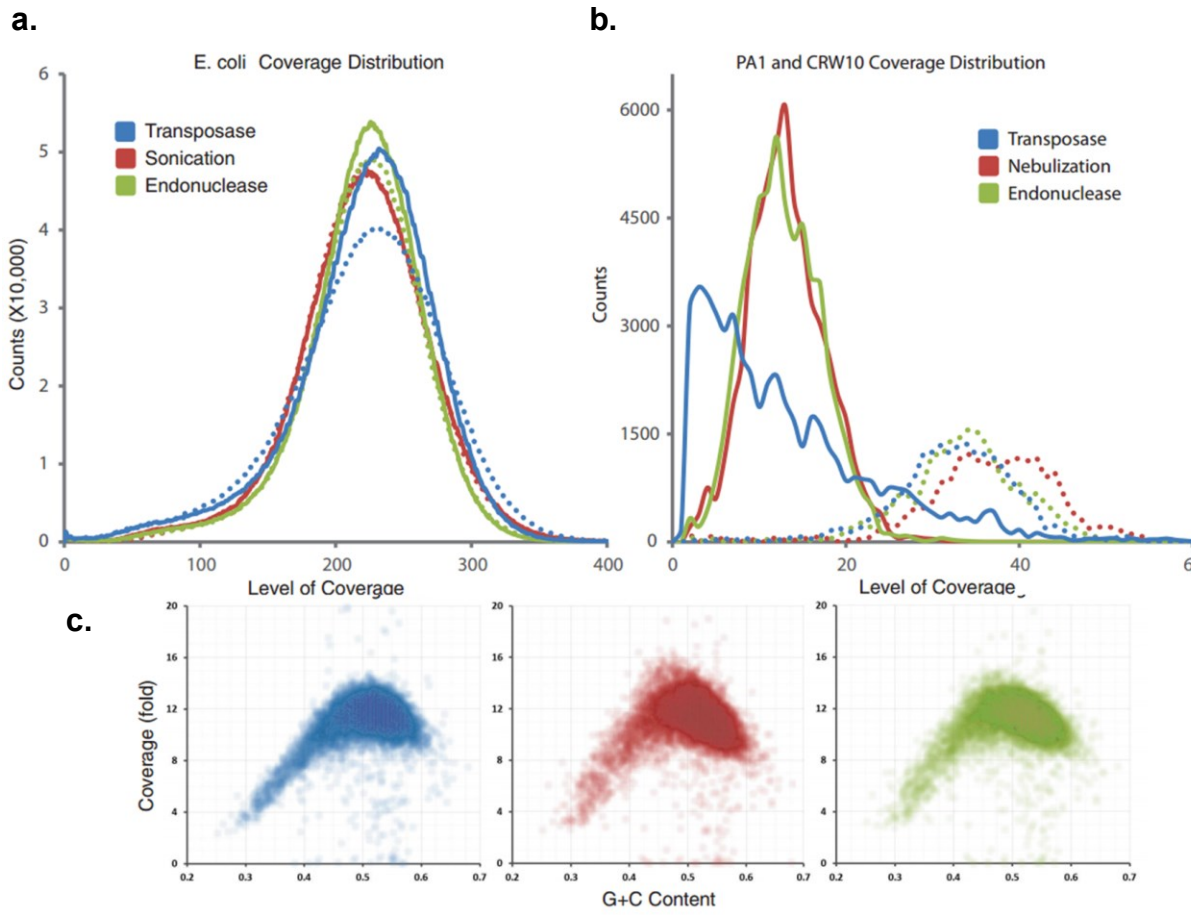


Figure 3.3. Comparison of coverage bias.
a. Coverage distribution across the E. coli genome with transposase (blue), sonication (red), and endonuclease (green) methods (solid lines) and replicates (dotted lines), normalized for total sequencing depth. **b.** Coverage distribution across the PA1 and CRW10 bacteriophage genomes with transposase (blue), nebulization (red), and endonuclease (green) methods (dotted lines represent replicate libraries). **c.** G+C bias for E. coli was assessed by calculating G+C content of the reference in 500 bp bins and plotting the coverage in each for transposase (blue), sonication (red), and endonuclease (green) methods, all of which show an approximately equivalent bias against the extremes.

In contrast, a more extended signature is observed for sites of transposase-catalyzed adaptor insertion, weakly resembling the reported insertion preference of the native Tn5 transposase (AGNTYWRANCT, where N is any nucleotide, R is A or G, W is A or T, and Y is C or T)[77]. However, when calculated in terms of per-position information content, the bias of transposase-catalyzed adaptor insertion is low, and only slightly greater than the other protocols. For E. coli data, maxima of per-position information content over ± 10 bp, on a two-bit scale for fixed positions, are 0.10, 0.11, and 0.16 for mechanical fragmentation,

endonuclease fragmentation, and transposase-catalyzed adaptor insertion, respectively. Average information content over ± 10 bp are 0.0056, 0.018, and 0.049, respectively. Equivalently low information contents were observed for human and phage libraries (Supplementary Table A.2.2). The effective bias associated with transposase-catalyzed adaptor insertion is thus greater than with standard library construction, but only modestly so. For E. coli and human libraries, signatures of bias were consistent in technical replicates for all three methods.

The greater insertion bias is problematic in a practical sense only if it has a significant impact on the distribution of genomic coverage. Consistent with the low calculated information content of the observed biases, the gross distributions of genomic coverage observed for the three methods are very similar (Figure 3.3a, b), the exception being the PA1 bacteriophage library, which may be skewed as a result of sequence context in a relatively small genome. Furthermore, similar biases in coverage are observed for different G+C content bins, with reduced representation at both extremes (Figure 3.3c). As PCR was used to prepare libraries constructed with all three methods, the consistent G+C bias probably arises at that step[71]. We initially predicted that the similar genomic coverage distribution associated with each method was due to factors introduced after the three protocols converge on common steps (solution phase PCR, cluster PCR, and sequencing). However, the correlation in coverage between methods at a per-base level was modest, with transposase-catalyzed adaptor insertion the least correlated with the other methods (Supplementary Table A.2.3).

In this comparative analysis, libraries generated by transposase-catalyzed adaptor insertion were sequenced directly after PCR (without size-selection), and the observed insert size distribution was considerably shorter than the other, size-selected, methods (transposase: 100 ± 47 bp, sonication: 256 ± 48 bp, endonuclease: 244 ± 56 bp; Supplementary Fig. A.3.2). To evaluate whether a lower-bound on insert size exists, tails of long-read (101 bp) pairs were aligned to one another and a mapping-independent size distribution constructed, revealing a sharp decrease at about 35 bp that is probably a secondary consequence of steric hindrance of adjacent, attacking transposases (Figure 3.4). This phenomenon also explains the approximately 10 bp peaks at the lower end of the insert size distribution resulting from the helical pitch of the DNA as it extends away from the transposase.

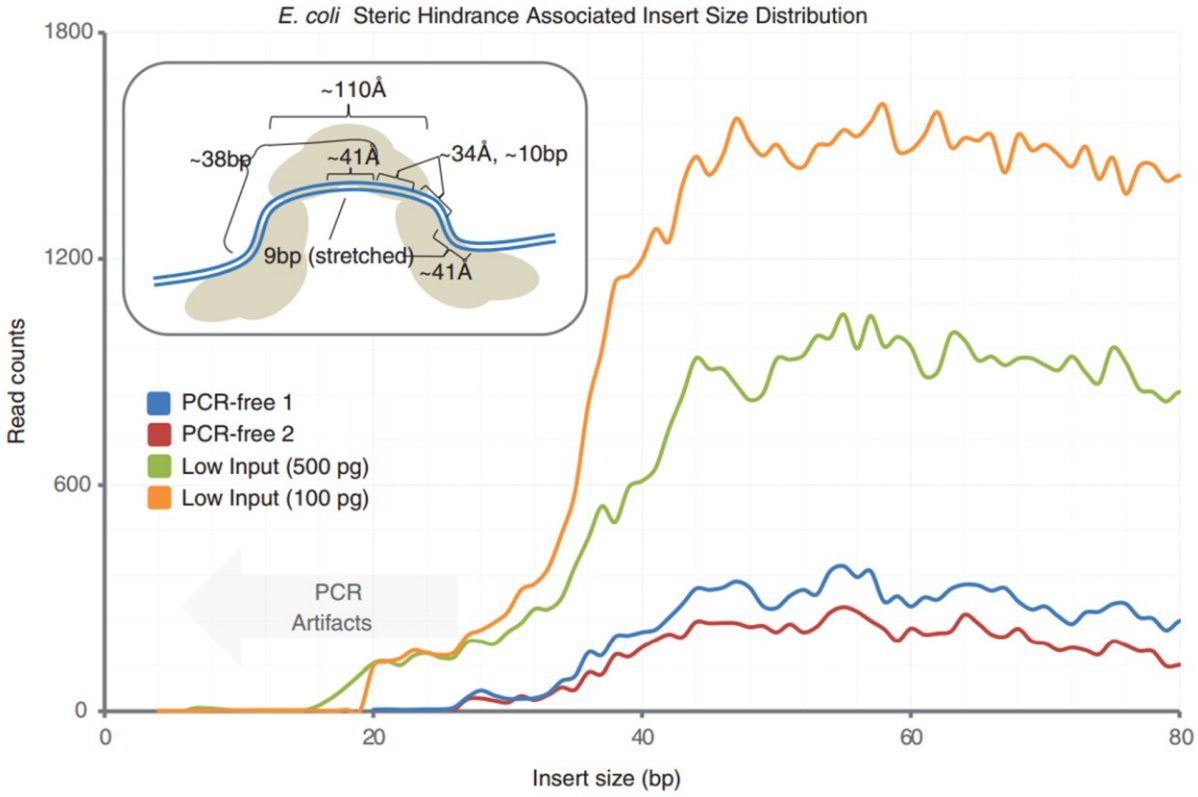


Figure 3.4. Insert size showing steric hindrance.
Insert size was generated from libraries spiked into a paired-end 101 bp run resulting in a large proportion of reads reading into the adaptor sequence. Tails of reads were then aligned to one another to discern the insert size between adaptors, resulting in a mapping-independent insert size at the lower extreme. All reads with an insert size less than 25 bp were PCR artifacts. Inset shows the noticeable drop below 40 bp is consistent with a model for complete saturation of transposition events on a given stretch of DNA. The roughly 110 Å transposase homodimer (grey) is bound to genomic DNA (blue), such that the core of the enzyme acts on a 9 bp region drawn out to 41 Å as well as approximately 10 additional bases of DNA flanking either side (~34 Å each) that are essentially protected from a subsequent transposase attack due to steric hindrance. Since the core region is duplicated in the process, the minimum spacing of transposition events is approximately 38 bp.

With alternative buffer and reaction conditions, other target size ranges can be achieved. For example, the transposon method adapted for Roche (454) library construction resulted in significantly longer fragments (300-800 bp; Supplementary Fig. A.3.3). To assess whether fragment size of libraries generated by transposase-catalyzed adaptor insertion could be constrained without resorting to gel-based size-selection, we evaluated alternative buffer and reaction conditions in combination with different approaches to post-PCR sample clean-up (Supplementary Fig. A.3.4). Notably, an automated chip-based size-selection yielded well-constrained libraries (insert size 162 ± 28 bp).

3.4.2 Whole genome sequencing of human and *Drosophila* genomes

To assess performance further, we conducted whole genome sequencing on transposase-based libraries from *H. sapiens* and *Drosophila melanogaster*. Human genomic DNA from a previously sequenced individual, YH1[75], was used to generate a series of libraries under different reaction conditions and size-selections that were then subjected to seven lanes of paired-end 90 bp (PE90) sequencing on the

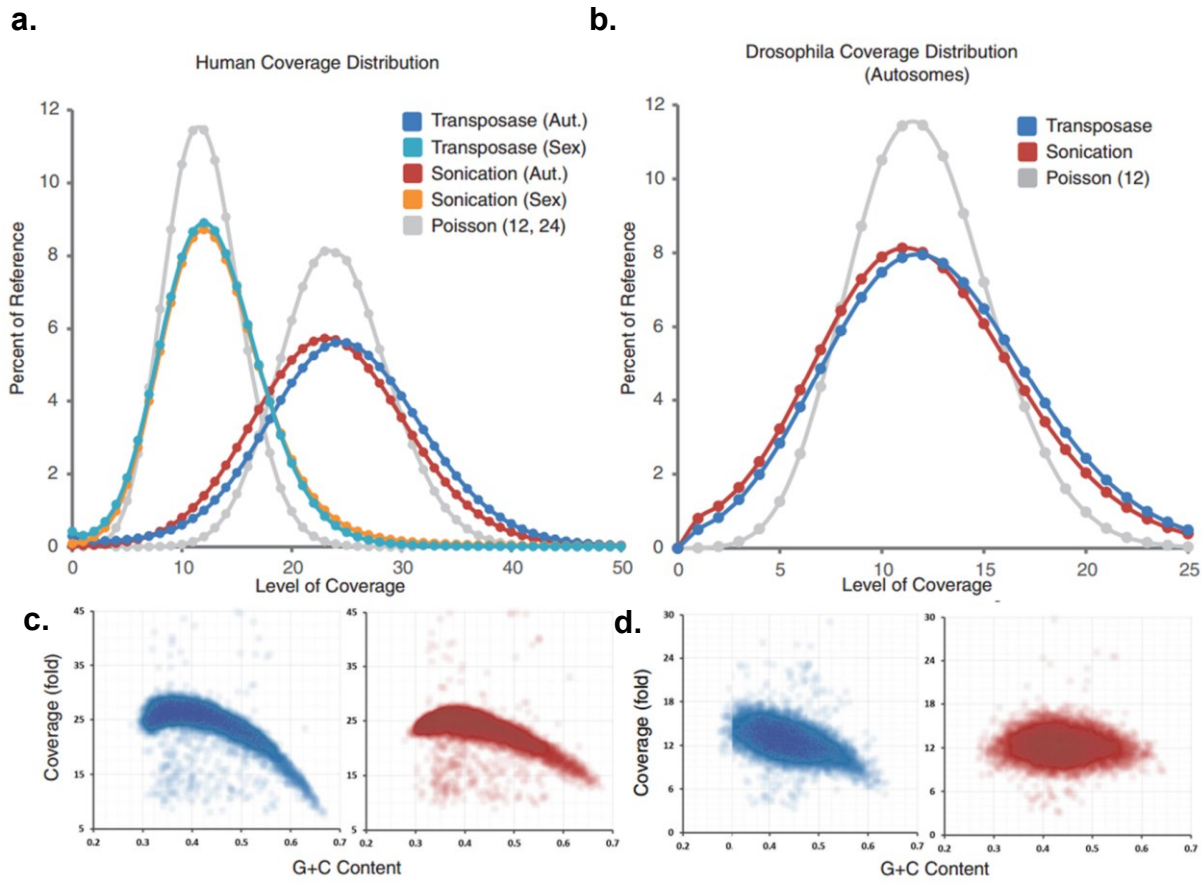


Figure 3.5. Sequence coverage of human and *Drosophila*.
a. Coverage distribution as a percentage of the genome for human (YH1) using transposase (dark blue, autosomes; light blue, sex chromosomes) and sonication (down-sampled to equivalent coverage; red, autosomes; orange, sex chromosomes) methods. Poisson (no bias) distributions (gray) with λ = 12 (sex chromosomes) and λ = 24 (autosomes) are also shown. Poisson distribution is the expected if there were absolutely no bias. (b) Coverage distribution as a percentage of *Drosophila* autosomes using transposase (blue, down-sampled to equivalent coverage) and sonication (*Drosophila* Population Genomics Project (DPGP), red) methods, as well as Poisson distribution with λ = 12 (gray). (c,d) Coverage with respect to G+C content of the reference in 10 kbp or 1 kbp bins for (c) human (YH1) and (d) *Drosophila* genomes respectively, for transposase (blue) and sonication (red) methods at comparable global genomic coverage.

Illumina HiSeq platform. Of 934 million reads, 781 million were mapped[76] to the human genome (GRCh36) for 25× coverage. Although a total of seven libraries were constructed and sequenced to assess reproducibility, the complexity of each individual library was sufficient enough that whole genome sequencing could be carried out using a single library. Variant calling on mapped YH1 data was performed with samtools[78] requiring consensus Q30 at called positions (Supplementary Fig. A.3.5). By these criteria, 3,556,679 SNPs were called (87% in dbSNP129; transition/transversion ratio (Ti/Tv) = 2.07), substantially greater than the 3,074,097 SNPs reported in initial sequencing of YH1. There were 2,922,525 SNPs shared between the analyses (91% in dbSNP129; Ti/Tv = 2.07), 634,154 SNPs unique to our analysis of this genome (70% in dbSNP129; Ti/Tv = 2.08), and 151,572 SNPs unique to the initial analysis of this genome (65% in dbSNP129; Ti/Tv = 1.18). The larger number of SNPs identified here may follow in part from greater mappability with longer read-lengths.

In this analysis, cell-line DNA derived from lymphoblasts was used; however, original sequencing of YH1 by Wang et al. (2008)[75] was carried out on blood DNA. Notably, 4,036 positions were called as mutations in the cell line and as the reference base in blood, both at a high quality score (30) and in uniquely mappable regions of the genome. Of the 1,720 SNPs at a quality over 50 (Ti/Tv = 0.95), a randomly selected 100 were subjected to validation in DNA from blood, DNA from the primary culture used to generate the cell line, and DNA from the cell line. Interestingly, 63 were confirmed as mutations only in the cell line (Ti/Tv = 1.1; one failed assay in primary culture). Of the 37 positions that failed validation, 31 were confirmed as the reference base in blood, primary culture, and cell line (Ti/Tv = 0.48), and the remaining six positions were variant in all three (Ti/Tv = 1.0; one failed assay in primary cell culture). Further experimentation is required to determine whether the validated mutations observed in the cell line represent only mutations occurring during immortalization or propagation of the cell line, or the eventual fixation of somatic mutations present at very low frequencies in primary culture.

Importantly, coverage of YH1's genome in sequencing of libraries derived from transposase-catalyzed fragmentation was relatively uniform when compared with the data generated on this same individual from conventional libraries (Figure 3.5a). The observed GC bias in whole human genome sequencing data from the two methods was comparable (Figure 3.5c); however, a modest decrease (23%) in

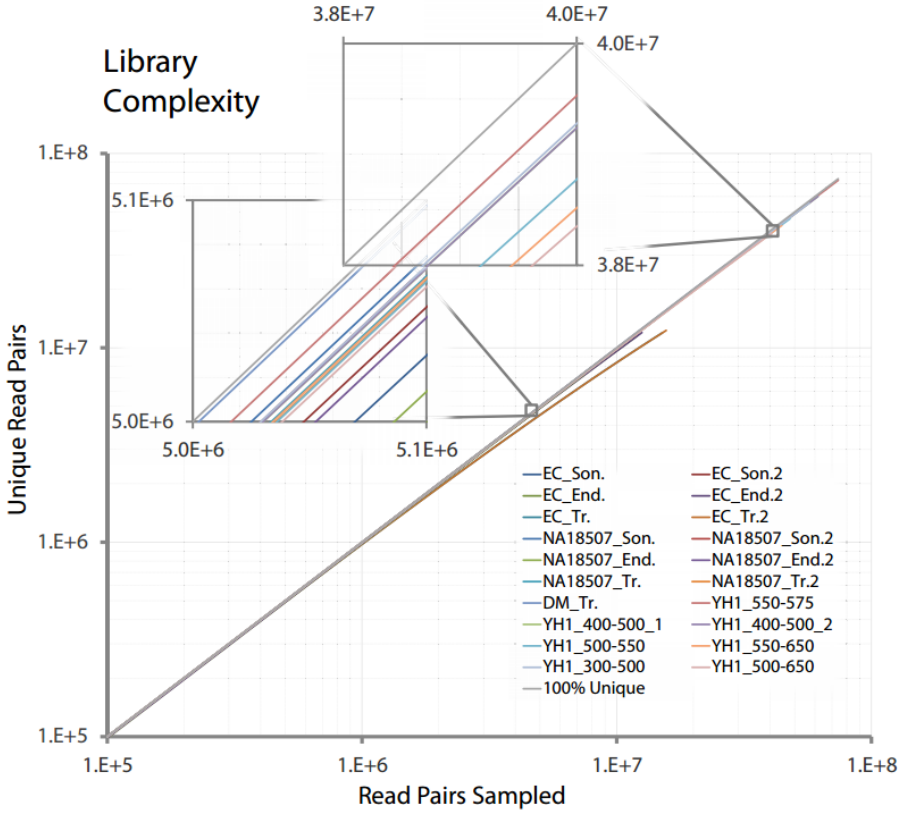


Figure 3.6. Library complexity.
Library complexity for each library shown by incremental, random sampling of 50,000 reads, without replacement, and plotting (on log-log scale) the number of uniquely occurring read-pairs with respect to total number of sampled read-pairs. Species: DM, *Drosophila*; EC, E. coli; NA18507 and YH1, human. Methods for fragmentation: End., endonuclease; Son., sonication; Tr., transposase. Size selection ranges are given for the YH1 libraries (all these were generated using transposase. Libraries ending in "2" are replicate libraries. 100% unique is in gray, i.e. the distribution if there were no duplicates of any sort.

coverage at bins with high GC content (≥60%) was observed with the transposase method. This decrease can potentially be mitigated by a PCR-free version of this method (discussed below), or by alternative PCR conditions (data not shown). For the *Drosophila* genome (dm3), one lane of PE45 sequencing of a transposase-based library on an Illumina GAIIx yielded 16× mean coverage. As with the human genome, the distribution of coverage was largely equivalent to that observed in sequencing *Drosophila* with standard libraries (Figure 3.5b,d) along with the modest decrease in coverage for regions of high GC content. For both human and *Drosophila* genomes, the signature of bias in the vicinity of insertion sites was similar to that observed in the comparative analysis.

Complexity, that is, the number of molecules of distinct origin, is a critical aspect of shotgun library quality, especially for libraries that will be deeply sampled or subjected to further bottlenecking, such as hybrid capture or size selection. Low complexity manifests as an excess of duplicate reads with identical mapping coordinates, which arise from the same progenitor molecule and can thus skew downstream analyses including SNP calling and genotyping. The complexity of each prepared library here was analyzed by incremental sampling of 50,000 read-pairs without replacement and plotting the number of read-pairs sampled versus the number of unique read-pairs (as determined by mapping location) within the sample. In this analysis, the extent of deviation from linearity provides a measurement of sample complexity. All human genome and *Drosophila* genome libraries sequenced here were found to be highly complex, each comprising over 96% unique read-pairs, even as sequencing depths approached 100 million read-pairs, indicating a single such library could be used for whole genome sequencing (Figure 3.6). The high complexities achieved are consistent with a high efficiency of conversion of input mass into sequence-compatible material as compared with conventional library construction. Less complexity was consistently observed for all E. coli libraries, but this is likely to be because we are simply saturating the set of possibilities for start-point pairs that are used to identify PCR duplicates by deep sequencing of a small genome.

3.4.3 Low input targeted sequence capture of the human exome

Exome capture is an increasingly mature technology, but standard protocols require several micrograms of input genomic DNA, which can be problematic when the sample is limiting (for example with tumor samples). We subjected a library from 50 ng human genomic DNA by transposome fragmentation to exome capture (Nimblegen SeqCap EZ Exome probes v1.0). Because the adaptor sequences are different from those in libraries prepared using mechanical shearing, custom blocking oligonucleotides were designed and used. After capture, the library was subjected to pre-sequencing real-time PCR with standard primers followed by sequencing on an Illumina GAIIx (SE36). The resulting reads were aligned to the human genome (GRCh36) with 78% mapping, of which 47% fell within 100 bp of a targeted exon (Supplementary Fig. A.3.6). A direct comparison with an equivalent number of mapped SE36 reads from a standard library, after capture with the same kit, revealed nearly identical complexity for on-target reads

(41% and 43% of an equivalent number of on-target SE36 reads with unique start sites for transposome-based and standard libraries, respectively), as well as comparable uniformity (87% and 82% of target bases covered with ≥1 reads for transposome-based and standard libraries, respectively). However, specificity was notably lower (47% of reads on or near target for transposome-based libraries versus 80% for standard libraries, likely due to the lack of optimization of blocking oligos and hybridization conditions). Nonetheless, we note that the standard protocol has been extensively optimized in the context of production-level scaling, and it is likely that specificity in the capture of transposome-based libraries can also be improved upon. Furthermore, the disadvantage of lower specificity is balanced by the advantage of significantly lower input requirements for genomic DNA entering a targeted capture workflow (50 ng for transposase-based libraries versus 3 µg for standard libraries).

3.4.4 Sub-nanogram library construction

To push the limits on library construction using reduced starting material, E. coli libraries were generated from 500 pg and 100 pg genomic DNA and sequenced as part of a barcode pool. For each library, expected numbers of read counts were observed (0.5 and 0.6 million mapped reads, respectively) without a noticeable drop in complexity (both libraries over 98% at 0.5 million read-pairs), or coverage uniformity. Next, we generated a library from 10 pg human genomic DNA, or roughly three copies of the human genome, which produced over 2 million uniquely mapped read-pairs. Although complexity was reduced because of the significant decrease in progenitor molecules entering PCR, the potential advantages of sequencing from material approaching a single equivalent of the human genome are substantial.

3.4.5 PCR-free library construction

Standard sequencing libraries for the Illumina platform have been generated without the use of PCR amplification in order to reduce associated biases[71,72]. We developed a similar approach for transposase-based methods by including sequences corresponding to the primers used for cluster formation, i.e. the Illumina adaptor sequences, into the adaptors that are added during the transposition reaction, as opposed to incorporating them during PCR (See Appendix A.1). After transposition, a nick translation is performed resulting in Illumina-ready libraries. This method was used to sequence E. coli CC118 and

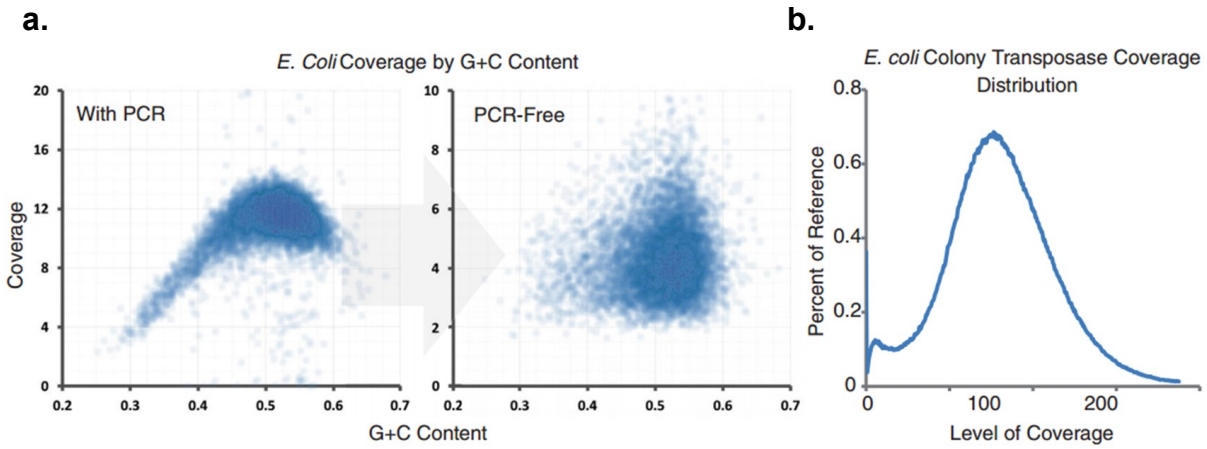


Figure 3.7. PCR-free reduction in G+C coverage bias and direct-from-colony coverage distribution.
a. Coverage with respect to G+C content in E. coli with and without PCR was assessed by calculating G+C content of the reference in 500 bp bins and plotting the coverage in each for transposase after PCR (left) and with the PCR-free method (right). A significant reduction in coverage bias at the extremes of G+C content is observed. **b.** Coverage distribution for E. coli library prepared directly from cell lysate without purification.

human NA18507 with two replicates of each using 100 ng and 200 ng starting material. A noticeable decrease in G+C coverage bias was observed (Figure 3.7a). Furthermore, complexity for each of these libraries was over 98%. The development of PCR-free, transposase-based library construction reduces the full amount of time required for converting DNA to a sequencing-ready shotgun library to less than 30 minutes.

N.4.6 96-plex sample indexing

Second-generation sequencing platforms suffer from poor granularity. For example, in the data described above, a single PE90 lane on the Illumina HiSeq platform yielded 20 Gb of mappable sequence, which is far in excess of what is required for many projects. Sample indexing (or 'barcoding') is a useful solution, but reported protocols still require most steps of library preparation to be carried out on individual samples prior to pooling, and also can suffer from non-uniform performance of individual barcodes[69]. With transposase-catalyzed adaptor insertion, sample indexing could potentially be introduced during adaptor insertion, or during the subsequent PCR step, that is, using a primer-embedded barcode sequence (Figure 3.1b). To evaluate the compatibility of this method with indexing, we attempted the latter

approach. Ninety-six barcodes (9 bp) were designed with a minimal edit distance of four between all pairs and additional constraints on base composition (Supplementary Fig. A.2.4). Performance was evaluated by subjecting DNA from 96 evolved derivatives of Pseudomonas aeruginosa to independent library construction, each with a different barcode-embedded primer during PCR. Post-PCR amplicons were quantified and pooled, followed by several lanes of massively parallel sequencing (PE76, with a third read to collect the 9 bp index). Samples were deconvolved using 9 bp indexes: 92%, 3%, and 3% reads were assigned with 0, 1, and 2 mismatches, respectively, and only 4% could not be unambiguously assigned. With the exception of a few outliers, the distribution of barcode assignments across the 96 was relatively uniform with 90% within a fourfold range (Supplementary Fig. A.3.7), as was the proportion of reads mapping to the reference, illustrating the robustness of the library protocol and the indexing scheme.

3.4.7 Constructing genomic libraries directly from bacterial colonies

In evaluating 96-plex sample indexing for 96-plex bacterial genome sequencing with transposase-catalyzed adaptor insertion, the burden of technical effort shifted from library preparation to the isolation of genomic DNA from each isolate. We speculated that integration of the transposase reaction into a 'colony PCR'-like workflow (in which cells from bacterial colonies are directly mixed into a PCR reaction without DNA isolation) could be used to further simplify library preparation for bacterial genome sequencing (Supplementary Fig. A.3.8). A pipet tip was used to transfer a small number of cells directly from an E. coli colony transformed with pUC19 to a transposase fragmentation reaction (with heat-lysing prior to addition of enzyme). An aliquot of this reaction served as input for PCR amplification without any intervening clean-up step. Sequencing (SE36) yielded 27 million reads, for 170× coverage of the E. coli genome (81% of reads) and 37,000× coverage of pUC19 (10% of reads). The remaining 8% mapped to the F plasmid, or an insert cloned into pUC19, or remained unmapped. Coverage of the E. coli genome was uniform (Figure 3.7b). We propose that direct preparation of genomic sequencing libraries from bacterial colonies with no DNA isolation or intervening purification steps will be useful for rapidly preparing sequencing-ready, indexed fragment libraries from large numbers of bacterial isolates.

3.5 Discussion

Massively parallel sequencing platforms generally require the conversion of genomic DNA (or other nucleic acid sample) into a fragment library that includes common adaptor sequences that are used to mediate clonal amplification and/or the priming of sequencing reactions. Practical limitations of conventional approaches to generating these libraries include high requirements for labor, time and cost, as well as the low efficiency of mass conversion into sequencing-compatible material. Here, we evaluate an alternative approach in which transposase catalyzes the fragmentation of target DNA and insertion of adaptor sequences in a 5-minute, small-volume reaction. The workflow is thus markedly simpler than the conventional approach yielding significant savings in terms of time and labor (summarized in 3.2). The input requirements are also over an order of magnitude lower than what is typically used with standard methods, and we demonstrate that high complexity libraries can be generated from as little as 100 pg of input DNA. Furthermore, input can be reduced to as low as just a few copies of the human genome and still produce a significant amount of sequence data. Taking advantage of the simplicity and low input requirements, we developed a method to construct libraries directly from bacterial colonies without DNA isolation or intervening clean-up steps.

Although there are significant advantages, this method nonetheless has its limitations. First, although there is a significant reduction in required steps and time, preparing very large numbers of libraries is still challenging without some degree of automation. Second, one has relatively limited control over the size distribution of fragmentation. In general, the trend is that the insert size distribution is smaller than desired when reacting to completion, and broader than desired when altering reaction conditions to increase the mean insert size. Third, genomes with high G+C contents show greater bias with this method than with conventional methods, although this is potentially correctable in part by the PCR-free approach or through modified PCR conditions. A related point is that we have also observed that performing transposase-catalyzed adaptor insertion on a single PCR product results in significantly greater bias than with shotgun libraries (J. Hiatt, personal communication), potentially secondary to a high molar concentration of a limited number of possible insertion sites. Fourth, we note that this library preparation does not solve an

ongoing challenge in the field, which is how best to efficiently construct high-complexity, long-insert mate-paired libraries.

A comparison of sequence composition in the vicinity of fragmentation sites identified a bias signature that was larger and more extended than that associated with conventional methods. However, we found that this had little impact at the level of coverage for bacterial, human, and *Drosophila* genomes. Nonetheless, the observed impact may be greater in genomes with compositions that correlate with this bias pattern, or in smaller genomes such as bacteriophage. In our view, for most goals, the disadvantage of the slightly greater bias is offset by the large advantages that we observed with respect to speed, simplicity, and low input requirements.

A further reduction in preparation time was achieved in developing a PCR-free, transposase-based library construction method. This approach also decreased biases resulting from amplification, notably in coverage with respect to G+C content. Looking ahead, we anticipate that the partnership of extremely fast, simplified methods for library construction with third-generation 'real-time' sequencing methods may represent a critical path forward in reducing the time between acquiring biological material and obtaining analyzed sequence data to less than 1 hour.

3.6 Conclusions

Current methods for preparing *in vitro* DNA fragment libraries for massively parallel sequencing are suboptimal for projects involving limited amounts of starting material or large numbers of samples. Here we have characterized an alternative method of library construction in which highly active transposase catalyzes the simultaneous fragmentation and adaptor ligation in a single 5-minute incubation. Comparison with conventional methods of library preparation, relying on mechanical or endonuclease fragmentation, finds that although transposase-catalyzed adaptor insertion demonstrates a slightly greater insertion bias, this has little impact at the level of genomic coverage and is offset by large advantages with respect to speed, simplicity, and low input requirements.

To fully take advantage of the method and expand on transposase-catalyzed adaptor insertion we applied it on a larger scale, including sequencing a human genome on a single flow-cell of a massively parallel

sequencing platform, and validating a 96-plex sample indexing scheme. We also show that transposase-catalyzed adaptor insertion can be integrated with exome sequencing workflows, with the advantage of significantly lower input requirements relative to conventional protocols for targeted sequence capture (50 ng versus several micrograms). In addition, we have generated libraries from as little as 10 pg of starting material and also developed a PCR-free library construction method in order to reduce associated biases and further reduce preparation time to less than 30 minutes. Finally, we demonstrate the direct preparation of genomic sequencing libraries from bacterial colonies with no DNA isolation or intervening purification steps.

3.7 Notes

3.7.1 Data access

All sequence data described here is being deposited in the NCBI Sequence Read Archive (SRA) under accession SRP004087.

3.7.2 Acknowledgements

We would like to acknowledge W Reznikoff for conceptual contributions; H Grunenwald and others at Epicentre for early access to materials; C Lee, J Hiatt, R Patwardhan, L Felker, R Qiu, B O'Roak, and other members of the Shendure Lab for assistance with protocol development and sequencing; the genome sequencing platform of BGI-Shenzhen, especially Wu Kui, Jiang Hui, Chen Mingfeng, Liu Lin, Yu Lili and Kaileung Yuan, for their contributions to YH1 and *Drosophila* genome sequencing; He Weiming, Liu Xingwang and Yang Lei for their genotyping work; and I. Meek and E. Wong-Ho (Caliper) for size selections on the LabChip XT. This work was supported in part by grants from the National Institutes of Health/National Heart Lung and Blood Institute (R01 HL094976 to JS), the National Institutes of Health/National Human Genome Research Institute (R21 HG004749 to JS), the National Institutes of Health/National Institute of Allergy and Infectious Disease Northwest Regional Center of Excellence for Biodefense and Emerging Infectious Diseases at the University of Washington (3U54AI05714), the Ministry of Science and Technology of China, 863 program (2006AA02A301), and an NSF Graduate Research Fellowship (to JOK).

Chapter 4 Application of transposase-mediated library preparation to whole genome bisulfite sequencing

This chapter is based on the following published paper:

Andrew Adey, and Jay Shendure. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Research*. 2012 June; 22(6): 1139–1143.

I performed all experiments and data analysis. Jay Shendure and I designed experiments and wrote the manuscript.

4.1 Summary

This chapter describes how I adapted transposase-based *in vitro* shotgun library construction ("tagmentation") for whole-genome bisulfite sequencing. This method, Tn5mC-seq, enabled a >100-fold reduction in starting material relative to conventional protocols, such that I generated highly complex bisulfite sequencing libraries from as little as 10 ng of input DNA, and ample useful sequences from 1 ng of input DNA. I demonstrated Tn5mC-seq by sequencing the methylome of a human lymphoblastoid cell line to ~8.6× high-quality coverage of each strand.

4.2 Introduction

DNA methylation is a widespread epigenetic modification that plays a pivotal role in the regulation of the genomes of diverse organisms. The most prevalent and widely studied form of DNA methylation in mammalian genomes occurs at the five carbon position of cytosine residues, usually in the context of the CpG dinucleotide. Microarrays, and more recently massively parallel sequencing, have enabled the interrogation of cytosine methylation (5mC) on a genome-wide scale[79]. However, the *in vivo* study of DNA methylation and other epigenetic marks, e.g., in specific cell types or anatomical structures, is sharply limited by the relatively high amount of input material required for contemporary protocols.

Methods for genome-scale interrogation of methylation patterns include several that are preceded by the enrichment of defined subsets of the genome[40,42,80], e.g., reduced representation bisulfite sequencing (RRBS) [40] and anti-methylcytosine DNA immunoprecipitation followed by sequencing (MeDIP-seq)[42]. An advantage of such methods is that they can be performed with limited quantities of starting DNA[81]. However, they are constrained in that they are not truly comprehensive. For example, the digestion-based RRBS method interrogates only ~12% of CpGs, primarily in CpG islands[43], with poor coverage of methylation in gene bodies[82] and elsewhere. Furthermore, RRBS does not target cytosines in the CHG or CHH (H = A,C,T) contexts, which have been shown to be methylated at elevated levels in the early stages of mammalian development[83]. While a small proportion of non-CpG methylation sites can be observed using RRBS, they are restricted to regions within or highly proximal to CpG-islands[84].

The most comprehensive, highest resolution method for detecting 5mC is whole-genome bisulfite sequencing (WGBS)[41,83,43]. Treatment of genomic DNA with sodium bisulfite chemically deaminates cytosines much more rapidly than 5mC, preferentially converting them to uracils[39]. With massively parallel sequencing, these can be detected on a genome-wide scale at single-base-pair resolution. This approach has revealed complex and unexpected methylation patterns and variation, particularly in the CHG and CHH contexts. Furthermore, as the costs of massively parallel sequencing continue to plummet, WGBS is increasingly affordable. However, a key limitation of WGBS is that the current protocols for library construction are based on ligation chemistry and call for 5 µg of genomic DNA as input[41,83,85] which is essentially prohibitive for many samples obtained *in vivo*.

We recently characterized a transposase-based *in vitro* shotgun library construction method ("tagmentation") that allows for construction of sequencing libraries from greatly reduced amounts of DNA (Fig. 4.1a)[59]. Briefly, the method utilizes a hyperactive derivative of the Tn5 transposase loaded with discontinuous synthetic oligonucleotides to simultaneously fragment and append adaptors to genomic DNA. The resulting products are subjected to PCR amplification followed by high-throughput sequencing. The increased efficiency of genomic DNA conversion to viable amplicons and the greatly reduced number of steps allow the construction of low-bias, highly complex libraries from <50 ng of genomic DNA.

4.3 Method overview

Here we describe a modified approach, which we call Tn5mC-seq, which retains the advantages of transposase-based library preparation in the context of WGBS. Because the target of the transposition reaction is double-stranded DNA, whereas bisulfite treatment yields single-stranded DNA, the method was extensively modified such that the tagmentation reaction could take place prior to bisulfite treatment (Fig. 4.1b). First, the adaptors to be incorporated were methylated at all cytosine residues to maintain cytosine identity during bisulfite treatment, with the exception of the 19-bp transposase recognition sequence (in order to minimize differential binding during transposome assembly). Second, an oligonucleotide replacement scheme (Supplementary Fig. B.2.1b)[86] was utilized to ensure that each strand would have adaptors covalently attached to both ends of the molecule. Specifically, this entails

initial transposition with a single adaptor in which the double-stranded transposase recognition sequence is truncated to 16 bp (Tm = 36°C), thereby facilitating its post-incorporation removal by denaturation. A second adaptor is then annealed and the gap repaired, resulting in each strand being covalently flanked by both a 3′ and 5′ adaptor. The fragmented, adapted, double-stranded genomic DNA is then subjected to standard bisulfite treatment for the conversion of unmethylated cytosine to uracil. Degradation during the conversion process likely remains a primary source of loss, but the increased efficiency of the prior steps and the lack of gel-based size selection result in an overall increase in the fraction of DNA that is converted, PCR-amplified, and sequenced.

4.4 Results

4.4.1 Ultra-low-input transposase-based WGBS library performance

We applied Tn5mC-seq to sequence the methylome of a lymphoblastoid cell line (GM20847) using libraries constructed from 1–200 ng of input genomic DNA. Each library was barcoded during PCR amplification and subjected to either a spike-in (5%) or majority (80%–90%) of a lane of sequencing on an Illumina HiSeq2000 (paired-end 100 bp [PE100]; v2 chemistry with custom sequencing primers). These data are summarized in Table 4.1 and Supplementary Fig. B.2.2. In addition, several PCR conditions were investigated to optimize amplification uniformity (Supplementary Fig. B.2.3), as well as a modified protocol (Tn5mC-seq 1.1) (Supplementary Figs. B.2.1d, B.2.4) that eliminates the need for custom sequencing primers and may increase library construction efficiency. Reads were aligned to an in silico converted hg19 (GRCh37) to both the top (C→T) and bottom (G→A) strands using BWA[76] followed by read trimming of unmapped reads and secondary alignment using the same parameters. Unaligned reads typically consisted of low-quality artifacts that likely arose during amplification due to the reduced base complexity of bisulfite converted amplicons.

For each library constructed using ≥10 ng of genomic DNA, over 100 million aligned reads were obtained (60%–75% of total filtered reads; see Methods) of high complexity (90%–97% nonduplicates). Despite the significantly reduced performance of libraries prepared from 1 ng, ~12 million reads were still aligned and the library was of reasonable complexity (78% nonduplicates). Post-alignment reads were merged and

quality filtered for a total of 51.7 Gb of aligned, unique sequence. The average read depth was 8.6× per strand with >96% of CpG and >98% of non-CpG cytosines covered genome-wide (Fig. 4.1c; Supplementary Fig. B.2.2). Because unmethylated nucleotides are incorporated during the gap-repair step (first 9 bp of the second read and last 9 bp before the adaptor as determined by insert size on the first read), the gap-repair regions must be excluded from methylation analysis. However, these bases also serve as an internal control for the conversion rate of the bisulfite treatment. We found this to be >99% for all libraries, and this was independently confirmed using unmethylated lambda DNA spike-ins to two libraries.

For comparison, ligation chemistry–based libraries were constructed using 1000, 100, and 10 ng of GM20847 DNA of the same isolation as the batch used for Tn5mC-seq. These libraries were prepared following the protocols outlined by Lister et al. (2009)[83] with the exception of PCR, which was performed using Kapa Robust due to its higher efficiency over other polymerase choices (Supplementary Fig. B.2.3) During amplification, the 100 and 10 ng preparations did not show significant amplification above a negative control background and were not carried through to sequencing, precluding a comparison of Tn5mC-seq and ligation chemistry–based library construction with identical inputs (a 1000 ng Tn5mC-seq preparation was also not feasible due to the dilute concentrations of the commercially available transposase, which would result in a reduced density of transposition events on a high input mass).

Post-alignment, the 1000 ng ligation chemistry–based library provided slightly more uniform coverage than Tn5mC-seq 1.1 (Supplementary Fig. B.2.1d) libraries constructed from 10 ng, particularly at the lower CpG densities that represent the majority of the genome (Supplementary Fig. B.2.5a). Comparable uniformity was also observed with respect to G+C content as well as for tetramer/pentamer sequence contexts (Supplementary Fig. B.2.5b, c). We also compared the methylation levels of CpGs well-covered by sequencing of libraries corresponding to both methods, and observed good agreement at positions with 5× or greater coverage (r2 = 0.55) as well as 10× or greater coverage (r2 = 0.82) (Supplementary Fig. B.2.5d).

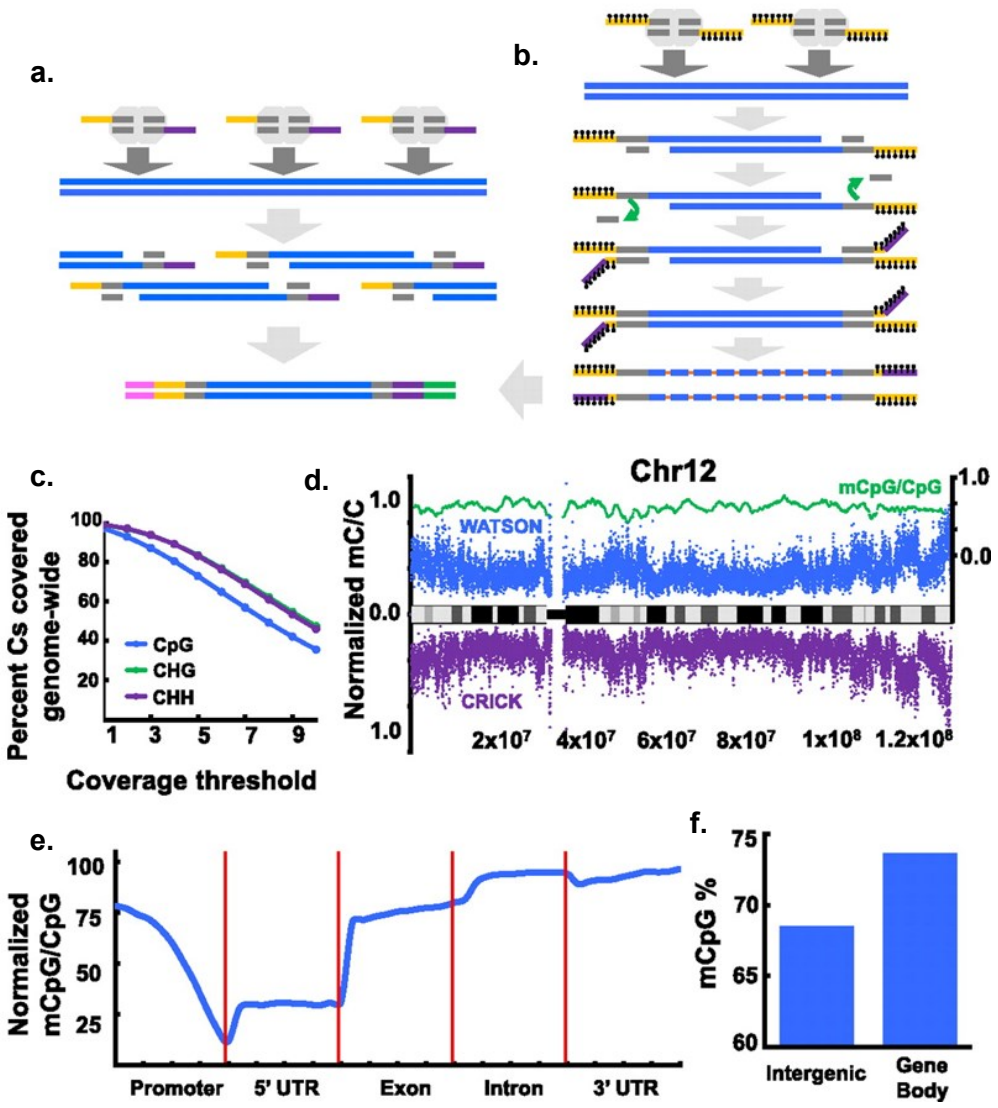


Figure 4.1. The Tn5mC-seq method and resulting methylation profiles.
a. Tagmentation-based DNA-seq library construction. Genomic DNA is attacked by transposase homodimers loaded with synthetic, discontinuous oligos (yellow, purple) that allow for fragmentation and adaptor incorporation in a single step. Subsequent PCR appends outer flowcell-compatible primers (pink, green). **b.** Tn5mC-seq library construction. Loaded transposase attacks genomic DNA with a single methylated adaptor (yellow). An oligo-replacement approach anneals a second methylated adaptor (purple), which is then subject to gap-repair. Bisulfite treatment then converts unmethylated cytosine to uracil (orange) followed by PCR to append outer flowcell-compatible primers (pink, green). Methylation is represented as black lollipops. **c.** Coverage of cytosine positions genome-wide. More than 96% of Cs in all three contexts are covered at least once. Slight decrease in CpG coverage is due to reduced read alignment ability at regions with a high density of methylation. **d.** Normalized methylated cytosine over total cytosine positions in 10-kbp windows across chromosome 12 (blue and purple, left axis), and normalized methylated CpG over total CpG in 100-kbp windows across chromosome 12 (green, right axis). **e.** Normalized methylated CpG over total CpG residues at annotated genic loci. Promoter is defined as 2-kbp region upstream of TSS. **f.** Elevated CpG methylation levels in gene body (intron, exon) compared to intergenic regions.

Name	Input DNA (ng)	Percent Aligning	Percent Unique	Unique Aligned Reads	Mean Insert Size (bp)
Tn5mC-C	200	68	93	127,098,152	198
Tn5mC-D	50	75	90	133,383,834	254
Tn5mC-E*	1	12	76	11,181,960	134
Tn5mC-F*	10	65	95	118,170,302	168
Tn5mC-G*	50	61	97	87,294,793	180
Tn5mC-H	1	11	78	12,393,357	126
Tn5mC-I**	10	62	n/a	29,546,077	n/a
Tn5mC-J	50	71	95	132,144,644	196
TOTAL				**651,213,119**	

*Valve failures in Read 2 resulted in extensive read trimming (50-70bp)

**Complete valve failure on Read 2.

Table 4.1 Summary of Tn5mC-seq libraries and sequencing
Raw reads were initially filtered for instrument valve failures at specific locations of reads and then removal of reads containing over three Ns or extremely low quality bases (phred score <= 2) in the first 50 bases. Alignment was then performed using BWA to in silico converted top and bottom strand references of hg19 (GRC37) followed by trimming and re-alignment. Duplicate reads were identified and removed according to their start position and insert size. The percentage of post-filtering reads that align for each library is shown, as is the percentage of these that are non-duplicates.

4.4.2 Lymphoblastoid cell line methylation

We were able to detect ~46 million 5mC positions (1% FDR; see Methods), accounting for 4.2% of total cytosines with coverage. The majority of methylation observed was in the CpG context (97.1%), and the global CpG methylation level was 69.1%. This level is similar to that of the fetal fibroblast cell line IMR90 sequenced by Lister et al. (2009; 67.7%)[83] and is consistent with the observation that CpG methylation levels are reduced in differentiated cell types. Additionally, CHG and CHH methylation levels were substantially lower than in ES cells, at 0.36% and 0.37%, respectively, again consistent with the differentiated cell type. On the chromosome scale, the methylation density correlated with banding patterns and increasing levels were observed extending distally through subtelomeric regions (Fig. 4.1d). An analysis of functionally annotated genic regions revealed a sharp decrease in CpG methylation through the promoter region followed by a minor increase in the 5′ UTR and then elevated levels of methylation throughout the gene body, particularly at introns (Fig. 4.1e,f), consistent with previously described CpG methylation profiles[83].

4.5 Discussion

We developed Tn5mC-seq as a novel method for rapidly preparing complex, shotgun bisulfite sequencing libraries for WGBS. In brief, the method utilizes a hyperactive Tn5 transposase derivative to fragment genomic DNA and append adaptors in a single step, as previously characterized for the construction of DNA-seq libraries[59]. In order for library molecules to withstand bisulfite treatment, the adaptors are methylated at all cytosine residues, and an oligonucleotide replacement strategy is employed to make each single-strand covalently flanked by adaptors. The high efficiency of the transposase and overall reduction in loss-associated steps permits construction of high-quality bisulfite sequencing libraries from as little as 10 ng that are comparable to ligation chemistry–based libraries generated from 100× more DNA, as well as useful sequence from 1 ng of input DNA. Additionally, the increased efficiency of transposase-mediated library construction may allow for preparation of WGBS libraries from poor-quality or degraded DNA samples.

Our results illustrate how derivatives of the transposase-based method for DNA-seq library preparation can enable key applications of next-generation sequencing where its advantages are perhaps even more relevant. The ability to generate such libraries from very low amounts of input genomic DNA substantially improves the practicality of whole methylome sequencing and removes a key advantage of less encompassing methods such as RRBS[40,43]. Specifically, low-input WGBS with Tn5mC-seq may make possible the comprehensive interrogation of methylation in many contexts where DNA quantity is a bottleneck, e.g., developing anatomical structures, microdissected tissues, or pathologies such as cancer, where the epigenetic landscape is of interest but tissue quantity limits high-resolution WGBS.

4.6 Notes

4.6.1 Data access

The sequence data presented in this study have been submitted to the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under accession no. SRP011746.

4.6.2 Acknowledgements

We thank Cholie (Charlie) Lee for performing all sequencing runs and the Shendure laboratory for helpful discussions. We also thank Nick Caruccio, Haiying Gruenwald, Brad Baas, and Igor Goryshin from Epicentre (Illumina) for help and ideas regarding transposase-based library preparation, as well as Eric Van Der Walt and colleagues at Kapa Biosystems for early access to reagents and protocols for library amplification. A.A. is funded by an NSF Graduate Research Fellowship. This work was supported in part by the Lowell Milken Prostate Cancer Foundation Young Investigator Award (J.S.).

Chapter 5 Transposase-based long-range contiguity applied to *de novo* genome assembly

This chapter is based on the following paper that is currently in preparation:

Andrew Adey, Jacob O. Kitzman, Joshua N. Burton, Riza Daza, Akash Kumar, Lena Christiansen, Mostafa Ronaghi, Sasan Amini, Kevin Gunderson, Frank J. Steemers, Jay Shendure. *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity.

I developed and implemented the algorithm behind the use of transposase contiguity for *de novo* genome assembly. Jacob Kitzman, Sasan Amini, Frank Steemers, Kevin Gunderson, Mostafa Ronaghi, Jay Shendure, and I developed the transposase contiguity method. Joshua Burton provided advice on the algorithmic details and performed the Hi-C based assembly stages. Riza Daza and Lena Christiansen performed transposase contiguity experiments, Akash Kumar performed long fragment read library construction. Jay Shendure and I wrote the manuscript and all authors reviewed it.

This chapter contains an additional section describing a novel algorithm I developed for the use of TC-Seq data for haplotype-resolution that is not present in the manuscript.

5.1 Summary

We developed a method that exploits transposase-based contiguity-preserving sequencing (TC-Seq) for *de novo* genome assembly. Our approach scaffolding, termed *fragScaff*, leverages coincidences between thousands of *in vitro* fragment pools as a source of contiguity information, increasing the scaffold N50 of *de novo* assemblies by up to 75-fold with high accuracy. As a proof-of-concept, we apply TC-Seq and *fragScaff* to substantially boost the contiguity of *de novo* assemblies of the human, mouse, and fly genomes. We also demonstrate that *fragScaff* is complementary to Hi-C-based contact probability maps, providing mid-range contiguity to support robust, accurate chromosome-scale *de novo* genome assemblies.

5.2 Introduction

The broad adoption of massively parallel sequencing (MPS) has resulted in a boom of *de novo* genome assemblies[87]. For the most part, these assemblies are of far lower quality than the human reference genome produced by the Human Genome Project (HGP)[4,5,88,89,90,91], largely due to a dearth of readily accessible MPS-based methods for generating mid-range and long-range contiguity information. Current next-generation assemblies rely on deep sequencing of shotgun fragment and 3 kbp mate-pair libraries to provide short-range contiguity, both of which are generated via simple, *in vitro* protocols. However, mid-range contiguity information requires the use of laborious, *in vivo* fosmid-end sequencing libraries[92], or fosmid pool sequencing[93], with the utility of these methods for *de novo* assembly closely tied to clone library complexity. Long-range contiguity information can be generated via contact probability maps (CPM), wherein Hi-C (three dimensional chromatin confirmation capture) read-pairs are used for chromosome-scale scaffolding[94]. However, the performance of CPM depends on input assembly scaffold size, with optimal results requiring an N50 of ~200 kbp or greater, a level of contiguity that assemblies based only on shotgun fragment and 3 kbp libraries usually fail to achieve [88,89,90,91]. As such, there remains a strong need for robust *in vitro* methods to capture mid-range contiguity information for *de novo* genome assembly.

5.3 Method overview

Transposase-mediated library construction, or "tagmentation", utilizes a hyperactive Tn5 transposase to both fragment and append universal adaptors in a single enzymatic step[95]. In recent years, tagmentation has been applied in diverse ways including for DNA-Seq[95], stranded RNA-Seq[86], whole-genome bisulfite sequencing[60], chromatin profiling[96] and *in situ* mate-pair library preparation directly on a sequencing flowcell[97]. We recently demonstrated a novel method, "Transposase Contiguity Sequencing" (TC-Seq) for haplotype-resolved genome sequencing[62]. TC-Seq utilizes an inherent property of the Tn5 transposase in which the enzyme remains tightly bound to the target DNA after tagmentation occurs, physically linking adjacent library molecules (Fig. 5.1a). Prior to PCR amplification, the high molecular weight stretches of linked templates are subjected to sub-haploid dilution and compartmentalization, followed by protein denaturation in order to free the templates for amplification. To increase the number of compartments, a two-tiered indexing approach is applied (Fig. 5.1b). The initial tagmentation is performed using 96 uniquely indexed transposase-adaptor complexes. The high molecular weight linked templates are then pooled, followed by limiting dilution into 96 indexed PCR reactions such that each PCR well contains templates from 96 originating transposase reactions, thus producing 96 × 96 = 9,216 distinct index combinations. The resulting alignments from each pool consist of clusters of read pairs that originated from the same high molecular weight transposed template fragment (Fig. 5.1c), analogous to fosmid[61] or *in vitro* [98] dilution pools but with a much higher effective number of pools.

5.4 Results

5.4.1 Scaffolding human assemblies with TC-Seq and *fragScaff*

We speculated that the large number of effective dilution pools in TC-Seq (n = 9,216), each of which contains thousands of long DNA fragments (Fig. 5.1d), might be an excellent source of mid-range contiguity information for *de novo* genome assembly. We therefore developed an algorithm, *fragScaff*, which determines shared pool fractions at the ends of each input contig or scaffold (iSC) in TC-Seq data as a means of linking proximal sequence for further *de novo* assembly (Fig 5.2a-c). We applied *fragScaff*

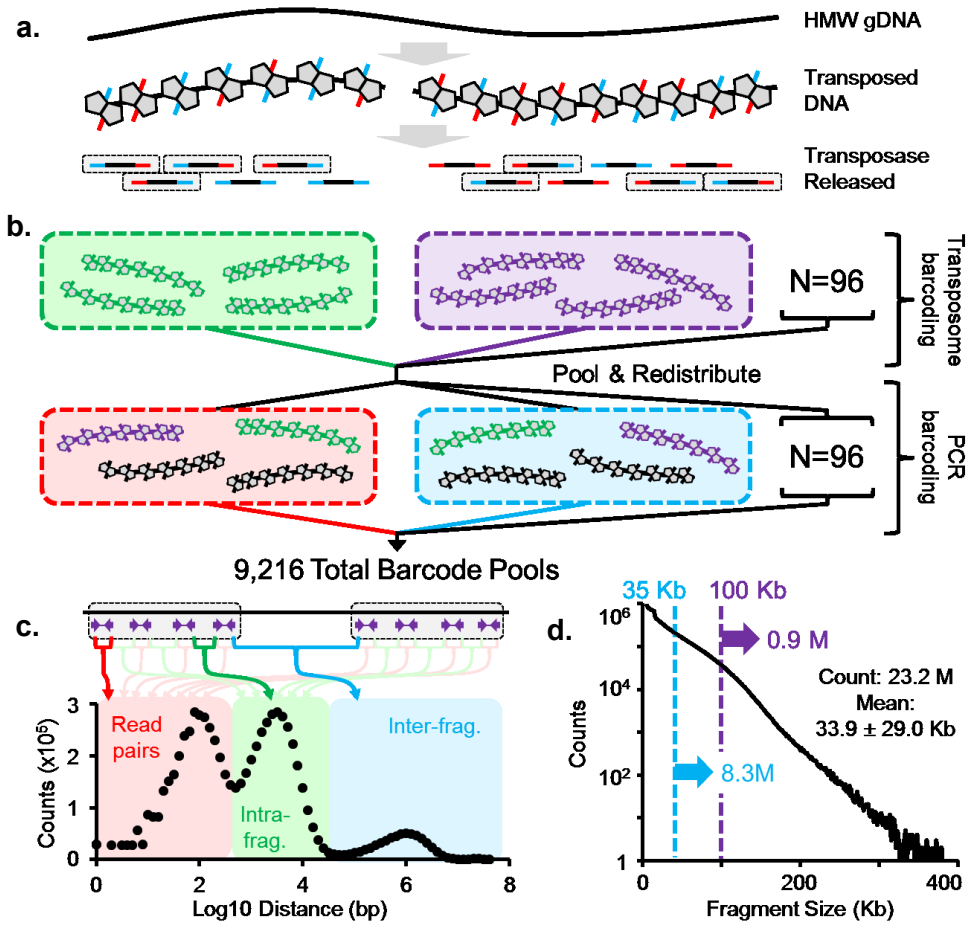


Figure 5.1. TC-Seq method and performance.
a. High Molecular Weight (HMW) genomic DNA is reacted with hyperactive Tn5 transposase loaded with universal adaptors. The transposase complex stays bound to the target DNA after completion of the reaction resulting in physical linking of adjacent template molecules. After transposase is removed, PCR amplification of viable templates (gray boxes) can be performed. **b.** 96-plex indexed tagmentation is performed followed by pooling and limiting dilution into 96-plex indexed PCR, thus allowing 96 × 96 = 9,216 indexed compartments. **c.** Tri-modal sequential read alignment distance distribution of TC-Seq. The first peak (~100 bp, red) corresponds to the read pair insert size, the second (~3.2 kbp, green) corresponds alignment distances between separate pairs of reads in the same long fragment (gray boxes), and the third (~1 Mb, blue) corresponds to alignment distances between separate pairs of reads at the ends of two different subsequent long fragments. **d.** Distribution of called fragment lengths for GM12878 TC-Seq.

to a human (GM12878) *de novo* assembly with an N50 scaffold size of 437 kbp generated by *ALLPATHS-LG* using only shotgun and 3 kbp mate-pair libraries previously produced by Gnerre *et. al.* (2011)[92], to exploit TC-Seq data generated on GM12878 DNA for further scaffolding (9,216 pools; mean 2,513 fragments ≥ 5 kbp per pool; 23.2 M fragments total). The resulting *fragScaff* assembly incorporated 98.2% of sum length of the iSC's, resulting in an N50 of 5,496 kbp (13× increase) with a 97.2% join accuracy, a 90.5% orientation accuracy, and 98.7% of base pairs properly placed (Table 5.1,

a

Organism	Input: Sequence used	Input: N50 (Kb)	Input: Scaffold count	fragScaff: Method	fragScaff: N50 (Kb)	fragScaff: N50 Imp.	fragScaff: Scaffold count	fragScaff: Placement Accuracy
Human	S*	47	127,088	TC-Seq	1,615	34×	27,619	99.4
Human	S + 3Kb	437	18,921	TC-Seq	5,469	13×	7,145	98.7
Human	S + 3Kb	437	18,921	TC-Seq	27,440	63×	5,357	97.5
Human	S + 3Kb	437	18,921	Fosmid	567	1.3×	15,303	88.0
Human	S + 3Kb	437	18,921	LFR	668	1.5×	14,476	60.5
Human	R, 15Kb	15	191,312	TC-Seq	361	24×	36,790	99.5
Human	R, 100Kb	100	28,817	TC-Seq	6,601	66×	4,223	99.3
Mouse	S + 3Kb	224	25,964	TC-Seq	1,414	6.3×	6,196	98.8
Fly	S	68	7,109	TC-Seq	459	6.8×	4,024	99.9

b

Organism	Input: Initial Assembly	Input: N50 (Kb)	CPM Only: Percent Clustered	CPM Only: Percent Ordered	fragScaff + CPM: Percent Clustered	fragScaff + CPM: Percent Ordered
Human	S*	47	89.7	41.5	99.4	99.1
Human	S + 3Kb	437	98.2	94.4	98.8	96.0
Human	R, 15Kb	15	35.9	0.2	91.9	88.2
Human	R, 25Kb	25	28.9	0.4	92.5	93.1
Human	R, 50Kb	50	70.8	16.5	93.6	95.5
Mouse	S + 3Kb	224	98.0	86.7	99.8	98.6
Fly	S	68	81.2	82.0	96.2	93.0

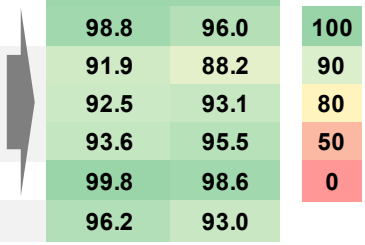


Table 5.1 *fragScaff* assembly improvements
a. Scaffolding improvements using *fragScaff*. Input assemblies were constructed using *ALLPATHS-LG* using shotgun sequence data (S) and in some cases 3 kbp mate-pair sequence data (3kbp), or simulated by segmenting the human reference genome (R, size). **b.** Performance of CPM on input assemblies with and without TC-Seq and *fragScaff* scaffolding prior to CPM. Improvement is most significant for smaller input assemblies.

Supplementary Table C.2.1, Supplementary Fig. C.3.1-3; performance metrics defined in Supplementary Methods C.1) including a perfectly ordered scaffold of 23.9 Mb. Additionally, we performed several less stringent iterations of *fragScaff* which resulted in N50 increases up to 63× (N50 = 27,440 kbp) while still maintaining a 97.5% base pair placement accuracy.

5.4.2 Scaffolding using fosmid or long-fragment-reads with *fragScaff*

We also attempted to apply *fragScaff* to data generated with fosmid[61] or long fragment read[98] (LFR) dilution pools. However, this resulted in low link counts of poor accuracy (N50 improvement: 1.3×, and 1.5×; join accuracy: 71.6% and 34.0%; sequence joined: 38.9% and 46.6% for fosmid and LFR respectively), predominantly due to the reduced pool counts (fosmid: 288, LFR: 96 vs. TC-Seq: 9,216),

which restrict the number of coincidences between pool content on which this method is based. Consistent with this, a marked performance decrease is also observed when down-sampling to a reduced number of TC-Seq pools (e.g. 47.9% (288 pools) vs 87.9% (9,216 pools) of sequence scaffolded for fly; see below).

5.4.3 Application of TC-Seq with *fragScaff* to short, simulated and nonhuman assemblies

A 437 kbp input assembly N50 is quite large compared to a typical *de novo* assembly of a complex genome based solely of shotgun and 3 kbp libraries[88,89,90,91]. Therefore we fragmented the iSC's at every gap to produce a contig-only assembly (N50 of 47 kbp) and then applied *fragScaff* with TC-Seq data. This resulted in an assembly that joined 97.7% of input contig sequence for an output scaffold N50 of 1,615 kbp (34× improvement) and a 97.6% join accuracy with 98.8% of base pairs properly placed. However, the reduced size of input contigs and exclusion of 3 kbp mate-pair data resulted in a substantial decrease in orientation accuracy to 68.0%. To further assess dependence of *fragScaff* on input contig N50, we fragmented the human reference genome *in silico* to sizes ranging from 15 to 300 kbp (Table 5.1a, Supplementary Table C.3.1). The scaffolding of these simulated input assemblies resulted in N50 improvements ranging from 24× to 75×, with greater than 99.3% of base pairs properly placed and 96.0% properly oriented in every assembly.

To evaluate *fragScaff* in additional contexts wherein a high quality reference genome is available for comparison, we also constructed TC-Seq libraries for *Mus musculus* (mouse) and *Drosophila melanogaster* (fruit fly). For both organisms, we used assemblies generated using *ALLPATHS-LG* from either shotgun and mate-pair sequencing (mouse, N50 = 224 kbp) or just shotgun sequencing (fly, N50 = 68 kbp)[92,94]. TC-Seq data and *fragScaff* were then used to increase the N50 to 1,414 kbp for mouse (6.3× improvement) and 459 kbp for fly (6.8× improvement) with 98.8% and 99.9% of base pairs properly placed for mouse and fruit fly respectively (Table 5.1a, Supplementary Table C.3.1).

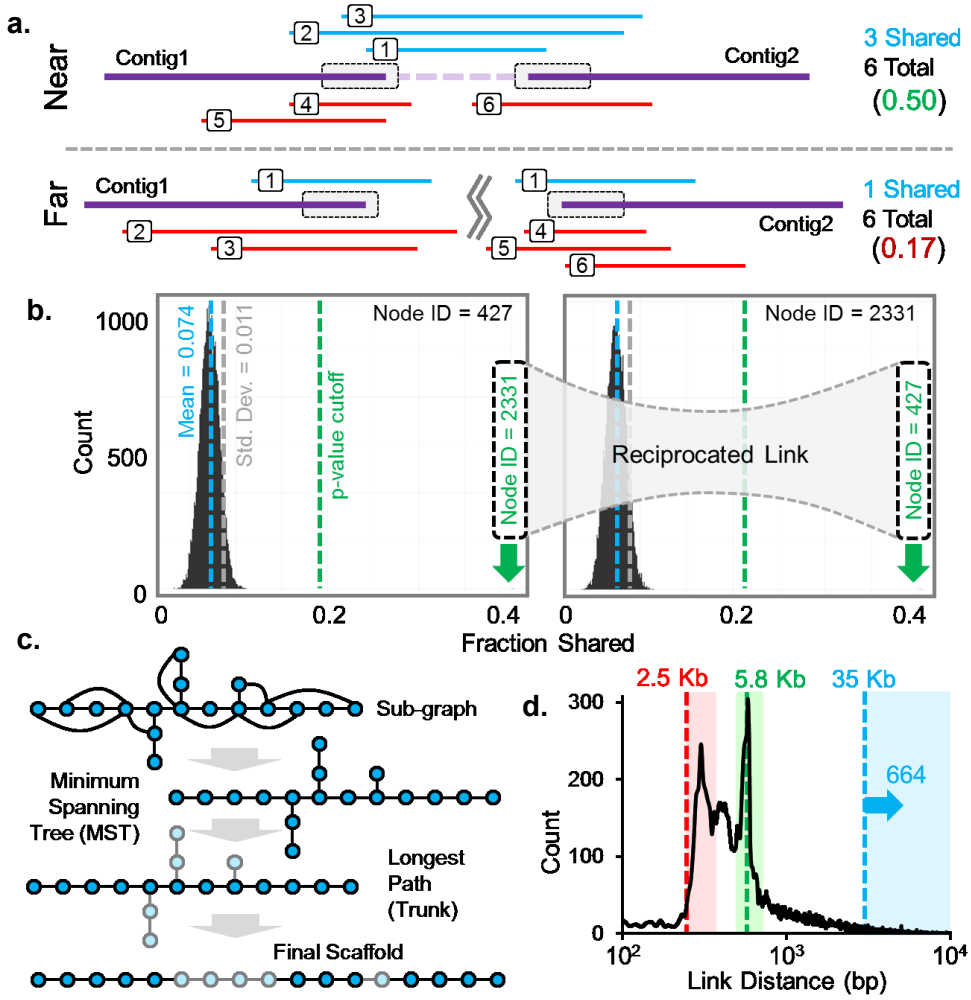


Figure 5.2. *fragScaff* assembly method.
a. The ends of contigs (gray boxes) are defined as nodes and the fraction of shared pools that hit two separate nodes is calculated. Nodes that belong near one another will have an increased fraction of shared pools from long fragments bridging the assembly gap, whereas distal nodes will only have a minimal fraction of shared pools due to separate fragments within pools. **b.** Shared pool fractions are calculated in an all-by-all fashion and shared fraction means and standard deviations are calculated for each node. Outlier nodes are identified using a *p*-value cutoff. If two nodes are each outliers in each other's distribution then the link is reciprocated and stored as an edge. **c.** Sub-graphs are reduced to their Minimum Spanning Tree (MST) and the longest path (Trunk) is found. Branches (light nodes) are then placed to produce the final output scaffold. **d.** Size distribution of properly linked contigs. Boxes indicate joins spanning gaps just beyond the 2.5 kbp mate-pair library (red), ~6 kbp L1 repeat elements (green), and joins longer than 35 kbp which cannot be achieved via fosmid mate-pair libraries (blue).

5.4.4 Bridging the mid-range gap for chromosome-scale *in vitro* assemblies

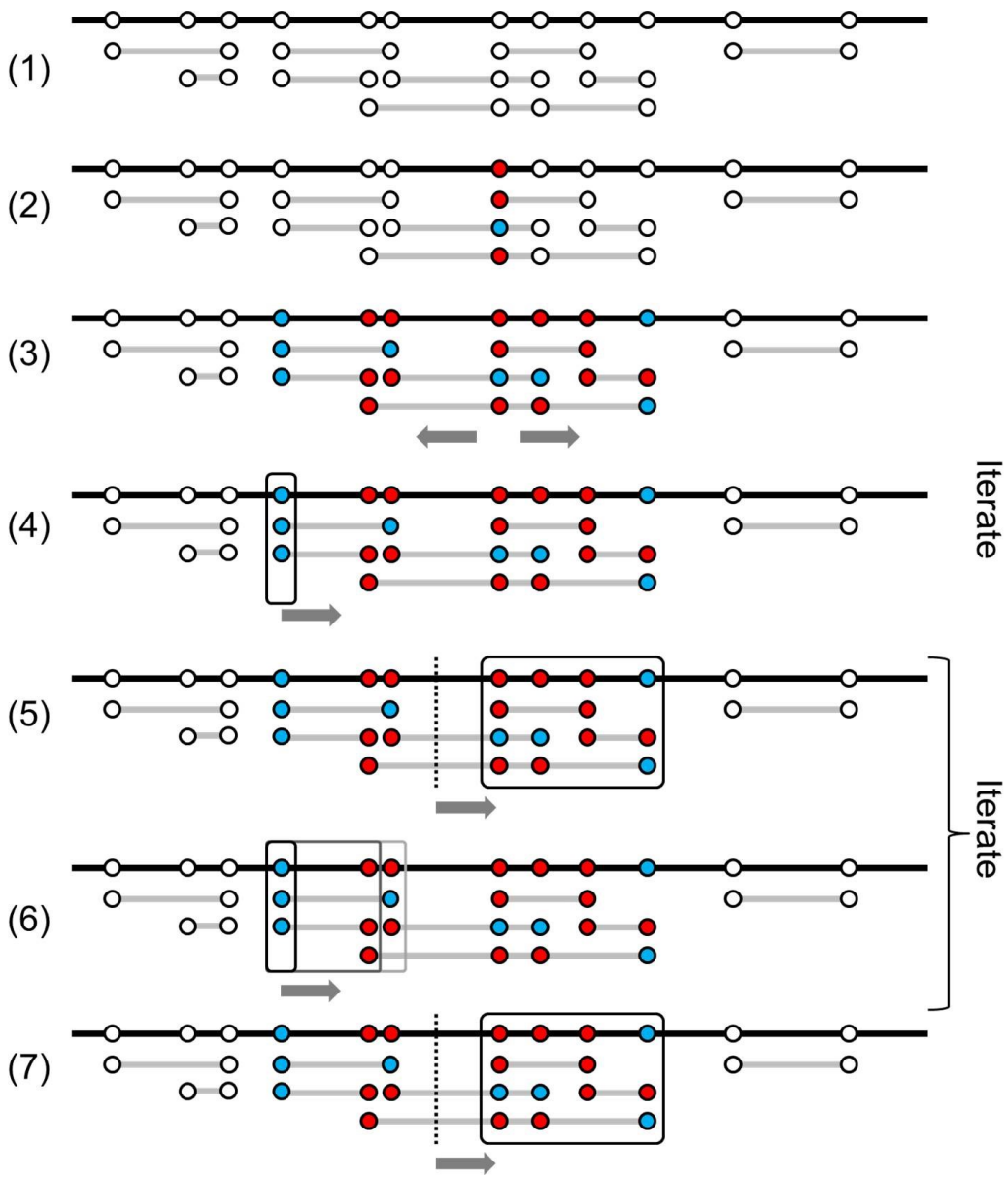
We recently reported the use of Hi-C data for CPM and chromosome-scale *de novo* assembly on human, mouse, and fruit fly genomes[94]. For input assemblies with a large N50, CPM produced high quality

scaffolds such as on human (iSC's with N50 = 437 kbp) where the resulting output clustered 98.2% of sequence with 94.4% ordered. However, on the smaller fly input assembly (N50 = 68 kbp) CPM was able to cluster only 81.2% and order 82.0%. Similarly, when we applied CPM to our human contig assembly (N50 = 47 kbp), 89.7% of sequence was clustered with only 41.5% ordered, and even less for the simulated 15 kbp contig assembly (35.9% clustered, 0.2% ordered). When we perform TC-Seq and *fragScaff* prior to CPM, the completeness and quality of the resulting chromosome-scale *de novo* assembly markedly improves for the human contig assembly (Table 5.1b, Supplementary Table C.3.2) with 99.4% of sequence clustered (gain of 9.7%) and more than doubling the amount of sequence ordered to 99.1% (gain of 57.6%). Similar improvements were also observed for simulated contig assemblies of fly and human, most strikingly for the 15 kbp human simulated contig input wherein the proportion of the iSc assembly that is ultimately clustered and ordered increases from 36% to 99%, and from 0.2% to 88%, respectively. These results demonstrate that TC-Seq and *fragScaff* improve the N50 sizes of assemblies generated from shotgun and short-range mate-pair libraries to a point suitable for chromosome-scale scaffolding with (Supplementary Fig. C.4.4).

5.4.5 Anchoring novel contigs and misassembly detection using TC-Seq

In Kitzman *et. al.* (2011)[61] we described a method that utilized the sub-haploid content of each fosmid pool to anchor *de novo* assembled contigs from reads that did not align to the reference genome, as well as a set of previously anchored sequences (GRCh36) described in Kidd *et.al.* (2010)[99] (Supplementary Fig. C.4.5). This method worked on the premise that each window in the genome is hit by a discrete set of pools, as is each novel contig, and the window with maximum pool overlap with a novel contig is the most probable anchor location. We applied this approach to anchor the same set of contigs to GRCh36 but used TC-Seq data generated on GM12878, thus increasing the number of pools, and therefore anchoring power, from 115 to 9,216. We were able to place 1,816 of the 2,363 contigs (76.9%), though a number of unplaced contigs are likely population-specific and not present in our sample. 1,226 of our placed contigs have published placements with 1,154 (94.1%) in agreement with our calls. The windows used for contig anchoring also allowed for misassembly detection by investigating the pool overlap fractions of immediately adjacent windows and windows one apart. For example, using this method, 1,411 suspicious

loci were identified in the assembly based exclusively on shotgun and TC-Seq data (N50 = 1.6 Mb), of which 629 (45%) were bona fide misassemblies upon comparison with the GRCh37 reference assembly (Supplementary Fig. C.4.6).



5.4.6 *fragPhase*: A novel haplotype-resolution algorithm designed for the high-fragment-count data produced by TC-Seq

As a final application of TC-Seq data, we developed an alternative haplotype phasing algorithm, *fragPhase* (Fig. 5.3, Supplementary Note C.2), which is more suited to the high number of fragments

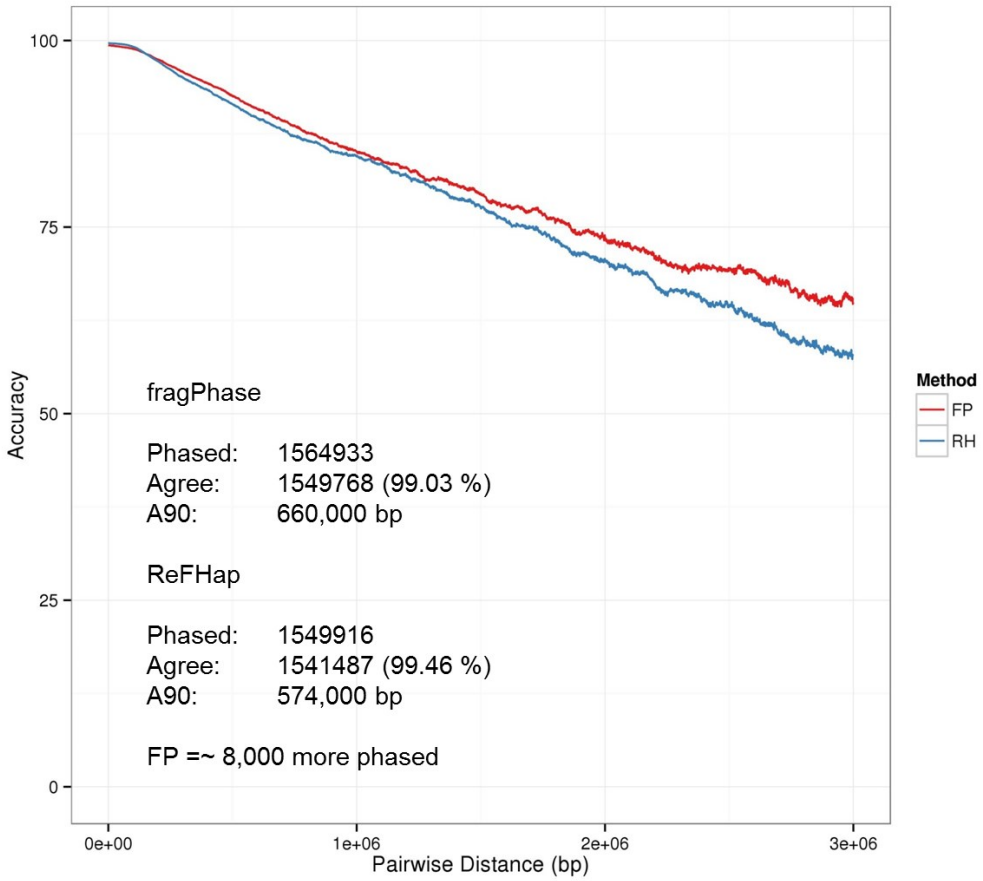


Figure 5.3. *fragPhase* performance.
Accuracy of pairwise variants by distance for *fragPhase* (FP) and ReFHap (RH).

produced (~4 million vs. ~0.3 million for fosmids) that results in excessive runtimes for the ReFHap algorithm[100]. Our approach assigns a graph vertex to every heterozygous SNV and identifies all connected components. A random phase is then assigned to the vertex with the highest degree followed by greedy assignment of all connected vertices to a haplotype as an initiation step. Next, a series of two greedy improvement steps are iterated until no further improvements can be made. The first improvement is checking the resulting score for switching each single vertex and making the switch if it improves the score. Sets of two and then three subsequent vertices are then checked and switched if the score is improved by doing so. The maximum number of vertices switched in each step can be adjusted; however limited improvements were observed when increasing the maximum from three to ten. The second step checks if switching all subsequent vertices provides a score improvement, again making the switch if it improves the score. After convergence, a final pass is taken that is the same as the second greedy improvement step and if the score of a switch is identical to the current score, the block is split at the

switch point. Lastly, the number of fragments are tallied that went into the phasing of each variant along with the competing information to produce quality scores. These quality scores can also be extended into an all-by-all pairwise quality score based on the amount of information in the form of direct fragments or paths of multiple fragments, which is of particular value for instances where the phase of two specific variants is crucial, eg. compound heterozygosity.In order to test *fragPhase*, I performed phasing using ReFHap[100] as a comparison on TC-Seq data generated on NA12878 that has previously been phased by pedigree methods. Both methods produce high quality subsequent variant phasing results with accuracies over 99%. Over greater distances I developed a metric – deemed A90, which corresponds to the distance at which all-by-all pairwise accuracies drop below 90%. This information is captured in Fig. 5.4 as a means of comparing both methods. For shorter ranges (up to ~100 kbp), ReFHap has higher accuracies, however beyond this point *fragPhase* produces higher quality results. A further advantage of *fragPhase* is that it takes approximately 20 minutes to perform phasing where ReFHap required over 72 hours.

5.5 Discussion

We demonstrate that *fragScaff* provides mid-range contiguity information comparable to 40 kbp fosmid-based mate-pair scaffolding, yet is entirely *in vitro* and can be tuned to allow for longer assemblies at a minor cost to accuracy depending on downstream requirements. The ease of TC-Seq library preparation and the amount of information obtained per base of sequencing performed makes it a very appealing tool for routine use in *de novo* genome assembly. The primary limitation of TC-Seq and *fragScaff* at present is that input contigs shorter than 5 kbp (typical end-node size, Figure 5.2a, dashed boxes) are difficult to place confidently due to a reduced number of pools covering the end-node. Additionally, input contigs of length <10 kbp lack sufficient separation between their end-nodes and often do not have enough differentiating information to properly assign orientation. Despite these limitations, the methods presented here bridge the gap between short-range assemblies and the scaffold sizes needed to perform chromosome-scale scaffolding with CPM. We envision that the combination of four types of sequencing libraries – shotgun fragment, 3 kbp mate-pair, TC-Seq, and CPM – will provide a robust strategy to *de*

novo assemble genomes to chromosome-scale contiguity using entirely *in vitro* methods that can be carried out cheaply and rapidly.

5.5 Notes

5.5.1 Data access

Human (Gm12878), Mouse and Fly TC-Seq data are in the process of submission to the Sequence Read Archive (SRA). Referenced SRA accessions include SRA024407 (Human, GM12878), SRA009956 (Mouse), and SRR516038 & SRR516001 (Fly) for input *de novo* assemblies.

5.5.2 Acknowledgements

We thank R. Patwardhan for generating *ALLPATHS-LG* assemblies and other members of the Shendure Lab for helpful discussion. We would also like to thank M. Wilken and the Reh lab for supplying a fresh mouse liver, and N. Peters and the Berg lab for supplying fly cultures. 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.).

Chapter 6 Application of long-range sequence information to haplotype-resolved the HeLa cancer cell line genome and epigenome

This chapter is based on the following published manuscript:

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.

Bold face indicates equal contribution. *Indicates corresponding authorship.

Josh Burton, Jacob Kitzman, Jay Shendure, and I devised experiments, carried out analyses and wrote the manuscript. Joe Hiatt, Alexandra Lewis, Beth Martin, Ruolan Qiu, Charlie Lee, and I maintained cell cultures, constructed libraries and performed DNA sequencing. Jay Shendure supervised all aspects of the study.

6.1 Summary

This chapter focuses on the application of haplotype-resolution technology to the aneuploidy HeLa cancer cell line. We not only perform haplotype-resolution, but use the long-range sequence information to piece together the complex rearrangement in HeLa around a locus in which the HPV-18 viral genome integrated. We further show that this integration resulted in the activation of the *MYC* proto-oncogene. We also investigate the stability of the HeLa genome by determining a mutation frequency as well as by sequencing multiple strains of HeLa. Lastly, we investigate the epigenome of HeLa in the context of haplotype and copy number.

6.2 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*[101]. The robust growth and unrestricted distribution of HeLa cells resulted in its broad adoption—both intentionally and through widespread cross-contamination[102] —and for the past 60 years it has served a role analogous to that of a model organism[103]. 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[104], partly because like many cancers, its extensive aneuploidy renders such analyses challenging. We carried out haplotype-resolved whole-genome sequencing[61] 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[105] and ENCODE Project[44] 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.

6.3 Results

6.3.1 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[61] (Supplementary Table D.3.1). To catalogue variants, we carried out conventional shotgun sequencing to 88-fold non-duplicate coverage and reanalyzed 11 control germline genomes in parallel[106] (Supplementary Tables D.3.2 and D.3.3). Although normal tissue corresponding to HeLa is unavailable, the total number of single-nucleotide variants (SNVs) identified in HeLa CCL-2 ($n = 4.1 \times 106$) and the proportion overlapping with the 1000 Genomes Project[107] (90.2%) were similar to controls (mean $n = 4.2 \times 106$ 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 (Supplementary Fig. D.4.1). Short tandem repeat profiles of HeLa also resembled controls, consistent with mismatch repair proficiency (Supplementary Fig. D.4.2).

After removing protein-altering variants that overlapped with the 1000 Genomes Project or the Exome Sequencing Project, 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 (Supplementary Fig. D.4.3). Although a previous study of the HeLa transcriptome[108] 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) was similar for HeLa (n = 1) and control genomes (mean n = 2.6). The gene-level overlap with the Sanger Cancer Gene Census (SCGC) 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* (B-cell CLL/lymphoma 11B (zinc finger protein)), *EP300* (E1A binding protein p300), *FGFR3* (fibroblast growth factor receptor 3), *NOTCH1* and *PRDM16* (PR domain containing 16), Supplementary Tables D.3.3–6). 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[109]. Mutations in *FGFR3* have been noted previously in cervical carcinomas, although infrequently and at different residues than observed here[110]. Both *EP300* and *NOTCH1* are recurrently mutated in diverse cancers and are involved in Notch signaling, a pathway that is dysregulated in HeLa[111]. *EP300*, which encodes the transcriptional co-activator p300, interacts directly with viral oncoproteins such as HPV-16 *E6* and HPV-16 *E7* (ref. [112]). 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.

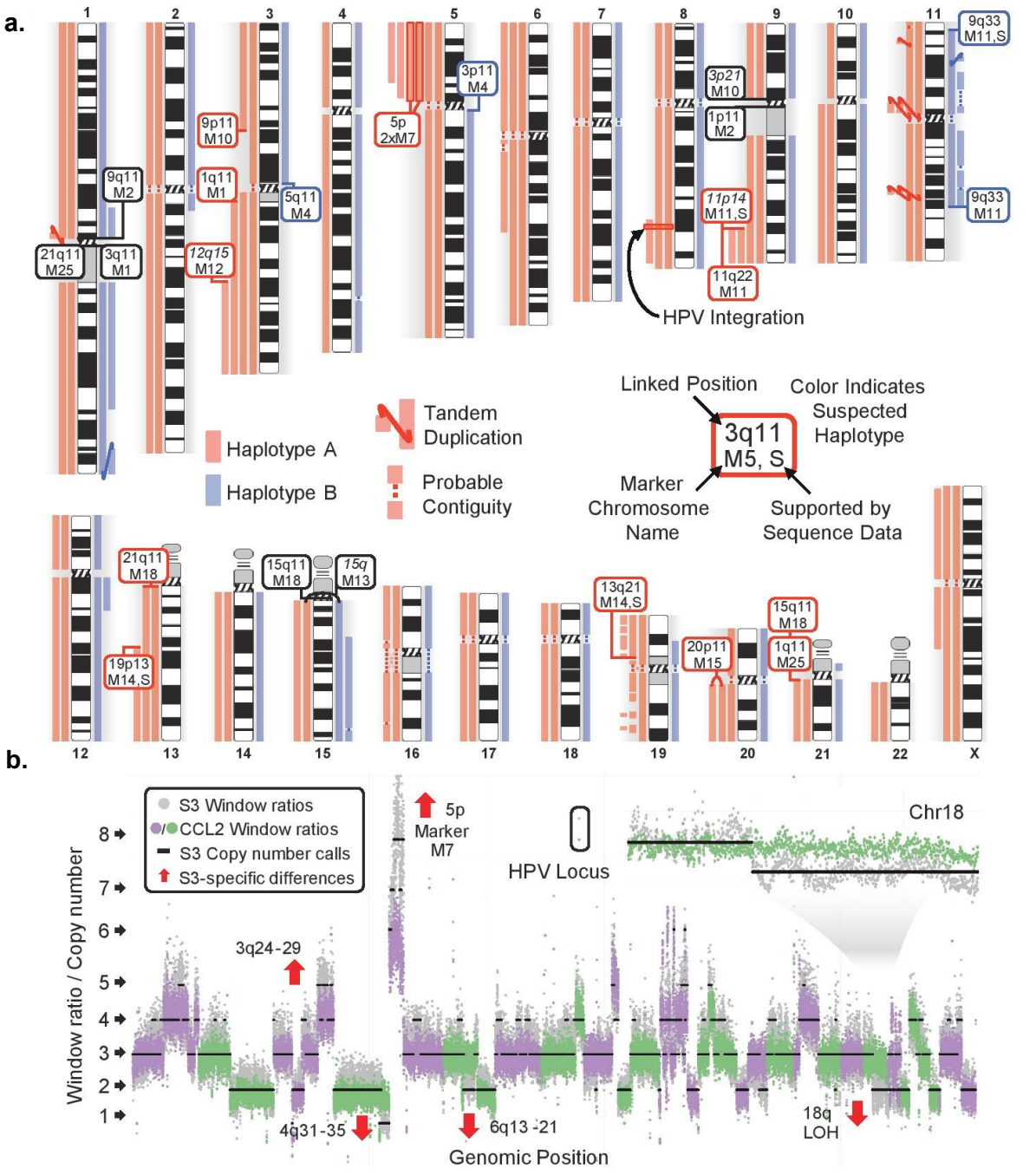


Figure 6.1. Haplotype-resolved copy number of the HeLa cancer cell line genome.
a. Copy-number profile of HeLa split by haplotypes. Links denote likely contiguity and tandem duplications. Boxes indicate marker chromosomes identified by copy-number breakpoints (boxes are colored by haplotype; black, unknown; 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.

6.3.2 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 (Fig. 6.1, Supplementary Fig. D.4.4 and Supplementary Table D.3.7). Read coverage profiles were segmented by a Hidden Markov Model (HMM) and recalibrated to account for widespread aneuploidy (Supplementary Figs D.4.5 and D.4.6). Sixty-one percent of the genome has a baseline copy number of three, and only a small minority (3%) has a copy number of greater than four or less than two (Supplementary Table D.3.8). 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) (Supplementary Fig. D.4.7 and Supplementary Table D.3.9), consistent with previous descriptions of LOH in cervical carcinomas[113]. The overall profile is consistent with published karyotypes of various HeLa strains[104], 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 kbp and 3 kbp mate-pair libraries (Supplementary Fig. D.4.8). 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 (Supplementary Figs D.4.9 and D.4.11, and Supplementary Table D.3.10). Only two genes that are impacted by HeLa-specific structural rearrangements (Supplementary Table D.3.11) intersected with SCGC (*STK11* (ref.[114]), FHIT), both of which are recurrently deleted in cervical carcinomas[114,115].

6.3.3 Haplotype-resolution of HeLa

Conventional whole-genome sequencing fails to resolve haplotype phase, an essential aspect of the description and interpretation of non-haploid genomes, including cancer genomes[52]. Recently, several groups have demonstrated genome-wide measurement of local[61] or sparse[116] 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[61]. 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 kbp, for a total physical coverage of 6.3× of the haploid reference genome (Supplementary Fig. D.4.12). The complement of likely inherited heterozygous variants (SNP and indel, n = 1.97 × 106) 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 (in the context of haplotype: the contig size above which 50% of the total length of the haplotype assembly is included) of 550 kbp 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 (Supplementary Fig. D.4.13). 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 one-third and two-thirds (Supplementary Fig. D.4.14). 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 megabases (Mb); Supplementary Table D.3.12). 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 (Fig. 6.1a, Supplementary Fig. D.4.15 and Supplementary Table D.3.13).

Phasing accuracy was independently confirmed by several methods. First, 99.7% of informative read pairs from 3 kbp 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 kbp; 90th percentile, 5.1 kbp 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% (Supplementary Fig. D.4.16). 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 (Supplementary Fig. D.4.17). 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 (Supplementary Figs D.4.18 and D.4.19).

6.3.4 Mutation frequency in HeLa

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; Supplementary Table D.3.14) were confirmed by clone-pool sequencing and allele frequency in shotgun sequencing (Supplementary Figs D.4.20 and D.4.21). 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 (Supplementary Fig. D.4.22). 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[117]. 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.

Four years after the initial establishment of the HeLa cell line, several additional strains were cloned[118]. 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; Supplementary Fig. D.4.23 and Supplementary Table D.3.15). 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.

6.3.5 Copy number profiles of HeLa strains

The copy-number profile of HeLa S3 broadly mirrors that of CCL-2 (Fig. 6.1b, and Supplementary Figs D.4.7 and D.4.24) as well as eight additional HeLa strains that we sequenced lightly (3.5 to 4.3×). We observed some strain-specific differences (Supplementary Figs D.4.25–27), 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 (Fig. 6.1b), chr9p, CCL-13 (Supplementary Figs D.4.28 and D.4.29)), 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 1970's[119] 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.

6.3.6 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 (refs [120,121]).

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 (Fig. 6.2a, b, and Supplementary

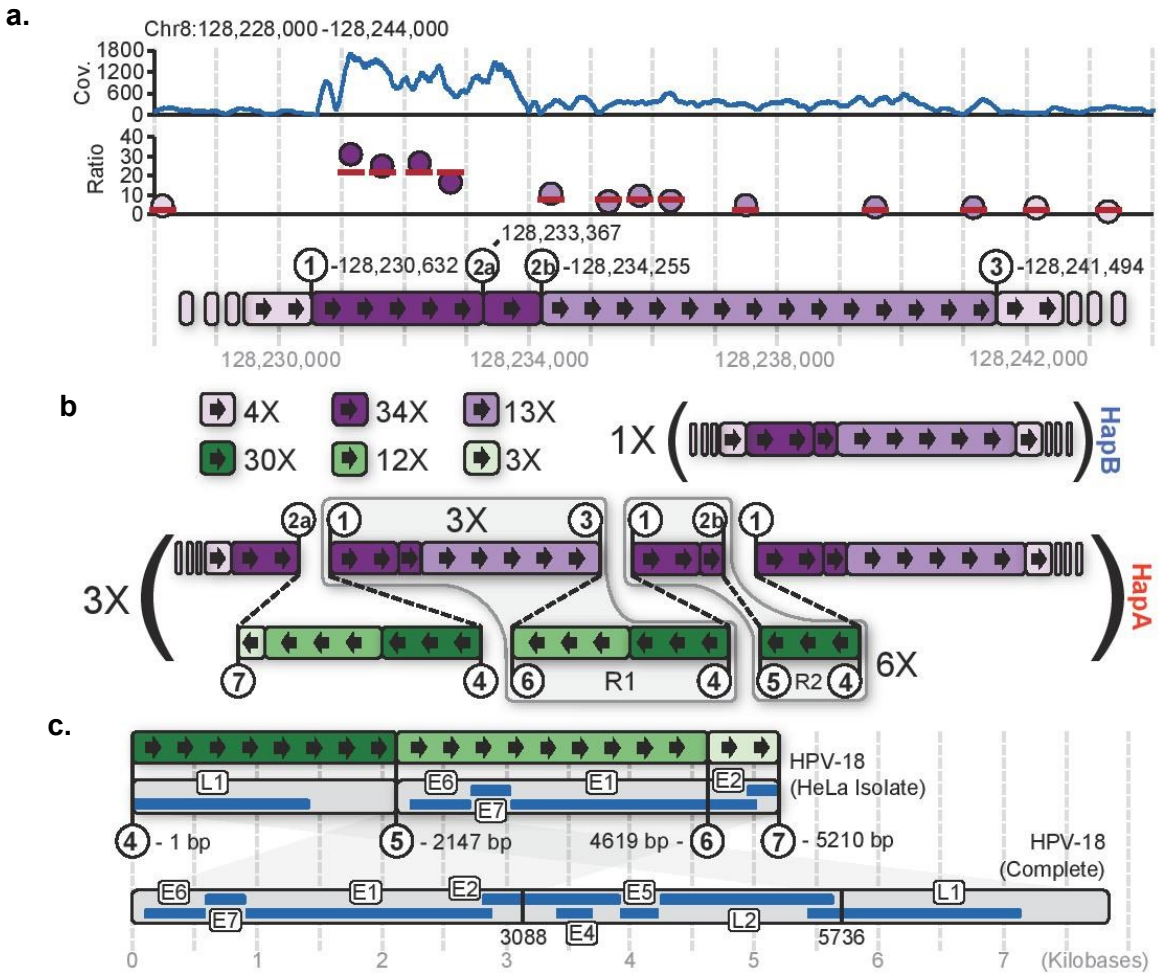


Figure 6.2. HeLa HPV integration locus.
a. Chromosome 8 read depth flanking the HPV integration site (top, blue line), windowed copy-number ratios (purple points, shaded by segment) and integer copy states (red bars, middle), and corresponding segments and breakpoints (circled numbers with genomic coordinates, bottom). **b.** roposed 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).

Figs D.4.30 and D.4.31). 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* (ref. [109]) (Fig. 6.2c). 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[122], possibly contributing to their high expression (Supplementary Fig. D.4.32).

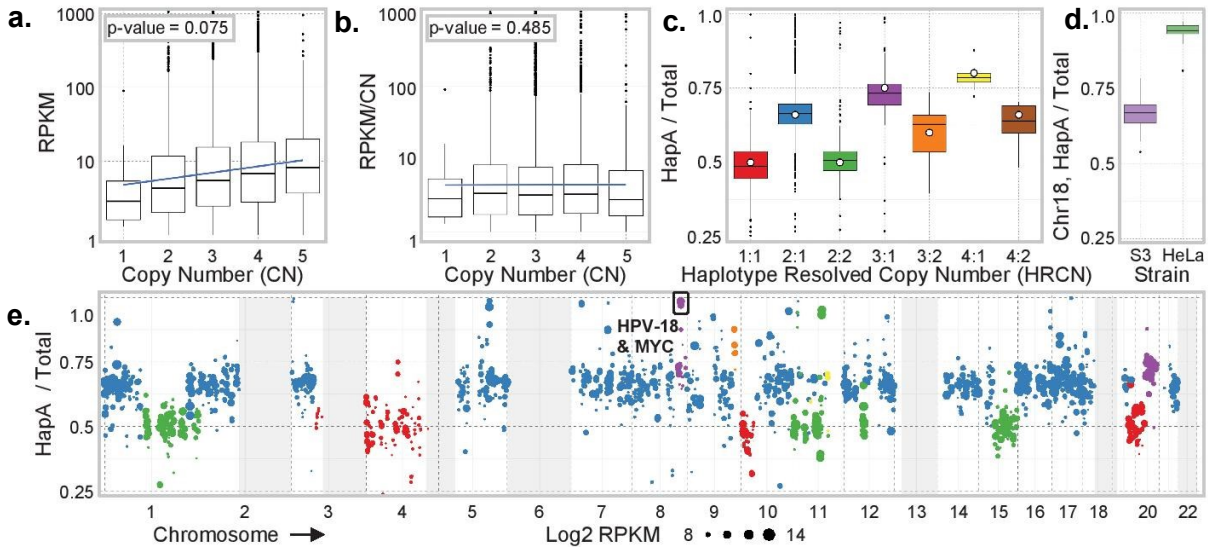


Figure 6.3. Gene expression by copy number and haplotype in HeLa S3.
a. Transcript abundance (reads per kilobase per million (RPKM), for genes with an RPKM ≥1) is positively correlated with gene copy. **b.** Expression per copy (RPKM per gene copy number) does not correlate with copy number. **c.** Fractional contribution of haplotype A to overall expression (Hap A/total) (RPKM averaged across megabase 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 represents the log2 total RPKM, grey boxes indicate HeLa S3 LOH).

6.3.7 Haplotype and copy number resolved epigenome of HeLa

Extensive sequencing-based functional genomic data have been generated on HeLa and other cancer cell lines by the ENCODE Project7, 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 functional data generated on S3 (ref. [44]), 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 (Supplementary Figs D.4.33–35). For each data set, aligned reads were phased by comparison to HeLa CCL-2[105,44] 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; Supplementary Figs D.4.36–38).

Across the HeLa genome, gene expression is significantly correlated with copy number ($P = 0.075$; Fig. 6.3a,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 (Fig. 6.3c,e). This trend is also observed for histone modifications, DNase hypersensitivity and transcription factor binding (Supplementary Figs D.4.39 and D.4.40). 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 (Fig. 6.3d). However, a small number 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; Supplementary Figs D.4.41–44). Interestingly, the HPV-18 insertion locus and proto-oncogene *MYC* (separated by approximately 500 kbp) were among the regions with the most highly haplotype-imbalanced regulation in the genome (Supplementary Fig. D.4.45).

6.3.8 *Cis*-activation of *MYC*

Phased RNA-seq data indicate that *MYC* is highly expressed, but almost exclusively from the HPV-18-integrated haplotype (mean ratio, 95:1; Fig. 6.4b and Supplementary Fig. D.4.46). Phased ENCODE tracks and long-range chromatin interaction data (ChIA-PET (chromatin interaction analysis with paired-end tag sequencing)[42]; Fig. 6.4a and Supplementary Fig. D.4.47) 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[123], acting in *cis* rather than in *trans* and possibly mediated by the epithelium-specific viral enhancer amplified to a copy number of

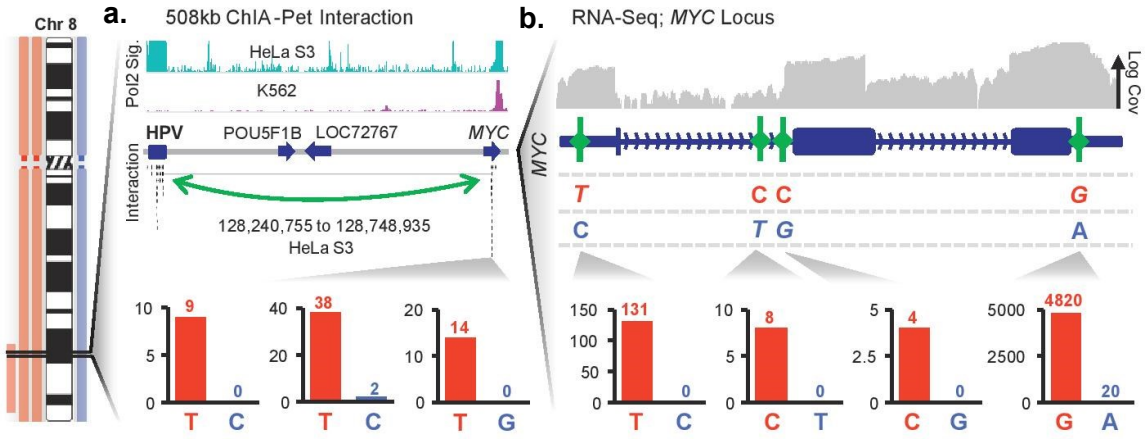


Figure 6.4. Haplotype-specific regulation near the HPV integration site.
a. Long-range chromatin interactions between the HPV and *MYC* loci demonstrated by ChIA-PET27 with the RNA polymerase II signal (top) shown for HeLa S3 and an HPV− cell line (K562). Chromatin interactions (middle) are indicated 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**.

approximately 30 within the R1 repeat structure (Fig. 6.2b)[122]. 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*[124].

6.4 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.

6.5 Notes

6.5.1 Data access

Primary accessions: GenBank/EMBL/DDBJ; DAAG00000000, DAAH00000000, DAAG01000000, DAAH01000000s

6.5.2 Acknowledgements

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 tumour 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 M. Kircher, M. Snyder, A. Kumar and R. 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.

Chapter 7 Addressing the limitations of high throughput sequencing: Successes, challenges and future direction

In this final chapter I discuss the successful outcomes of the work presented in this dissertation and how these advances have – and will – impact the field of genomics and epigenomics. I also describe the challenges that were encountered during this work, how some were overcome, and how others still remain. Lastly, I describe my plans to continue on in the direction this work has taken me by developing and implementing methods for precision epigenomics.

7.1 Introduction

The fields of genomics and epigenomics have been largely fueled by the development of new technologies that allow us to answer questions that would have otherwise been impossible. Perhaps the most driving of these technologies is our ability to sequence nearly the entire complement of DNA present within an organism. Since the development of DNA sequencing technologies, there has been a continual decrease in the cost to resequence a genome that has plummeted faster than Moore's Law, which would correspond to halving the cost of sequencing every two years (data from NHGRI). Much of this progress has been a result of next generation sequencing (NGS) platforms, which can now resequence a human genome for just over $1,000. However, the increase in throughput and decreased costs have come at the price of short read lengths that prevent accurate resolution of structural variants or haplotype phasing and the use of protocols that require substantial amounts of input, which often mean compromising sample purity to enable entry to the NGS workflow.

7.2 Successes

7.2.1 Low-input and high throughput DNA-Seq library construction via transposase-mediated fragmentation and adaptor incorporation

The first challenge facing the field that this work addressed was the development of a method to substantially reduce both the input requirements with respect to DNA mass, as well as the throughput

bottleneck for next generation sequencing library construction. I characterized in extensive detail the use of a hyperactive derivative of the Tn5 transposase to accomplish the majority of steps in library construction. This method, later commercialized as “Nextera”, and is now fully integrated into the Illumina next generation sequencing platform, performs comparably to the standard method of library construction and allowed for a reduction in pre-PCR library construction time from 8-10 hours down to a single, 5 minute enzymatic reaction followed by an optional clean up.

I demonstrated the comparable performance of transposase-mediated library construction by sequencing bacterial, viral, fly, and human genomes to high coverage and analyzing potential biases, including G+C bias in coverage depth, uniformity of coverage depth, and library complexity (the fraction of unique molecules in the final library). On all fronts this method produced nearly identical results, with the majority of bias being shared between the platforms due to the final PCR step. This therefore served as the motivating factor for me to develop a PCR-free version of transposase-mediated library construction, as such approaches had previously shown reduced biases for standard methods of library construction[71,72]. This extension resulted in not only a substantial decrease in library construction time (<30 minutes from input DNA to sequencer-ready libraries), but substantially mitigated the biases introduced during PCR amplification, particularly with respect to G+C content as highlighted in Figure 3.7. To further demonstrate the utility of transposase-mediated library construction, I generated libraries that were then used for exome capture and sequencing, produced useable sequence information from as little as 10 picograms of DNA (or approximately three haploid genome equivalents), demonstrated the ability for 96-plex sample indexing, and a colleague showed that libraries could be produced directly from bacterial colonies.

Since the publication of Adey *et. al.* (2010)[59], transposase-mediated library construction has seen widespread adoption, particularly where sample quantities are limited, or a rapid protocol is favorable. Additionally, this method has been utilized as the final stage in library construction for single-cell DNA sequencing[125], as well as for single-cell RNA sequencing[126]. Furthermore, I have taken advantage of the reduced input requirements in my own work in the form of a collaboration that involved whole genome sequencing of flow cytometry sorted, highly-pure populations of Hodgkin’s Reed-Sternberg cells of limited

quantity which has shown promising results with respect to the identification of somatic point and copy number alterations (Salipante & Adey *et. al.* 2014, in preparation).

7.2.2 Extension of transposase-mediated library construction to whole-methylome interrogation from limited starting material

Shortly after publishing the initial work on transposase-mediated library construction for DNA sequencing, Rick Myers' group released a method that used this technology to produce strand-specific RNA-Seq libraries[86]. At the same time I was working on a different adaptation of transposase-mediated library construction to allow for a significant reduction of input requirements for sequencing the entire methylome of an organism via whole genome bisulfite sequencing (WGBS)[60]. This method required several key modifications to the protocol to allow for compatibility with whole genome bisulfite sequencing. For the standard ligation-based library construction, methylation of the adaptors was all that was required[29]; however in transposase-mediated library construction a gap is introduced on the bottom strand of the molecule. This gap is repaired during the first cycle of PCR in DNA-Seq protocols but for WGBS bisulfite conversion must occur prior to PCR which involves denaturation of the DNA; therefore the gap results in library molecules with an adaptor at only one end that cannot be PCR amplified. To get around this, I developed an approach that fills in the gap and appends the second adaptor prior to bisulfite conversion and subsequent PCR.

To demonstrate that this method of library construction performed comparably to ligation-based methods, I sequenced the methylome of a human lymphoblastoid cell line. I additionally generated ligation-based WGBS libraries on the same cell line for a direct comparison. Both methods produced very similar profiles that were consistent with previously published studies on similar cell types yet the transposase-mediated method allowed for a reduction of input DNA requirements by 100-fold down to a mere 10 nanograms and thus opened up a number of opportunities for whole methylome profiling where limited cell counts are a constraint, such as in development or cancer.

Later in the same year, Kun Zhang's group utilized the input advantages of my method in their Nature publication regarding Tet1's control of meiosis by regulating meiotic gene expression[127]. Additionally, the

high throughput sequencing core at the German Cancer Research Center (DKFZ) in Heidelberg, Germany reproduced the method in a production setting and we jointly published a Nature Protocols paper to allow other groups to readily implement the method[128]. Continuing on the low-input epigenetic interrogation trend, a Stanford group published a method of profiling open chromatin using transposase-mediated library construction this last year[96], further emphasizing the versatility and ease of the platform.

7.2.3 The acquisition of long-range information via transposase contiguity to substantially improve *de novo* genome assemblies entirely *in vitro*

Early in the development of transposase-mediated library construction, the idea arose of leveraging the inherent properties of the transposase to provide long-range sequence information. This idea spawned over a dozen potential approaches to achieve the goal[129]. One of these methods utilized a continuous, barcoded transposon, such that each insertion event would retain a unique barcode and subsequent restriction digest would allow each resulting fragment to terminate in the barcode of the insertion event. A similar approach involved constructing linked, uniquely barcoded transposons to produce the same type of "fragmentation site barcoding" data. These methods would then allow for the *in silico* stitching-together of reads that were originally adjacent in the target DNA and effectively produce strobe reads over a long range. Unfortunately, none of these methods were able to accomplish their aim with any meaningful efficiency.

An alternative approach, to which I contributed, that was led by a colleague, Jerrod Schwartz, involved the physical stretching of high molecular weight DNA directly on an Illumina sequencing flowcell[97]. The ends of the DNA were modified with adaptors such that they could hybridize to the surface primers on the flowcell. Subsequent transposition throughout the long molecule would allow the terminal, flowcell-bound ends to form complete library molecules that could undergo bridge-PCR and sequencing which produced cluster-pairs at the ends of the long originating molecule that can be linked by physical proximity. This method successfully produced mate-pair sequencing libraries *in situ* for ranges up to eight kilobases; however it failed to produce information over longer ranges.

The latest approach, transposase contiguity sequencing, or TC-Seq, detailed in Chapter 5, relies on an inherent property of the Tn5 transposase, in that the Tn5 enzyme remains tightly bound to the transposed target DNA after catalyzing the "tagmentation" reaction. This results in long chains of library molecules that are physically bound to one another in the order from which they were originally derived. These long chains can then be subject to limiting dilution such that any of the pools represent only two to three percent of the target genome. In order to achieve a sufficient number of these pools, a two-tiered indexing approach was utilized that combines the incorporation of one of 96 barcodes at both the transposition and PCR steps to produce 96 × 96 = 9,216 distinct compartments. Each compartment contains a couple thousand sets of reads derived from single fragments that range in size from a kilobase to over a megabase with each fragment having been derived from a single haplotype.

The first application of TC-Seq was in the context of haplotype-resolution. This work, which is presented in collaboration with Illumina in Amini *et. al.* (2014, under review)[62] produced a haplotype resolved human genome with error rates and completeness metrics comparable to the fosmid-based methods that we first described in Kitzman *et. al.* (2011)[61]; however all methods used in the study were carried out entirely *in vitro* without the need for time-consuming cloning methods. One complication that arose during the data analysis portion of the haplotype-resolution was that ReFHap[100], the software used to differentiate between the two haplotypes, was designed for fosmid-based data that contains upwards of ~200,000 fragments as opposed to the >20 million fragments produced by TC-Seq. This resulted in excessively long run times for the haplotype-resolution to complete and motivated me to design a novel algorithm that is more suited to the increased pool counts. The algorithm, *fragPhase*, described in 5.4.6, allowed for a reduction in run time from >72 hours to 20 minutes, increased the number of phased variants by approximately 8,000, and produced higher accuracies at longer ranges. However, *fragPhase* did not perform as well at distances below 100 kilobases with accuracies roughly one percent lower than with RefHap.

The second application of TC-Seq was work, which I led, to apply the long-range information to the problem of *de novo* genome assembly, which I describe in greater detail in Chapter 5 of this dissertation. This approach leveraged fragment coincidences between pools as a means of identifying adjacent

contigs in an algorithm called *fragScaff* which allowed for up to 75-fold increases in N50 sizes (measurement of assembly contiguity). I applied TC-Seq and *fragScaff* to human, mouse, and fly genomes that were assembled from easy, *in vitro*, shotgun and 3 kbp mate-pair libraries as well as on simulated assemblies that were generated by *in silico* fragmentation of the human reference genome to set sizes. Each output scaffolded assembly from *fragScaff* observed a substantial increase in contiguity with high accuracy. Furthermore, I demonstrated that the method can be used to fill the mid- to long-range gap between short-range methods and the chromosome-scale contact probability map (CPM) method we describe in Burton *et. al.* (2013)[94]. Assemblies that were first scaffolded using *fragScaff* prior to CPM produced much more complete assemblies of higher accuracy and substantially improved contiguity when compared to CPM alone, particularly when the initial assemblies were of low contiguity.

7.2.4 Application of long-range sequence information to produce the most comprehensive view of a cancer genome to date

We had previously described fosmid-based haplotype resolution methods in Kitzman *et. al.* (2011)[61] and applied this technology to successfully haplotype resolve the diploid genome of a Gujarati Indian individual. This raised the question of whether or not the method could be executed on a non-diploid genome, such as that of an aneuploid cancer cell line. I decided that HeLa would be the ideal case to first attempt haplotype-resolution of a cancer genome given its ubiquitous use, and role akin to that of a model organism in cell biology. Second, HeLa is incredibly complex with widespread aneuploidy and a number of complex structural rearrangements[104]; therefore if HeLa can be successfully haplotype-resolved using these methods, then the approach should be able to be translated to any cancer genome. Lastly, a large amount of epigenetic data sets had previously been generated on HeLa S3 (a sub-strain of HeLa) that could be re-analyzed in the context of haplotype and copy number[44].

Using high-depth shotgun sequence data I was able to perform high-resolution copy number calling by designing and implementing my own Hidden Markov Model based algorithm. These calls were combined with the fosmid-based sequence data to produce a haplotype-resolved copy number profile of the HeLa genome.

Contrary to expectations, the HeLa genome is relatively stable in terms of both point and structural variation by comparison to normal controls, and nine additional sequenced strains. Furthermore, there were no point variants identified that could confer the robust growth properties of HeLa. However, the long-range information provided by haplotype-resolution methods allowed me to assemble a highly amplified and rearranged locus on chromosome 8q24.21 that included human papillomavirus 18 (HPV-18) sequence. The viral genome integration is present on only one of the two haplotypes at the locus, the same haplotype in which the proto-oncogene *MYC* is highly expressed in a haplotype-specific manner approximately 500 thousand bases away. These data along with the application of haplotype information to chromatin conformation sequencing revealed strong *cis* activation of *MYC* by the HPV-18 epithelial-specific enhancer present at high copy number at the integration locus. Copy number and haplotype-resolution also allowed me to globally investigate the effects of gene dosage and allelic imbalance on epigenetic signatures, and gene expression.

The current standards in genome sequencing are unable to discern complex structural variation and do not provide a framework for identifying *cis* regulatory interactions. The application of long-range contiguity methods was crucial in the identification of the driving lesion that ultimately resulted in the immortalization of the HeLa cell line, an event that is likely similar to other, undiscovered lesions in other cancers. These approaches and their demonstrated utility serve as model for the comprehensive characterization of complex genomes and their regulatory landscape.

7.3 Challenges

7.3.1 Cost of acquiring long-range sequence information: Is it worth it?

In spite of the advantages gained from acquiring sequence information of high contiguity, the cost of these methods is above that of standard shotgun sequencing. This is a particularly steep obstacle for studies that require high sample counts. Specifically for phasing applications, such as fosmid-pool approaches or transposase contiguity sequencing (TC-Seq, described in Chapter 4), standard shotgun sequencing must first be performed to identify variants that are phasing candidates. Furthermore, targeted methods are not compatible with the current phasing approaches and thus require whole

genome sequencing. Attempts have been made to rely solely on fosmid-pool or TC-Seq data for both identification as well as phasing, but with limited success. Based on current sequencing cost estimates, employing a haplotype-resolution method increases the cost-per-genome by approximately 50% (fosmid, which also includes a time-consuming, laborious cloning step performed by a skilled scientist) or 100% (TC-Seq, though further developments should push this down to 50%, uses a simple library construction protocol). With these cost estimates, haplotype-resolution should be reserved for studies where it is either required for the experimental aims, such as in the context of fetal genome sequencing[130,131,132], or for population genetics applications[133].

In contrast to the proportionately high cost of implementing long-range sequencing technologies in order to perform haplotype-resolution, acquiring this information for *de novo* genome assembly is quite cheap, particularly for TC-Seq. The bulk of the cost for genome assembly lies in the acquisition of extremely high shotgun sequence coverage[92] and therefore the incremental cost increase for the quality of information obtained is much more favorable. Furthermore, the amount of TC-Seq raw sequence read counts that need to be obtained for assembly is approximately half that when compared to haplotype-resolution. This results in an approximate 10% increase in the amount of sequence to be obtained that can produce an up to 75-fold increases in assembly scaffold lengths which makes a strong case for the use of TC-Seq in the context of *de novo* genome assembly.

7.3.2 Limitations of pool-based scaffolding

In Chapter 4 I demonstrated the ability for TC-Seq and the *fragScaff* algorithm to scaffold together assembled contigs to provide a greatly increased level of contiguity in human mouse and fly genome assemblies. However, there remain some significant limitations to this method, particularly surrounding the pool-based nature of the approach and how repetitive elements in the genome can confound this method in two primary ways: i) repeat elements or segmental duplications that occur at the ends of contigs, or nodes, can create an increased count of pools that contain reads aligning to the node and thus produce incorrect joins, and ii) the method ultimately scaffolds contigs, yet does not fill in the gaps between these contigs.

In order to address the first limitation, *fragScaff* does not include nodes that have excessively high pool hit counts (default is the top 95%). Additionally, I am in the process of developing a method that will first perform repeat masking on the input contigs and produce a bed file of regions to exclude aligned reads which should remove some of the false joins, yet if the repeat element within the node is too large it will instead just prevent any links from occurring and no join will be made, though that is still preferred over an incorrect join.

The second limitation requires an alternative approach to place the appropriate sequence between the joined contigs as opposed to a series of N's. One potential method that uses existing data is to identify the shared pools that were used to make a join and attempt to identify a specific repeat element with an exact sequence matching the reads within those pools; however this method quickly falls apart if the repeat elements in the genome are highly similar (ie. ≥95%) or at very high frequency.

The ideal means to address both limitations is the development of a new method that does not require fragment pools, but instead stretches out the joined library molecules produced by TC-Seq on a sequencing flowcell to produce strobe-like sequence across a long range thus effectively reducing the fragment count per "pool" to only one and removing ambiguity. However, this next iteration of TC-Seq still requires substantial development before a working method is in place.

7.3.3 Technical challenges of low-input interrogation of other epigenetic marks

The methods I present in this dissertation have allowed for a substantial reduction of starting material for DNA-Seq experiments as well as for the interrogation of DNA methylation across the genome. It is now possible to interrogate the DNA landscape with respect to both point variation and copy number variation in single cells[134,125,135], as well as the transcriptional profile of single cells[126,136,137,138]. However, extending the realm of single-cell interrogation to epigenetic information poses a number of challenges.

First, the primary, and most quantitative method of DNA methylation detection involved conversion of unmethylated cytosines to uracils by sodium bisulfite treatment. This process is very harsh on the DNA and results in breaks as well as depurination. When working with extremely limited quantities of DNA, the resulting material that has not been destroyed is now a much smaller subset of the genome. In order to

get around this, I propose single-cell DNA methylation profiling of a large number of cells in order to allow sparse clustering. This approach is detailed later under future directions and will hopefully overcome this technical challenge.

Second, interrogating other aspects of the epigenome often involve enrichment based methods. This results in read depth profiles with increased depth where enrichment occurred and decreased depth where it did not. When the theoretical maximum level of unique coverage for a single, diploid cell is two, such methods of analysis are not possible. In order to work around this, single-molecule methods, or methods that provide equivalent information are likely the best option. However, such approaches have not yet been developed and will likely require a great deal of refining to enable the acquisition of meaningful epigenetic signatures from extremely limited or single-cell sample sizes.

Lastly, methods for single-cell analysis require the leveraging of information over large cell counts, thus increasing the amount of sequencing required for any given experiment. With the costs of DNA sequencing via NGS methods continuing to decrease this may not remain a significant challenge, but as costs currently stand, the price may be inhibitory for many groups.

7.4 Future directions

My work thus far has centered around two major themes: (I) the cost-efficient, rapid interrogation of the genome and epigenome of samples with limited starting material, and (II) the use of long-range sequence information to investigate structural variation, haplotypes, and *cis* interactions. I plan to continue with these themes to investigate the prevalence and role of epigenetic heterogeneity within a sample.

7.4.1 Epitypes: Identification and characterization

Cell fate determination follows an intricate, highly-specific path of epigenetic reconfiguration. This dynamic process involves the deposition and removal of chromatin marks that form the regulatory landscape of a differentiated cell. Recent studies have investigated epigenetic configurations across a breadth of cell types[54,139,53]; however the frequency and distribution of these marks, particularly DNA methylation, reveal heterogeneity within the population and suggest epigenetic sub-types, or “epitypes”

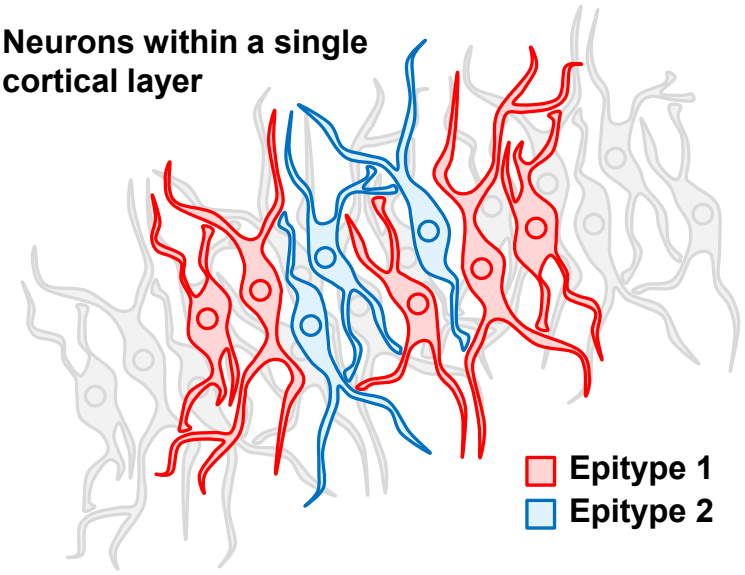


Figure 7.1. Neuron epitypes.
Even with highly-specific cell-type isolation, epigenetic heterogeneity will be present.

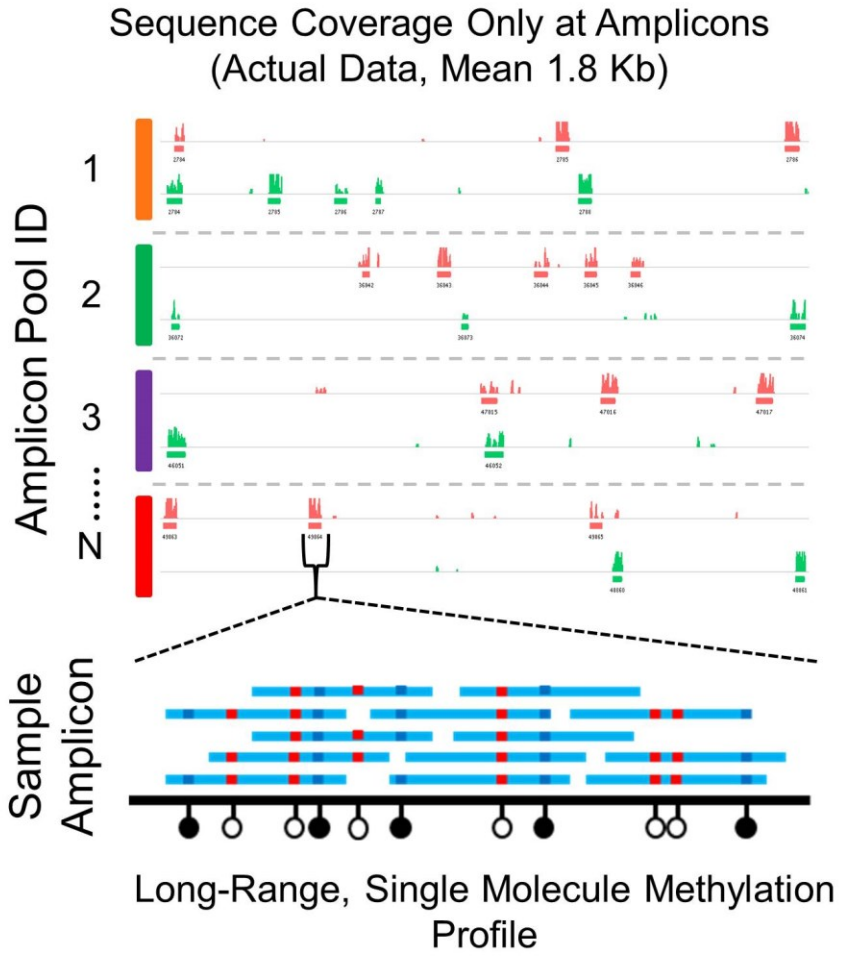
(Fig. 7.1). This observation is also present in the DNA methylation profiles of isolated cortical neurons where non-CG methylation (mCH) is prevalent and accumulates during synaptogenesis[140,141]. To understand the complex role that methylation plays in neuron epigenetic programming, high precision methods must be employed to identify these distinct states within a cell type. Furthermore, any hope of elucidating the impact of an aberrant programming event on neurodevelopmental disorders will require a comprehensive view of the epigenetic landscape beyond that of our present, limited scope.

7.4.2 A new approach to understanding DNA methylation

Current strategies for methylation profiling rely heavily on high-throughput sequencing and thus suffer from the crux of the technology: short reads that are derived from a population of molecules. This severe limitation prohibits the ascertainment of long-range information and thus fails to capture the context of the mark both locally and globally within the genome. Furthermore, the readout of these methods often does not fit the expected tri-modal distribution (100% methylated, 50% methylated (allele specific), or 0% methylated). I hypothesize that this is predominantly due to epitype heterogeneity within the population of cells which can be assessed by methods that provide regional and global context to the observed profiles. I plan to further develop technologies to interrogate methylation architecture on local and chromosome-level scales. This will entail (I) the further development of a working technology I created that provides

long-range, single-molecule methylation information, and (II) performing single-cell methylation profiling and sparse clustering to elucidate epitypes within a population of cells. I will apply these approaches to mature mammalian neurons and then at progressive stages of brain development.

7.4.3 (I) Interrogation of dense, local methylation architecture



I recently performed proof-of-principle experiments on a novel technology (LR-mC, Fig. 7.2) that provides up to 3 kilobases (kbp) of sequence contiguity genome-wide for the identification of 5-methylcytosine. This approach involves the construction of adaptor-ligated 1-3 kbp fragments that are bisulfite converted and then subject to limited-dilution PCR (~20% of the genome in viable amplicons per reaction). The PCR products are then converted to sequencer-ready libraries via barcoded transposase-based library preparation, pooled and sequenced. Each pool is therefore comprised of islands of coverage that consist

of reads which originate from the same initial molecule and thus provide methylation status over an extended range (1-3 kbp) for that single molecule. Thus far, I have successfully demonstrated this technology at limited pool counts and in the future I plan to scale up the method for complete interrogation of mammalian methylomes.

This technology will allow me to gain insight into adjacent methylation site dynamics that is not possible with current technologies. For example, a series of cytosines are all methylated with ~30% of the coverage containing a methylated residue: are 30% of the molecules methylated at all sites? or are 30% of the sites in each molecule methylated? The answers to this question will provide insight into which regions of the genome conform to an epigenetic heterogeneity, or epitype model, and those that are likely stochastically methylated over a functional region. One key advantage to implementing this strategy is that regardless of the result, it will further our knowledge of the architecture and basic mechanisms of DNA methylation. These data can also be integrated with other epigenetic data sets, including those produced for the ENCODE project, to investigate the relationship to the greater regulatory landscape. Furthermore, the increased contiguity will provide an ideal platform to perform haplotype assignment for the robust analysis of allele-specific methylation and imprinting in the genome, a task that has been particularly difficult in humans where heterozygous sites are much less dense than in mouse models which are the current standard[140].

As with many experiments, particularly in the field of epigenetics, careful sample selection is paramount. Genome-wide methylation in mammalian brain tissue is a largely unexplored domain, with the discovery of a significant increase in non-canonical (mCH) methylation in frontal cortex neurons throughout early development and synaptogenesis having been recently described[140,141]. With that in mind, I plan to first isolate neurons from individual cortical layers of a mouse forebrain via micro dissection and $NeuN^+$ flow sorting. Initial whole genome bisulfite sequencing on a size-selected (600 bp) library and paired-end 300 bp read sequencing will allow for 600 bp of contiguity. Light sequencing of library pools on a bench top machine will then aid in the selection of which isolate to explore more deeply using LR-mC, as well as give a sense of what to expect from the longer contiguity approach. Extremely precise cell type selection

is ideal; however demonstration of this technology and the biological insights that can be gained is possible on samples that are much more easily obtained via commercial avenues.

7.4.4 (II) Single-cell methylation analysis for genome-wide contiguity

Long-range methylation methods will allow for a better understanding of local epigenetic dynamics, yet contiguity is limited to a few kbp. To gain insights into genome-wide effects, I will perform multiplex single-cell methylation sequencing of primary tissue (Fig. 7.3). Recently, a group has published on single-cell reduced representation bisulfite sequencing which generated extremely sparse information[142]. The application of similar techniques applied to whole genome bisulfite sequencing (WGBS) will be technically simpler (one entire genome equivalent of viable input mass as opposed to CG-island-only viable input mass) and provide more coverage per cell. However, the majority of aligned reads will be derived from loci outside of CG islands, and therefore most relevant to cell types where non-CG methylation (mCH) is prevalent, such as in mammalian neurons or in the very early stages of mammalian development. I plan to carry out this project on the same cell types as LR-mC, or 600 bp library sequencing to allow integration of the two distinct levels of contiguity.

The majority of covered loci between any pair of single cells will not likely overlap, though aggregation of

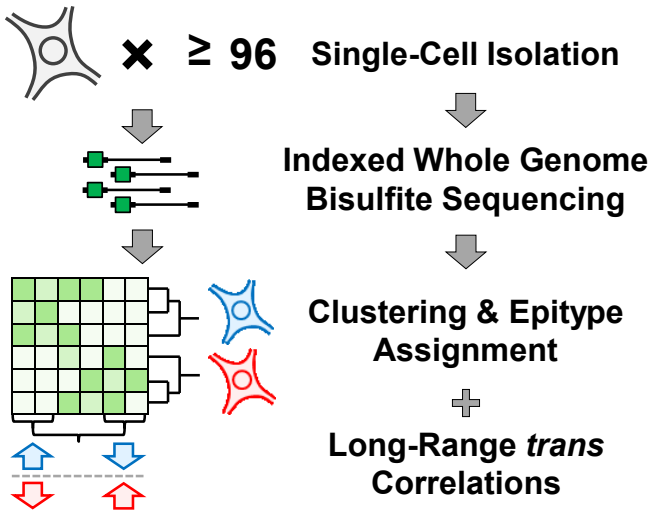


Figure 7.3. Single-cell methylation.

signal across functional elements along with sufficient cell counts will facilitate sparse clustering into respective epitypes. It is also entirely possible that epitypes are not readily distinguishable and more of a gradient in epigenetic space. As with the possible results from my local LR-mC approach, either outcome would provide a better understanding into DNA methylation dynamics. Beyond global classifications, I also plan to investigate *trans* correlations of the methylation status of functional elements. Furthermore, large-scale somatic copy number alterations that have been described in mammalian neurons could be ascertained from these data[125], though likely only amplifications or deletions on the multi-megabase scale, such as entire chromosome arms. If such events have an impact on the epigenetic landscape that includes the perturbation of DNA methylation, detection of these changes may be possible.

7.4.5 Precision epigenomics

Our current strategies lack the ability to differentiate between epigenetically distinct sub-populations within a tissue type. This shortcoming is particularly significant in developmentally complex cell types, such as neurons, or in cancers where heterogeneity can be extreme (eg. glioblastoma). The technologies I will employ provide a clear path forward towards understanding the methylation architecture of mammalian cells with a high level of precision and will facilitate detailed epigenetic cell fate mapping. These insights will not only advance our comprehension of epigenetic reprogramming in development, but allow us to discern subtle shifts in an epigenetic landscape that mark transition into a disease state.

7.5 Conclusions

The work presented in this dissertation addresses two primary shortcomings of next-generation sequencing methods: (I) library construction bottlenecks with respect to both time and input requirements and (II) the read length limitations that inhibit comprehensive interrogation of genomes. I apply these methods to produce the first haplotype-resolved cancer genome in which I also utilize long-range sequence information to discern the structure of a complex genomic rearrangement. While a truly complete view of all genomes and epigenomes of a multicellular organism is still beyond our reach, these approaches and the strategies I present as future directions will take us several steps closer to that goal.

Appendix A: Supplementary material for Chapter 3

A.1 Supplementary methods for Chapter 3

Genomic DNA: Genomic DNA for H. sapiens NA18507 (Yoruban) was prepared by the Coriell Institute. *E. coli* CC118 (MC1000 (*araD139* Δ(*ara leu*)*7697* Δ*lacX74 phoA*Δ*20 galE galK thi rpsE rpoB argEam recA1*))[143], genomic DNA, and CRW10 and PA1 phage genomic DNA were extracted using Qiagen (Valencia, CA, USA) MidiPrep buffers P1, P2, P3 and were cleaned by phenol, purified from low melting point agarose and dissolved in Tris/EDTA (TE) buffer. YH1 genomic DNA was extracted from a lymphoblastoid cell line of Yanhuang[75] using protein K and phenol/chloroform[144] and further subjected to RNase treatment/purification. The D. melanogaster genomic DNA was extracted from whole bodies of several individuals by Puregene blood core kit B (Qiagen). The molecular weights of both *Drosophila* and YH1 genomic DNAs were confirmed to be larger than 23 kbp by gel electrophoresis (not shown), with no detection of degradation or RNA/protein contamination, and quantified by Quant-iT dsDNA HS assay kit (0.2-100 ng; Invitrogen, Q32854, Carlsbad, CA, USA); of each, an aliquot was diluted to 25 ng/µl for use in library construction. P. aeruginosa PAO1 strains were selected for tobramycin resistance at 16 mg/l (41 strains), ciprofloxacin resistance at 4 mg/l (47 strains), or no antibiotic resistance (8 strains). DNA was then isolated using the Wizard® SV 96 genomic DNA purification system (Promega, Madison, WI, USA). Concentrations of isolated DNA were measured using a Nanodrop (Thermo Scientific, Waltham, MA, USA).

Mechanical shearing: Shearing was performed on two 5-µg aliquots of E. coli CC118 genomic DNA brought up to 200 µl with TE. Sonication was performed by two 15-min treatments with a Bioruptor sonicator (Wolf Laboratories, Pocklington, UK) at maximum settings in a cold room, switching out the water between treatments to keep samples cool. After sonication, each sample was cleaned up using a QIAquick PCR purification kit, eluting in 30 µl Buffer EB (Qiagen). CRW10 and PA1 phage DNA was fragmented using 5 µg aliquots brought to 100 µl with TE and mixed with 500 µl nebulization buffer (Roche/454, Branford, CT, USA) in a nebulizer cup. Nebulization was carried out using 45 psi (310 kPa) nitrogen for 1 min on ice. The sheared DNA was cleaned and concentrated using MinElute columns and eluting in Buffer EB (Qiagen).

Enzymatic fragmentation: Enzymatic fragmentation was performed on six 1 µg aliquots of genomic DNA from both E. coli CC118 and H. sapiens NA18507 added to 2 µl 10× fragmentation reaction buffer, 0.2 µl 100× BSA (NEBiolabs, Ipswich, MA, USA), 2 µl NEB fragmentase enzyme, and nuclease-free water (Ambion, Austin, TX, USA) to 20 µl. Reactions were gently vortexed and spun-down, then incubated on ice for 5 min followed by a time-course incubation at 37°C, removing samples at 15, 20, 30, 45, 60, and 120 min for each of the two sets. Reactions were stopped by placing on ice followed by purification using a QIAquick PCR purification kit, eluting in 30 µl buffer EB (Qiagen). 4 µl of each sample was run on a Novex TBE gel (Invitrogen) to observe size distribution. The 60-min time point showed the desired size range and duplicate samples were prepared by the same method for each organism. Fragmentation of phage DNAs were performed using 1 µg aliquots of PA1 and CRW10 in a 20 µl reaction volume including fragmentation reaction buffer to 1×, BSA, and 2 µl NEB fragmentase enzyme (NEB). The reaction was incubated on ice for 5 min followed by incubation at 37°C for 20 min and stopped with 5 µl 500 mM EDTA and cleaned using MinElute columns (Qiagen).

Post-fragmentation library preparation: Post-fragmentation library preparation on the duplicate samples of E. coli CC118 sonication and fragmentase (60 min), and *H. sapiens* NA18507 fragmentase (60 min), was carried out as per standard Illumina methods, including a size selection using Novex TBE gels (Invitrogen), excising the 400-500 bp band. Final PCR amplification was carried out on a Bio-Rad (Hercules, CA, USA) MiniOpticon using SYBR Green I as a dye to monitor amplification. 1 µl of each final library was run on a Novex TBE gel (Invitrogen) for library size confirmation. Nebulized or fragmentase-treated phage samples were size-selected using SPRI beads (Beckman Coulter, Danvers, MA, USA) and used to construct libraries according to standard protocols, including end-polishing, adaptor ligation, fill-in, and single-strand isolation. Adaptor sequences included multiplex identifiers (barcodes).

Transposase-based library preparation: Transposase-based library preparation for *E. coli* CC118 and *H. sapiens* NA18507 Illumina-compatible libraries used 50 ng of genomic DNA brought up to 15 µl in nuclease-free water (Ambion) followed by the addition of 4 µl 5× LMW Nextera reaction buffer and 1 µl Nextera enzyme mix (Illumina-compatible; Epicentre), followed by a gentle vortex and brief centrifugation. Each reaction tube was incubated at 55°C in a thermocycler with a heated lid for 5 min followed by placement on ice and immediate purification using a QIAquick PCR purification kit and elution in 20 µl buffer EB (Qiagen). Suppression PCR was then carried out using 10 µl of the eluate as template with 11.5 µl nuclease-free water (Ambion), 25 µl 2× Nextera PCR buffer, 0.5 µl SYBR Green, 1 µl 50× Nextera primer cocktail (Illumina-compatible), 1 µl 50× Nextera adaptor 2 (barcodes 1-2 for E. coli and 3-4 for H. sapiens), and 1 µl Nextera PCR enzyme. The reaction was cycled in a Bio-Rad MiniOpticon to monitor the reaction under the following conditions: (1×) 3:00 min at 72°C and (1×) 0:30 min at 95°C, followed by 13 cycles of [0:10 min at 95°C, 0:30 min at 62°C, 3:00 min at 72°C] for *E. coli* barcodes 1 and 2 and H. sapiens barcodes 3 and 4 (12 cycles for *H. sapiens* barcode 4). 1 µl of each post-PCR library was electrophoresed through a Novex TBE gel (Invitrogen) for library size confirmation. Size selection of the E. coli CC118 transposase libraries was carried out at Epicentre Biotechnology using Agencourt AMPure (> 300 bp size selection), Zymo DNA purification (no size selection), or Caliper (350 ± 10% bp size selection) methods. The *D. melanogaster* library was constructed by pooling two standard Nextera reactions following the manufacturer's protocol (Epicentre). For each reaction, 50 ng genomic DNA was initially tagmented (*in vitro* transposase-catalyzed adaptor insertion) at 55°C for 5 min, and then MinElute purified. This was followed by PCR amplification with same conditions as with H. sapiens and *E. coli* libraries for 12 cycles. 400-450 bp gel-based size selection was carried out prior to sequencing.

A total of seven *H. sapiens* YH1 libraries were constructed, differing in mass of DNA, number of PCR cycles, and selected DNA fragment size. These included two (about 500 bp and about 550 bp) produced from pooling five standard Nextera reactions, three (400-500 bp, 500-550 bp and 550-600 bp) produced from pooling two modified reactions with nine cycles of PCR enrichment, and another two libraries (300-500 bp and 500-650 bp) from a single tagmentation reaction using 500 ng starting DNA with five cycles of PCR enrichment. The insert-size distribution and final yields for the *Drosophila* and *H. sapiens* YH1 libraries were validated separately using a 2100 Bioanalyzer (DNA 1000 and 7500 kit; Agilent, Santa Clara, CA, USA) and quantitative PCR.

P. aeruginosa PAO1 Illumina-compatible shotgun libraries were prepared for each strain using Epicentre Biotechnologies' Nextera DNA sample preparation kits with a customized, unique 9 bp barcode sequence for each strain. The tagmentation reaction consisted of 200 ng PAO1 DNA, 25 µl Nextera high molecular weight buffer, 1 µl Nextera transposase enzyme, and water to a total volume of 20 µl. The reaction was incubated for 5 min at 55°C, cleaned using Qiagen MiniElute columns, and eluted in 11 µl water. PCR reactions included 5 µl of the fragmented DNA, 17 µl water, 25 µl Nextera PCR buffer, 1 µl Nextera PCR enzyme, 1 µl of a Nextera primer cocktail containing two short primers (at 10 µM each) and one long Illumina-compatible adaptor (at 5 µM), and 1 µl of the barcode containing Illumina adaptor (at 5 µM). PCR conditions used were the same as above using 12 cycles of amplification, followed by MinElute clean-up as before. Samples were run on a Novex TBE polyacrylamide gel to confirm library quality, and DNA concentrations measured using a Nanodrop.

For the Roche (454)-compatible libraries, standard Nextera reaction conditions were used with 50 ng CRW10 (barcode 11) or PA1 (barcode 10) bacteriophage DNA, 454-Titanium compatible kit components and standard PCR methods, cycling 15 times. The PCR products were purified using Qiagen MinElute columns. Library fragment sizes were assessed using an Agilent Bioanalyzer DNA1000 chip.

Targeted sequence capture of the human exome: Libraries were prepared by transposase-catalyzed adaptor insertion by previously described methods using 50 ng genomic DNA (BK229.03, SFARI-SSC), 1 µl transposomes, 4 µl 5× HMW buffer, and water to 20 µl. Samples were incubated at 55°C for 5 min then cleaned up (AMPure) and eluted in 20 µl followed by the addition of 25 µl 2× Nextera PCR buffer, 1 µl 50× Nextera primer cocktail, 1 µl Nextera PCR enzyme, 0.5 µl 100× SYBR Green, and 1 µl of a barcoded adaptor (Table S4 in Additional file 3) with water to 50 µl. Reactions were carried out on a Bio-Rad MiniOpticon using recommended cycling conditions for 12 rounds. Each tube was cleaned up (AMPure, Agencourt, Boston, MA, USA) and checked for size and quantity on an Agilent Bioanalyzer DNA 1000

chip. One sample was selected for capture using all of the 414.6 ng for hybridization to Nimblegen (Madison, WI, USA) SeqCap EZ Exome probes v1.0 as per Nimblegen protocols using custom blockers (Nextera_Block1: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CCT CCC TCG CGC CAT CAG AGA TGT GTA TAA GAG ACA G-3', Nextera_Block1_REV: 5'-CTG TCT CTT ATA CAC ATC TCT GAT GGC GCG AGG GAG GCG TGT AGA TCT CGG TGG TCG CCG TAT CAT T-3', Nextera_Block2: 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT GCC TTG CCA GCC CGC TCA GAG ATG TGT ATA AGA GAC AG-3', Nextera_Block2_REV: 5'-CTG TCT CTT ATA CAC ATC TCT GAG CGG GCT GGC AAG GCA GAC CGA TCT CGT ATG CCG TCT TCT GCT TG-3') for 72 h at 47°C. After hybridization, wash was performed as per Nimblegen protocols with streptavidin-coupled magnetic beads. Finally, PCR amplification was performed on exome captured library (Post_Cap_Short_For_Amp: 5'-AAT GAT ACG GCG ACC ACC GAG ATC T-3', Post_Cap_Short_Rev_Amp: 5'-CAA GCA GAA GAC GGC ATA CGA GAT-3'; 1× [0:30 min at 98°C], 17× [0:10 min at 98°C, 0:30 min at 65°C, 0:45 min at 72°C]) followed by clean up (AMPure) and sequencing on an Illumina GAIIx SE36 run.

PCR-free library preparation: Adaptor sequences (NoPCR1: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CCT CCC TCG CGC CAT CAG AGA TGT GTA TAA GAG ACA G-3', and NoPCR2: 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT GCC TTG CCA GCC CGC TCA GAG ATG TGT ATA AGA GAC AG-3') were designed to contain the original 'Nextera' adaptor sequences, but with an additional 5' overhang of either P1 or P2 on adaptor 1 or adaptor 2, respectively (i.e. sequences to make compatible with cluster PCR on Illumina flow-cell), thus eliminating the need to add them during a PCR step. The 5' phosphorylated reverse compliment of the 19 bp mosaic end (ME: 5'-Phos-CTG TCT CTT ATA CAC ATC T-3') sequence was hybridized to NoPCR1/2 by combining 5 µl of each NoPCR1 and NoPCR2 with 10 µl ME reverse complement all at 100 µM with 80 µl TE, followed by denaturation at 95°C for 5 min then slow cooling to room temperature for a final annealed adaptor concentration of 10 µM. Transposomes were assembled by incubating 5 µl annealed adaptors at 10 µM with 5 µl 100% glycerol and 10 µl Ez-Tn5 transposase (Epicentre) and allowed to incubate at room temperature for 20 min.

Tagmentation was carried out using previously prepared E. coli (CC118) or human (NA18507) genomic DNA using either 100 or 200 ng of DNA, 5 µl prepared transposomes, 2 µl 5× Nextera HMW buffer (Epicentre), and water to 10 µl. Reactions were incubated at 55°C for 5 min, followed by the addition of 25 µl 2× FailSafe PCR master mix (Epicentre), 1 µl FailSafe DNA polymerase (Epicentre), and 14 µl nuclease-free water (Ambion), and subsequent 5 min incubation at 72°C for nick translation. Tubes were then cleaned up using Qiaquick MinElute PCR purification columns (Qiagen), eluting in 12 µl buffer EB.

For each reaction, 2 µl was used as template for a real time PCR on a MiniOpticon (Bio-Rad) using 0.5 µl SYBR Green, 25 µl 2× Nextera PCR master mix, 1 µl Nextera PCR enzyme and nuclease-free water to 50 µl. Alongside the NoPCR reactions, libraries of known concentrations were used as template at successive dilutions to be used as a standard for rough library quantification. Standard Nextera cycling conditions were used, without the initial 72°C extension step. PCR reactions were cleaned up by Qiaquick PCR purification columns and run on a Noved 6% TBE PAGE gel (Invitrogen) for library size verification. After quantification, libraries were sequenced as per standard Illumina GAIIx protocol as a paired-end 36 bp run.

Low input transposase-based library preparation: For the 500 pg and 100 *pg E. coli* (CC118) libraries and 10 pg human library (NA18507, Coriell), genomic DNA (in 1 µl volume) was incubated with 1 µl Nextera Illumina-compatible transposomes (Epicentre) at a 1 to 50 dilution (1 µl Nextera enzyme, 24 µl TE, 25 µl 100% glycerol), 1 µl 5× Nextera HMW buffer, and 2 µl nuclease-free water (Ambion). To avoid contamination, all dilutions and reaction preparation was carried out in a PCR hood. Reactions were incubated at 55°C for 5 min followed by addition of 25 µl 2× Nextera PCR buffer, 0.5 µl SYBR Green, 1 µl 50× Nextera primer cocktail, and 1 µl 0.5 µM barcode adaptor 2 (barcodes A6, A9, or A4 for 500 pg and 100 pg E. coli DNA, or 10 pg human DNA, respectively) and cycled under standard Nextera conditions in a MiniOpticon (Bio-Rad) real-time PCR thermocycler. Both reactions were removed after 20 cycles and cleaned up using Qiaquick MinElute columns, eluting in 20 µl EB. Libraries were run on a 6% Novex TBE PAGE gel (Invitrogen) for size verification and sequenced as barcoded spike-ins as per standard Illumina GAIIx protocol as a paired-end 101 bp (plus 9 bp barcode) run for E. coli libraries and a paired-end 36 bp (plus 9 bp barcode) run for human.

Direct colony-based library preparation: Fusion-Blue chemically competent *E. coli* (Clontech, Mountain View, CA, USA) were transformed with pUC19 bearing a 2 kbp insert of human genomic DNA, and then plated on Luria broth (LB) + ampicillin. A small number of cells were picked from a single bacterial colony with a 10 µl pipette tip, and then transferred with dipping into 15 µl nuclease-free H2O. The suspended cells were heat-lysed at 95°C for 5 min, then placed on ice for 2 min. Nextera 5× LMW reaction buffer and enzyme (Illumina-compatible; Epicentre) were added to the sample, followed by brief mixing and incubation at 55°C for 5 min. The reaction was then stopped by heating to 70°C for 15 min. Sequencing-compatible primer sites were added in a 50 µl PCR reaction using 5 µl of the transposase reaction directly as template without intervening purification. PCR was carried out with 31.6 µl H2O, 10 µl Kapa 2G robust A buffer 5× (Kapa Biosystems, Cape Town, South Africa), 1 µl dNTP mix (10 mM each), 0.25 µl SYBR Green 100×, 1 µl 50× Nextera primer cocktail, 1 µl 50× Nextera adaptor 2, and 0.20 µl Kapa 2G robust polymerase; cycling conditions were as described by Epicentre. The amplification reaction was cleaned up with a Qiaquick PCR clean-up column (Qiagen) and eluted into 50 µl EB.

Sequencing: Sequencing of the *H. sapiens* NA18507, *E. coli* CC118, and *D. melanogaster* libraries was done on an Illumina Genome Analyzer IIx as paired-end 36, 36, and 45 bp runs, respectively, using standard read primers for sonication and fragmentase libraries, run in individual lanes, and Nextera read primers for Nextera libraries. *H. sapiens* YH1 libraries were run on an Illumina HiSeq2000 as paired-end 90 bp run using Nextera read primers. Phage libraries contained library-specific barcodes and were run as multiplexed samples using GS FLX Titanium sequencing protocols. *P. aeruginosa* libraries were pooled by combining 100 ng of each strain library, and sequenced on an Illumina Genome Analyzer IIx with a paired-end 76-cycle run.

Short read mapping: Short read mapping was done on the *E. coli* CC118, *D. melanogaster and H. sapiens* (NA18507 & YH1) Illumina GAIIx or HiSeq2000 sequenced samples by converting the raw sequence files to fastq format and then mapping to the GRCh37 (hg19) reference using the BWA[76] alignment software. After mapping, PCR duplicates were removed, as well as read-pairs with an insert size shorter than that of the read length.

Long read assembly: Long read assembly from bacteriophage samples sequenced on the Roche GS FLX Ti was done using Roche's newbler assembler under default parameters. Individual reads from each dataset were mapped against the assembled genome using gsMapper (Roche Software Release: 2.3 (091027_1459)).

Fragmentation site characterization: Fragmentation site characterization was carried out by stacking all regions of the genome flanking forward strand mapping start locations and the reverse complement of reverse strand start locations followed by calculating nucleotide frequencies at each position relative to the fragmentation site, thus generating a positional weight matrix (PWM). The PWMs then were imported into the SeqLogo (Oliver Bembom, Dept. of Biostatistics, University of California, Berkeley, 2008) package for Bioconductor in R and used to generate positional information content (IC) and sequence logos using the equation outlined by T. D. Schneider et al.[145146]:

$$\mathrm{IC}(w)=\log_2(J)+\Sigma_{Jj=1}\, p_{wj}\,\log_2(p_{wj})=\log_2(J)-entropy(w)$$

where J is the number of variables in the alphabet (4; A, C, G, or T), and j is the base at position w. This equation does not factor in the background nucleotide frequencies.

Normalization of Illumina GAIIx coverage: Normalization of Illumina GAIIx coverage for the *E. coli* (non-size-selected) data was done by dividing the coverage at each position in the genome by the total number of mapped bases and then multiplying by a constant (the average number of mapped bases was close to 1 Gb, therefore 109 was used as the constant).

Coverage distribution histograms: Coverage distribution histograms were generated by calculating the number of times each base of the genome was sequenced and plotting the frequency of each level of coverage.

Coverage by G+C content: Coverage by G+C content plots were generated by binning the reference genome into 500 bp bins for E. coli, 10 kbp bins for human, and 1 kbp bins for *Drosophila* (other sizes were also investigated, resulting in very similar distributions) and calculating the G+C content of each bin, followed by plotting the coverage of that bin.

Library complexity: Library complexity was calculated by random sampling of 50,000 read-pairs without replacement and plotting the number of uniquely occurring read-pairs versus the total number of sampled read-pairs.

Insertion size distributions: Insertion size distributions were generated by taking the distance between the start mapping location of the first read and the end mapping location of the second read for every read-pair and plotting the frequency of occurrences of each insert size.

SNP calls: SNP calls for the YH1 genome were generated using the SAMtools[78] variant caller with a maximum coverage of 1,000 and minimum quality score of 30. Prior to variant calling, read-pairs with an insert size less than 90 bp and reads not properly paired were removed to reduce noise. Calls were then compared with the SNPs reported by Wang et al. (2008)[75] and to dbSNP build 129.

Cell-line-discordant SNP calls:

In order to minimize false calls due to mapping errors, a repeat-masked version of GRCh36 was used along with further masking with respect to mappability according to the UCSC Genome Browser 'Rosetta 35mer uniqueness' and excluding all regions with a score of 0 (this score means that the sequence maps perfectly to multiple locations in the genome). This track was used because it was generated using the BWA aligner that was used in our analysis, and because the original YH1 sequence data is made up of 36 bp reads. Out of this newly masked genome, positions were called that had a SNP quality score in the cell line over 30, a reference call in blood over 30, and coverage less than 100× in both datasets. Of those with a quality score of 50 in both for their calls, 100 were randomly chosen for validation.

Validation was carried out using a mass spectrometry assay (Sequenom, San Diego, CA, USA). Primers for PCR amplification and extension were successfully designed for 100 mutation sites using the Sequenom MassArray Assay Design v3.1. PCR amplification, shrimp alkaline phosphatase treatment of unincorporated dNTPs, probe extension and resin desalting were carried out in sequence using the conditions described elsewhere[147]. Sequenom genotyping was performed in parallel for genomic DNA from YH1 blood and the same batch of lymphoblastoid cell lines as was used for sequencing. A negative control and technical replicate were also run in parallel for each typed position. Genotyping of all 100 testing sites passed the filter criteria of: (1) no failing extension, (2) no false positive in the negative control, (3) consistency between two technical replicates. Genotyping was further performed for the 100 positions in the YH primary lymphoblastoid cell lines using the same method, with 98 meeting the above filter criteria.

Exome analysis: Sequence reads from transposase-based libraries subjected to human exome capture were aligned to the human reference (GRCh36) using BWA[76]. Each aligned base was deemed to be on target if it was within 100 bases of a targeted sequence. At each position within target regions, coverage was assessed and any position with a depth of at least one was considered covered. Comparison to standard exome methods was made by trimming read 1 of a PE76 lane from a GAIIx down to 36 bases and aligning it the same way. Because the standard library had fewer reads mapped, 28 million reads were randomly taken from both libraries and above analysis performed. Complexity was interrogated by taking an equivalent number of on-target SE36 reads generated by each method and calculating the percentage with unique start-points.

Barcode design: Barcode design yielded a set of 96 × 9 bp sequences in which each 9 bp sequence contained no homopolymer run of three or more bases, had a GC content of ≤ 60%, was a edit distance of at least four away from all other members of the set of 96, and screened negative when compared with other adaptor and primer sequences used here. Also, we took care to ensure that each base (A, G, C, or T) was represented at least once in each 9 bp barcode, and at least once in each position along the 9 bp.

Barcode deconvolution: Barcode deconvolution for pooled, multiplexed *Pseudomonas* samples was carried out by computing the Levenshtein edit distance between the obtained index read and each of the 96 barcode sequences used. The corresponding read-pair was assigned to a barcode when that barcode was within edit distance of 2 of the index read, with the next closest matching barcode being at least two further edits away.

A.2 Supplementary tables for Chapter 3

Summary of libraries sequenced in comparative analysis

Organism	Method	Size Selection	Sequencing Platform	Raw Read Pairs	Mapper	Raw Mapped	PCR Duplicates Removed†	Unique Mapped %	Complexity
E. coli CC118	Sonication, replicate 1	none	Illumina GAIIx (PE36)	14,764,330	BWA	13,096,300	12,719,017	86%	97%
	Sonication, replicate 2	none	Illumina GAIIx (PE36)	15,177,716	BWA	13,283,051	13,018,271	86%	98%
	Endonuclease, replicate 1	none	Illumina GAIIx (PE36)	15,131,182	BWA	13,374,850	12,775,619	84%	96%
	Endonuclease, replicate 2	none	Illumina GAIIx (PE36)	16,171,788	BWA	14,124,278	13,441,857	83%	95%
	Transposase, replicate 1	none	Illumina GAIIx (PE36)	19,812,714	BWA	16,439,998	13,209,901	67%	80%
	Transposase, replicate 2	none	Illumina GAIIx (PE36)	23,239,433	BWA	17,501,733	14,094,636	61%	81%
	Transposase	none	Illumina GAIIx (PE76)	47,445,883	BWA	32,763,991	24,138,660	51%	74%
	Transposase	Ampure >300	Illumina GAIIx (PE76)	49,202,822	BWA	28,370,120	21,618,428	44%	76%
	Transposase	Caliper 350+/-10%	Illumina GAIIx (PE76)	14,781,578	BWA	12,898,508	9,439,353	64%	73%
	Transposase NoPCR (100ng)	none	Illumina GAIIx (PE36)	N/A	BWA	512,202	510,620	N/A	100%
	Transposase NoPCR (200ng)	none	Illumina GAIIx (PE36)	N/A	BWA	1,492,999	1,475,918	N/A	99%
	Transposase NoPCR (50ng), 1	none	Illumina GAIIx (PE101)	spike in	BWA	417,268	414,738	N/A	99%
	Transposase NoPCR (50ng), 2	none	Illumina GAIIx (PE101)	spike in	BWA	416,216	413,342	N/A	99%
	Transposase (500pg)	none	Illumina GAIIx (PE101)	spike in	BWA	882,236	814,763	N/A	92%
	Transposase (100pg)	none	Illumina GAIIx (PE101)	spike in	BWA	1,232,098	1,097,646	N/A	89%
H. sapiens NA18507	Sonication, replicate 1**	none	Illumina GAIIx (PE36)	8,830,236	BWA	8,515,999	8,451,436	96%	99%
	Sonication, replicate 2**	none	Illumina GAIIx (PE36)	8,779,733	BWA	8,471,772	8,403,393	96%	99%
	Endonuclease, replicate 1	none	Illumina GAIIx (PE36)	10,137,423	BWA	8,499,831	8,295,592	82%	98%
	Endonuclease, replicate 2	none	Illumina GAIIx (PE36)	12,477,053	BWA	10,790,123	10,637,932	85%	99%
	Transposase, replicate 1	none	Illumina GAIIx (PE36)	18,378,252	BWA	15,814,133	15,602,480	85%	99%
	Transposase, replicate 2	none	Illumina GAIIx (PE36)	18,559,340	BWA	15,877,850	15,653,955	84%	99%
	Transposase NoPCR (100ng)	none	Illumina GAIIx (PE36)	N/A	BWA	1,044,539	1,037,373	N/A	99%
	Transposase NoPCR (200ng)	none	Illumina GAIIx (PE36)	N/A	BWA	488,026	485,966	N/A	99%
	Transposase (10pg)††	none	Illumina GAIIx (PE36)	N/A	BWA	5,731,004	2,182,087	N/A	38%
CRW10	Nebulization	none	Roche GS FLX Ti*	46,745	gsMapper	43,707	43,707	94%	N/A
	Endonuclease	none	Roche GS FLX Ti*	6,444	gsMapper	5,983	5,983	93%	N/A
	Transposase	none	Roche GS FLX Ti*	47,986	gsMapper	43,917	43,917	92%	N/A
PA1	Nebulization	none	Roche GS FLX Ti*	6,183	gsMapper	6,121	6,121	>99%	N/A
	Endonuclease	none	Roche GS FLX Ti*	9,979	gsMapper	9,412	9,412	94%	N/A
	Transposase	none	Roche GS FLX Ti*	43,639	gsMapper	40,637	40,637	93%	N/A
D. melanogaster w1118	Transposase (2rxn)	Gel 400-450	Illumina GAIIx (PE45)	31,272,321	BWA	30,918,821	28,142,311	90%	91%
H. sapiens YH1	Transposase (5rxn)	Gel 550-575	Illumina HiSeq2000 (PE90)	103,341,810	BWA	82,511,867	81,183,426	79%	98%
		Gel 400-500	Illumina HiSeq2000 (PE90)	68,022,215	BWA	62,183,263	60,709,520	89%	98%
	Transposase (2rxn)	Gel 400-500	Illumina HiSeq2000 (PE90)	67,823,613	BWA	62,293,546	60,817,189	90%	98%
		Gel 500-550	Illumina HiSeq2000 (PE90)	58,597,591	BWA	51,335,787	49,590,370	85%	97%
		Gel 550-650	Illumina HiSeq2000 (PE90)	52,625,788	BWA	47,144,748	45,258,959	86%	96%
	Transposase (1rxn)	Gel 300-500	Illumina HiSeq2000 (PE90)	68,342,929	BWA	58,105,589	56,850,508	83%	98%
		Gel 500-650	Illumina HiSeq2000 (PE90)	54,039,263	BWA	44,013,398	42,200,046	78%	96%
H. sapiens BK229.03	Transposase – Exome Capture	none	Illumina GAIIx (SE36)*	44,472,457	BWA	34,654,705	N/A	78%	N/A

* Single end read only
** Data for sonication method for human libraries taken as single lanes of data from Bentley et al. Nature 2008
† PCR Duplicates removed using the samtools rmdup function.
†† Duplicates removed by using outer mapping coordinates.

Table A.2.1. Summary of libraries sequenced in comparative analysis of library construction methods.

Information Content (-10 to +15bp)

		Average	Maximum
E. coli CC118	Sonication	0.007	0.102
	Endonuclease	0.015	0.108
	Transposase	0.046	0.157
H. sapiens NA18507	Sonication	0.031	0.048
	Endonuclease	0.024	0.144
	Transposase	0.049	0.153
CRW10	Nebulization	0.049	0.063
	Endonuclease	0.050	0.134
	Transposase	0.052	0.109
PA1	Nebulization	0.018	0.035
	Endonuclease	0.032	0.085
	Transposase	0.026	0.053

Table A.2.2. Information content of sequence bias in vicinity of fragmentation sites for all examined library construction methods.

Pearson Correlation of Coverage

Spearman Rank Order Correlation of Coverage	Son.1	Son.2	End.1	End.2	Tr.1	Tr.2
Son.1		0.816	0.653	0.678	0.336	0.366
Son.2	0.854		0.688	0.691	0.320	0.370
End.1	0.693	0.696		0.795	0.383	0.376
End.2	0.720	0.720	0.818		0.352	0.401
Tr.1	0.409	0.421	0.435	0.428		0.835
Tr.2	0.489	0.497	0.490	0.493	0.877	

Table A.2.3. Correlation of coverage between compared library construction methods.
Son. represents sonication replicates, End. represents endonuclease replicates, and Tr. represents transposase-mediated library construction replicates. Spearman correlations are to the right of the diagonal in red shading and Pearson correlations to the right in blue shading.

Standard "Adaptor 2"

CAAGCAGAAGACGGCATACGAGATCGGTCTGCCTTGCCAGCCCGCTCAG

Format for modified "Adaptor 2"

CAAGCAGAAGACGGCATACGAGAT#########CGGTCTGCCTTGCCAGCCCGCTCAG

Adaptor 2 with embedded barcode (9 bp x 96 sequences)	Cell # for IDT plate	Barcodes alone
CAAGCAGAAGACGGCATACGAGATTACGAAGTCCGGTCTGCCTTGCCAGCCCGCTCAG	A1	TACGAAGTC
CAAGCAGAAGACGGCATACGAGATGACGAGATTCGGTCTGCCTTGCCAGCCCGCTCAG	B1	GACGAGATT
CAAGCAGAAGACGGCATACGAGATACCGTAAGACGGTCTGCCTTGCCAGCCCGCTCAG	C1	ACCGTAAGA
CAAGCAGAAGACGGCATACGAGATTAGTGGCAACGGTCTGCCTTGCCAGCCCGCTCAG	D1	TAGTGGCAA
CAAGCAGAAGACGGCATACGAGATCATTAACGCCGGTCTGCCTTGCCAGCCCGCTCAG	E1	CATTAACGC
CAAGCAGAAGACGGCATACGAGATTCGTTGAAGCGGTCTGCCTTGCCAGCCCGCTCAG	F1	TCGTTGAAG
CAAGCAGAAGACGGCATACGAGATTAGTACGCTCGGTCTGCCTTGCCAGCCCGCTCAG	G1	TAGTACGCT
CAAGCAGAAGACGGCATACGAGATCTCAGATCACGGTCTGCCTTGCCAGCCCGCTCAG	H1	CTCAGATCA
CAAGCAGAAGACGGCATACGAGATTTCACCGTACGGTCTGCCTTGCCAGCCCGCTCAG	A2	TTCACCGTA
CAAGCAGAAGACGGCATACGAGATGTCATGCATCGGTCTGCCTTGCCAGCCCGCTCAG	B2	GTCATGCAT
CAAGCAGAAGACGGCATACGAGATAGGACAGTTCGGTCTGCCTTGCCAGCCCGCTCAG	C2	AGGACAGTT
CAAGCAGAAGACGGCATACGAGATATGGTGTCTCGGTCTGCCTTGCCAGCCCGCTCAG	D2	ATGGTGTCT
CAAGCAGAAGACGGCATACGAGATGGATGTTCTCGGTCTGCCTTGCCAGCCCGCTCAG	E2	GGATGTTCT
CAAGCAGAAGACGGCATACGAGATCTTATCCAGCGGTCTGCCTTGCCAGCCCGCTCAG	F2	CTTATCCAG
CAAGCAGAAGACGGCATACGAGATGTAAGTCACCGGTCTGCCTTGCCAGCCCGCTCAG	G2	GTAAGTCAC
CAAGCAGAAGACGGCATACGAGATTTCAGTGAGCGGTCTGCCTTGCCAGCCCGCTCAG	H2	TTCAGTGAG
CAAGCAGAAGACGGCATACGAGATCTCGTAATGCGGTCTGCCTTGCCAGCCCGCTCAG	A3	CTCGTAATG
CAAGCAGAAGACGGCATACGAGATCATGTCTCACGGTCTGCCTTGCCAGCCCGCTCAG	B3	CATGTCTCA
CAAGCAGAAGACGGCATACGAGATAATCGTGGACGGTCTGCCTTGCCAGCCCGCTCAG	C3	AATCGTGGA
CAAGCAGAAGACGGCATACGAGATGTATCAGTCCGGTCTGCCTTGCCAGCCCGCTCAG	D3	GTATCAGTC
CAAGCAGAAGACGGCATACGAGATAGCAGATGTCGGTCTGCCTTGCCAGCCCGCTCAG	E3	AGCAGATGT
CAAGCAGAAGACGGCATACGAGATTCCTAACGTCGGTCTGCCTTGCCAGCCCGCTCAG	F3	TCCTAACGT
CAAGCAGAAGACGGCATACGAGATAACAGTCCACGGTCTGCCTTGCCAGCCCGCTCAG	G3	AACAGTCCA
CAAGCAGAAGACGGCATACGAGATCCTTGAGAACGGTCTGCCTTGCCAGCCCGCTCAG	H3	CCTTGAGAA
CAAGCAGAAGACGGCATACGAGATTTAAGCCTGCGGTCTGCCTTGCCAGCCCGCTCAG	A4	TTAAGCCTG
CAAGCAGAAGACGGCATACGAGATTTAGACCACCGGTCTGCCTTGCCAGCCCGCTCAG	B4	TTAGACCAC
CAAGCAGAAGACGGCATACGAGATTGTCTAGTGCGGTCTGCCTTGCCAGCCCGCTCAG	C4	TGTCTAGTG
CAAGCAGAAGACGGCATACGAGATTAGATCGAGCGGTCTGCCTTGCCAGCCCGCTCAG	D4	TAGATCGAG
CAAGCAGAAGACGGCATACGAGATTGAATGCCACGGTCTGCCTTGCCAGCCCGCTCAG	E4	TGAATGCCA
CAAGCAGAAGACGGCATACGAGATGTGCAATGTCGGTCTGCCTTGCCAGCCCGCTCAG	F4	GTGCAATGT
CAAGCAGAAGACGGCATACGAGATAGTGGCATACGGTCTGCCTTGCCAGCCCGCTCAG	G4	AGTGGCATA
CAAGCAGAAGACGGCATACGAGATATGATCGGTCGGTCTGCCTTGCCAGCCCGCTCAG	H4	ATGATCGGT
CAAGCAGAAGACGGCATACGAGATAGTCTACCTCGGTCTGCCTTGCCAGCCCGCTCAG	A5	AGTCTACCT
CAAGCAGAAGACGGCATACGAGATGATCAACTGCGGTCTGCCTTGCCAGCCCGCTCAG	B5	GATCAACTG
CAAGCAGAAGACGGCATACGAGATATCGGTAGTCGGTCTGCCTTGCCAGCCCGCTCAG	C5	ATCGGTAGT
CAAGCAGAAGACGGCATACGAGATCGTATGATGCGGTCTGCCTTGCCAGCCCGCTCAG	D5	CGTATGATG
CAAGCAGAAGACGGCATACGAGATTTACTGACGCGGTCTGCCTTGCCAGCCCGCTCAG	E5	TTACTGACG
CAAGCAGAAGACGGCATACGAGATCTGTCGTAACGGTCTGCCTTGCCAGCCCGCTCAG	F5	CTGTCGTAA
CAAGCAGAAGACGGCATACGAGATTCAACTGGTCGGTCTGCCTTGCCAGCCCGCTCAG	G5	TCAACTGGT
CAAGCAGAAGACGGCATACGAGATATCGATCTCCGGTCTGCCTTGCCAGCCCGCTCAG	H5	ATCGATCTC
CAAGCAGAAGACGGCATACGAGATGCAACTATGCGGTCTGCCTTGCCAGCCCGCTCAG	A6	GCAACTATG
CAAGCAGAAGACGGCATACGAGATGATGACTTCCGGTCTGCCTTGCCAGCCCGCTCAG	B6	GATGACTTC
CAAGCAGAAGACGGCATACGAGATGACGTTACACGGTCTGCCTTGCCAGCCCGCTCAG	C6	GACGTTACA

CAAGCAGAAGACGGCATACGAGATCATCTGCTACGGTCTGCCTTGCCAGCCCGCTCAG	D6	CATCTGCTA
CAAGCAGAAGACGGCATACGAGATATTAGTCGGCGGTCTGCCTTGCCAGCCCGCTCAG	E6	ATTAGTCGG
CAAGCAGAAGACGGCATACGAGATTAGCGTACTCGGTCTGCCTTGCCAGCCCGCTCAG	F6	TAGCGTACT
CAAGCAGAAGACGGCATACGAGATCCAAGCAATCGGTCTGCCTTGCCAGCCCGCTCAG	G6	CCAAGCAAT
CAAGCAGAAGACGGCATACGAGATCCGTAATTGCGGTCTGCCTTGCCAGCCCGCTCAG	H6	CCGTAATTG
CAAGCAGAAGACGGCATACGAGATAGAATTGCCCGGTCTGCCTTGCCAGCCCGCTCAG	A7	AGAATTGCC
CAAGCAGAAGACGGCATACGAGATACCTGTAACCGGTCTGCCTTGCCAGCCCGCTCAG	B7	ACCTGTAAC
CAAGCAGAAGACGGCATACGAGATCATCAGTGTCGGTCTGCCTTGCCAGCCCGCTCAG	C7	CATCAGTGT
CAAGCAGAAGACGGCATACGAGATGAATCCTCACGGTCTGCCTTGCCAGCCCGCTCAG	D7	GAATCCTCA
CAAGCAGAAGACGGCATACGAGATGCTGTATACCGGTCTGCCTTGCCAGCCCGCTCAG	E7	GCTGTATAC
CAAGCAGAAGACGGCATACGAGATGAAGGCTATCGGTCTGCCTTGCCAGCCCGCTCAG	F7	GAAGGCTAT
CAAGCAGAAGACGGCATACGAGATGGAATCGATCGGTCTGCCTTGCCAGCCCGCTCAG	G7	GGAATCGAT
CAAGCAGAAGACGGCATACGAGATGCTTATGGTCGGTCTGCCTTGCCAGCCCGCTCAG	H7	GCTTATGGT
CAAGCAGAAGACGGCATACGAGATTGACGCATTCGGTCTGCCTTGCCAGCCCGCTCAG	A8	TGACGCATT
CAAGCAGAAGACGGCATACGAGATCACGATTCTCGGTCTGCCTTGCCAGCCCGCTCAG	B8	CACGATTCT
CAAGCAGAAGACGGCATACGAGATTATTGCCTCCGGTCTGCCTTGCCAGCCCGCTCAG	C8	TATTGCCTC
CAAGCAGAAGACGGCATACGAGATAAGTCAGAGCGGTCTGCCTTGCCAGCCCGCTCAG	D8	AAGTCAGAG
CAAGCAGAAGACGGCATACGAGATATAGCTGAGCGGTCTGCCTTGCCAGCCCGCTCAG	E8	ATAGCTGAG
CAAGCAGAAGACGGCATACGAGATTGCTCACAACGGTCTGCCTTGCCAGCCCGCTCAG	F8	TGCTCACAA
CAAGCAGAAGACGGCATACGAGATGTCTTCTGACGGTCTGCCTTGCCAGCCCGCTCAG	G8	GTCTTCTGA
CAAGCAGAAGACGGCATACGAGATTTGCCGATTCGGTCTGCCTTGCCAGCCCGCTCAG	H8	TTGCCGATT
CAAGCAGAAGACGGCATACGAGATCTCGAATACCGGTCTGCCTTGCCAGCCCGCTCAG	A9	CTCGAATAC
CAAGCAGAAGACGGCATACGAGATTGGCTTCTACGGTCTGCCTTGCCAGCCCGCTCAG	B9	TGGCTTCTA
CAAGCAGAAGACGGCATACGAGATAAGGCCATTCGGTCTGCCTTGCCAGCCCGCTCAG	C9	AAGGCCATT
CAAGCAGAAGACGGCATACGAGATAAGTTGACCCGGTCTGCCTTGCCAGCCCGCTCAG	D9	AAGTTGACC
CAAGCAGAAGACGGCATACGAGATCTGAACTGACGGTCTGCCTTGCCAGCCCGCTCAG	E9	CTGAACTGA
CAAGCAGAAGACGGCATACGAGATCTAGGTGTACGGTCTGCCTTGCCAGCCCGCTCAG	F9	CTAGGTGTA
CAAGCAGAAGACGGCATACGAGATCCATCTTAGCGGTCTGCCTTGCCAGCCCGCTCAG	G9	CCATCTTAG
CAAGCAGAAGACGGCATACGAGATCTACGACATCGGTCTGCCTTGCCAGCCCGCTCAG	H9	CTACGACAT
CAAGCAGAAGACGGCATACGAGATTCCAACATGCGGTCTGCCTTGCCAGCCCGCTCAG	A10	TCCAACATG
CAAGCAGAAGACGGCATACGAGATGCTATCATCCGGTCTGCCTTGCCAGCCCGCTCAG	B10	GCTATCATC
CAAGCAGAAGACGGCATACGAGATACAGCTTCACGGTCTGCCTTGCCAGCCCGCTCAG	C10	ACAGCTTCA
CAAGCAGAAGACGGCATACGAGATAGTCATTGCCGGTCTGCCTTGCCAGCCCGCTCAG	D10	AGTCATTGC
CAAGCAGAAGACGGCATACGAGATAGATCTCGACGGTCTGCCTTGCCAGCCCGCTCAG	E10	AGATCTCGA
CAAGCAGAAGACGGCATACGAGATATGCTCTTGCGGTCTGCCTTGCCAGCCCGCTCAG	F10	ATGCTCTTG
CAAGCAGAAGACGGCATACGAGATTTAGTGCGTCGGTCTGCCTTGCCAGCCCGCTCAG	G10	TTAGTGCGT
CAAGCAGAAGACGGCATACGAGATTCCTAGTTCCGGTCTGCCTTGCCAGCCCGCTCAG	H10	TCCTAGTTC
CAAGCAGAAGACGGCATACGAGATGGTGCATTACGGTCTGCCTTGCCAGCCCGCTCAG	A11	GGTGCATTA
CAAGCAGAAGACGGCATACGAGATACTGAGGATCGGTCTGCCTTGCCAGCCCGCTCAG	B11	ACTGAGGAT
CAAGCAGAAGACGGCATACGAGATTAGCAGTCACGGTCTGCCTTGCCAGCCCGCTCAG	C11	TAGCAGTCA
CAAGCAGAAGACGGCATACGAGATTCACTCGAACGGTCTGCCTTGCCAGCCCGCTCAG	D11	TCACTCGAA
CAAGCAGAAGACGGCATACGAGATACCAATCAGCGGTCTGCCTTGCCAGCCCGCTCAG	E11	ACCAATCAG
CAAGCAGAAGACGGCATACGAGATGATATGGACCGGTCTGCCTTGCCAGCCCGCTCAG	F11	GATATGGAC
CAAGCAGAAGACGGCATACGAGATTGAGAGATCCGGTCTGCCTTGCCAGCCCGCTCAG	G11	TGAGAGATC
CAAGCAGAAGACGGCATACGAGATTGCCATTAGCGGTCTGCCTTGCCAGCCCGCTCAG	H11	TGCCATTAG
CAAGCAGAAGACGGCATACGAGATACTAACGCACGGTCTGCCTTGCCAGCCCGCTCAG	A12	ACTAACGCA
CAAGCAGAAGACGGCATACGAGATATGTAGCACCGGTCTGCCTTGCCAGCCCGCTCAG	B12	ATGTAGCAC
CAAGCAGAAGACGGCATACGAGATGGTCGATATCGGTCTGCCTTGCCAGCCCGCTCAG	C12	GGTCGATAT
CAAGCAGAAGACGGCATACGAGATGCGAGTTATCGGTCTGCCTTGCCAGCCCGCTCAG	D12	GCGAGTTAT
CAAGCAGAAGACGGCATACGAGATGACTGAGTACGGTCTGCCTTGCCAGCCCGCTCAG	E12	GACTGAGTA
CAAGCAGAAGACGGCATACGAGATAGATACTCCCGGTCTGCCTTGCCAGCCCGCTCAG	F12	AGATACTCC
CAAGCAGAAGACGGCATACGAGATGCTAGAGTTCGGTCTGCCTTGCCAGCCCGCTCAG	G12	GCTAGAGTT
CAAGCAGAAGACGGCATACGAGATAATGTAGCGCGGTCTGCCTTGCCAGCCCGCTCAG	H12	AATGTAGCG

Table A.2.4. Barcode primer design and oligos used for 96-plex sample indexing.
"Standard Adaptor 2" refers to the non-indexed primer, "Format for modified adaptor 2" designates where index bases are placed by the use of "#" symbols. The 96 indexed primers contain the index sequence in red with the flanking adaptor priming sequence in black.

A.3 Supplementary figures for Chapter 3

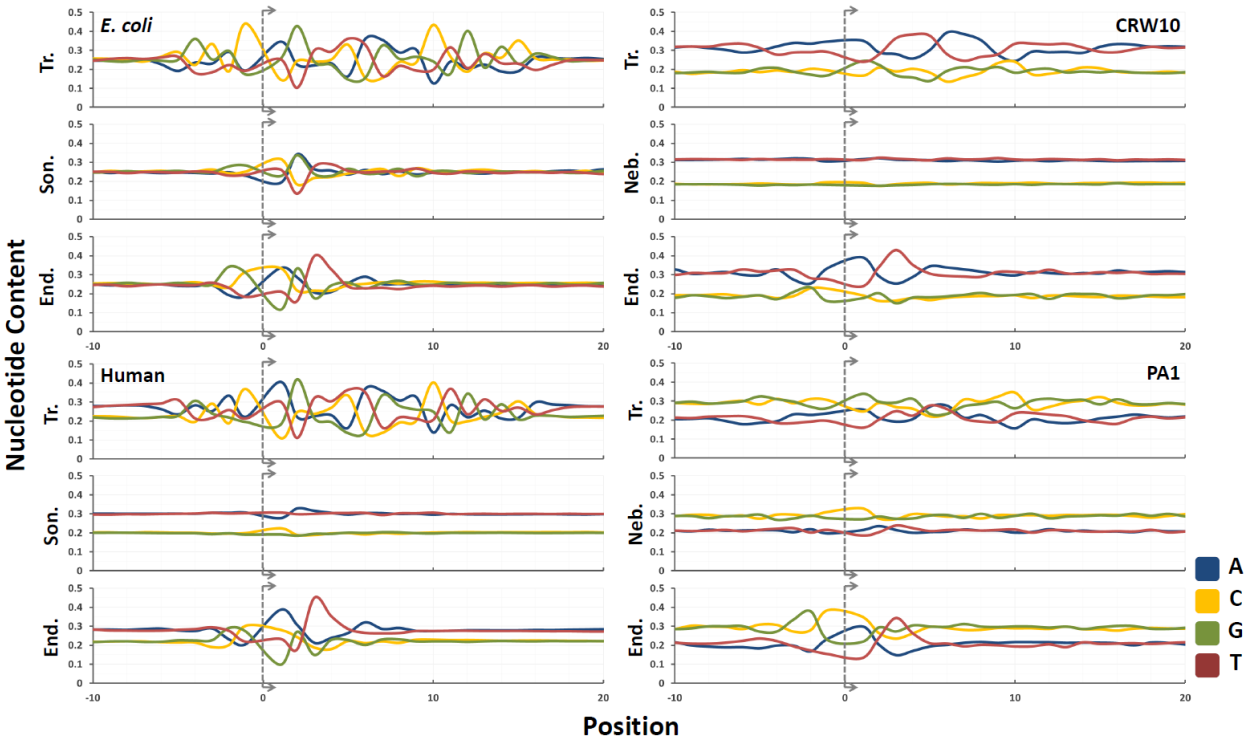


Figure A.3.1. Fragmentation site profiles.
Nucleotide content over a 30 bp interval (-10, +20) corresponding to read starts for each method of library construction (Tr. = Transposase, Son. = Sonication, Neb. = Nebulization, End. = Endonuclease). Observed biases are shown for Illumina GAIIx sequenced libraries (E. coli (CC118), human (NA18507)) and Roche (454) GSFLX sequenced phage libraries (CRW10, PA1).

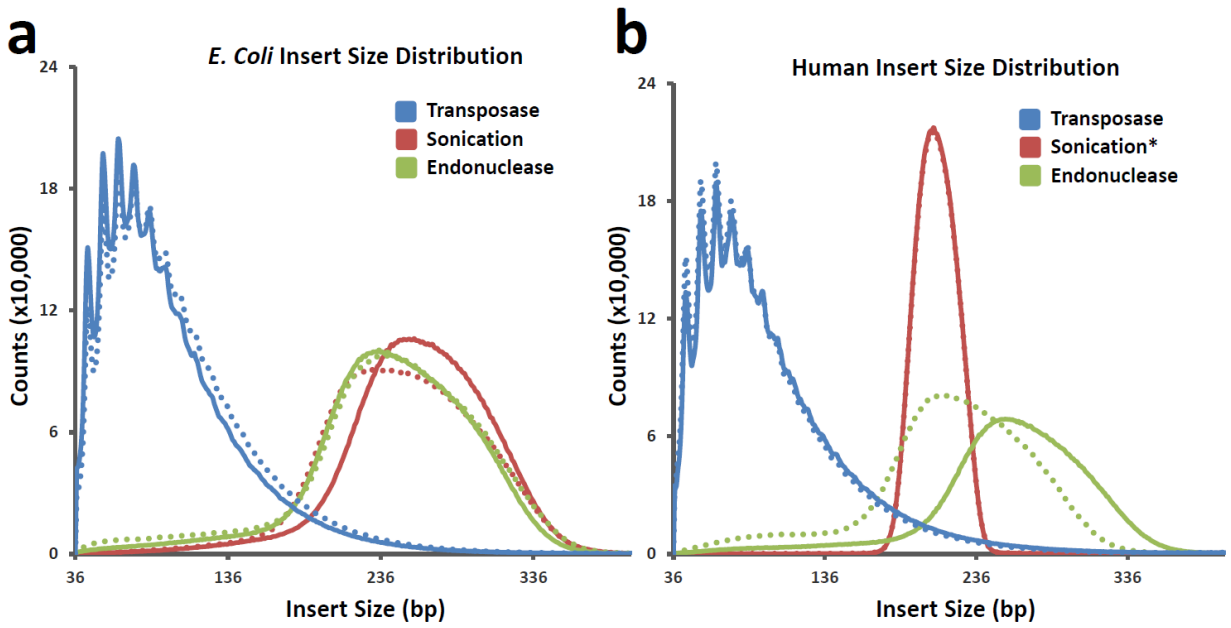


Figure A.3.2. Insert size distributions.
E. coli CC118 (**a**) and *H. sapiens* NA18507 (**b**) libraries of each method sequenced on the Illumina GAIIx platform with replicates (dashed). *The H. sapiens library prepared by sonication is from Bentley et al. and had two size selections resulting in a tight distribution. In both organisms, the transposase based method shows a slight periodicity at ~10 bp intervals, likely due to physical constraints on the ability for a transposase to attack certain positions with regards to the helical pitch of the DNA as it extends away from the bound transposase.

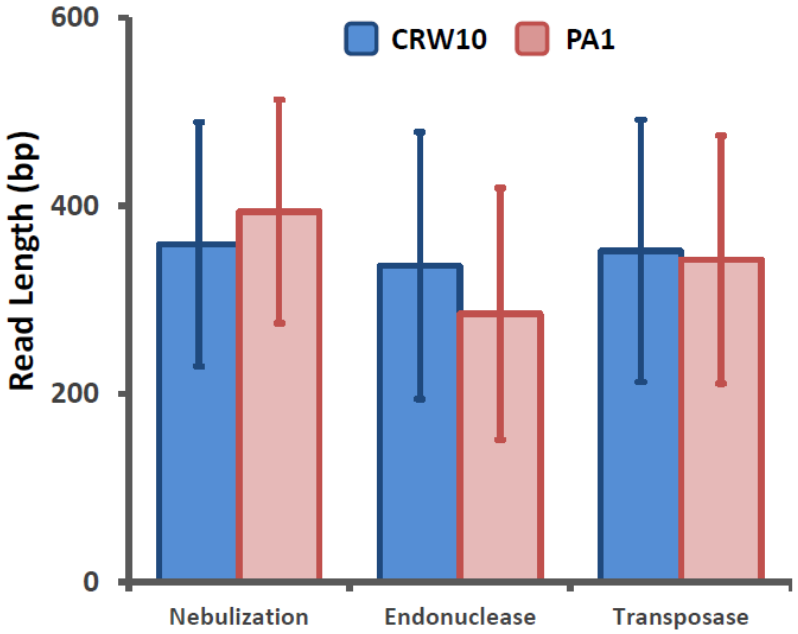


Figure A.3.3. Average 454 read lengths.
Generated for Roche (454) GS FLX Ti libraries for CRW10 and PA1 bacteriophages using nebulization, endonuclease, and transposase methods.

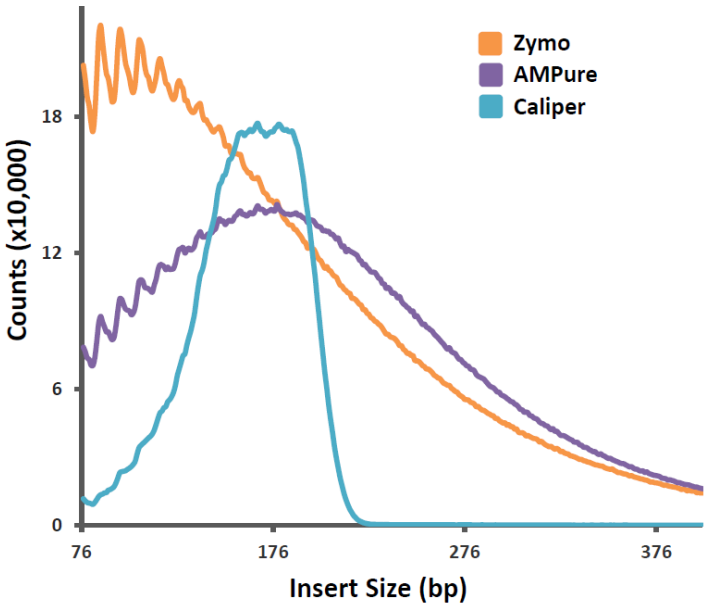


Figure A.3.4. Insert size distributions for various size selection methods.
Zymo column clean-up (orange), AMPure bead clean-up (300 bp cutoff, purple), and Caliper chip-based size selection (350 +/- 10 bp, teal).

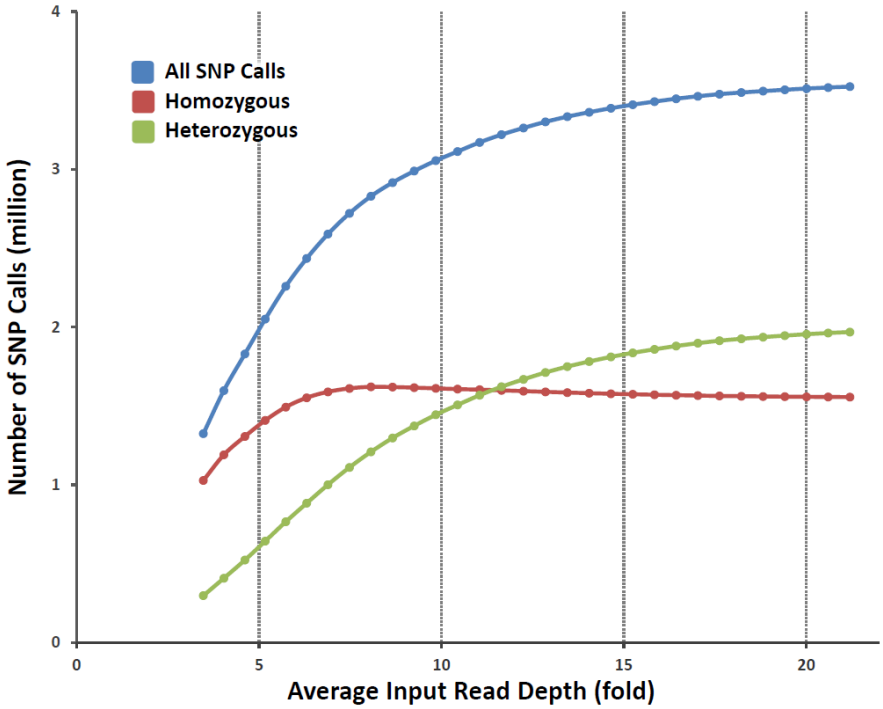


Figure A.3.5. Number of SNP calls in YH1 genome by read depth.
SNP calls versus sequence depth were calculated by pulling random sets of 10 million mapped read pairs (properly paired, insert size >=90 bp) without replacement and calling SNP positions of a minimum quality of 30. Trend observed is consistent with previously reported whole genome analyses (Bentley et. al. (2008))[18].

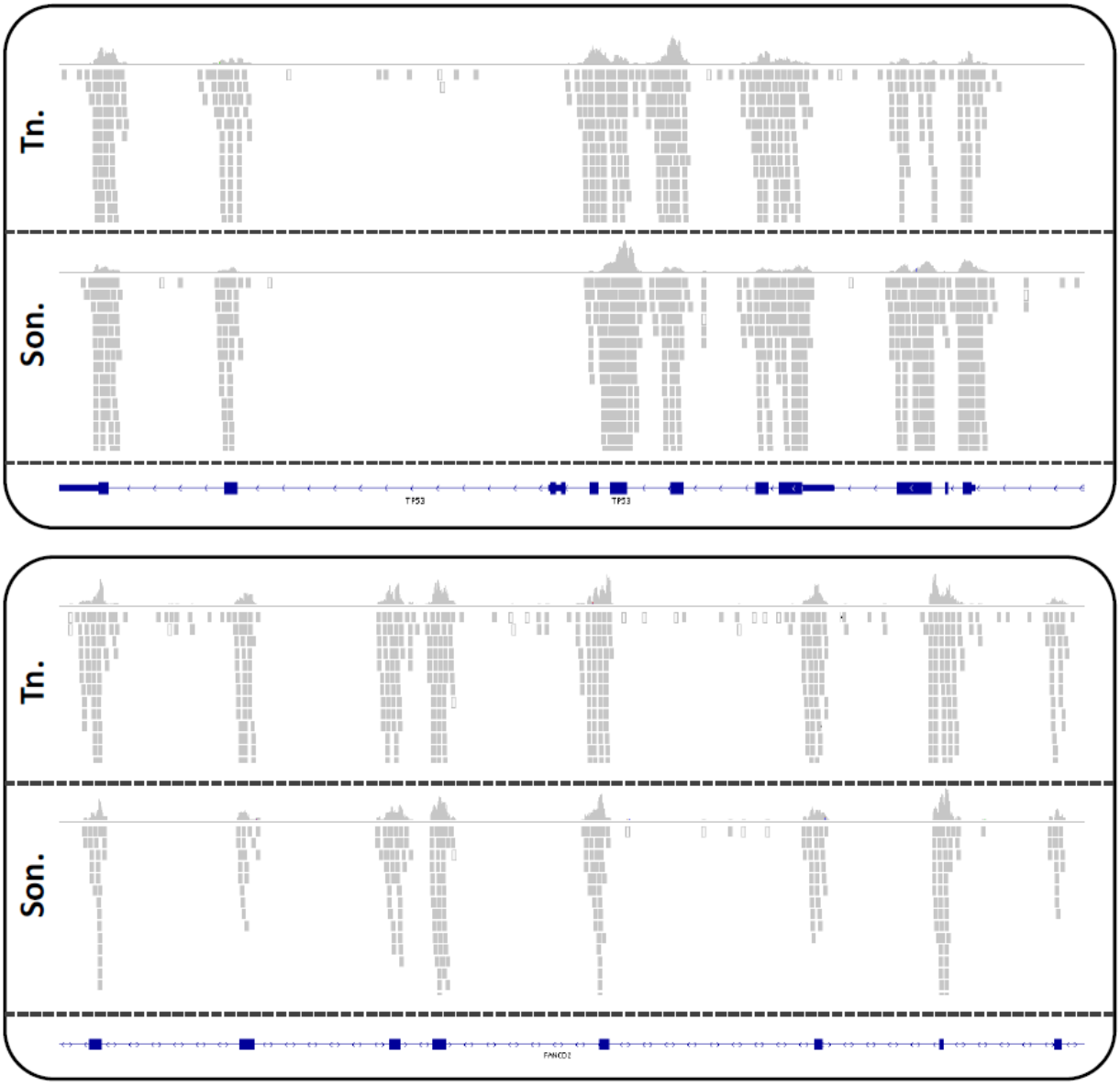


Figure A.3.6. Sample coverage profiles for transposase-mediated exome sequencing library construction.

Standard (sonication-based) libraries from 3 µg of starting material and transposase-based libraries from 50 ng of starting material were subjected to whole human exome target capture and sequenced on an Illumina GAIIx. Shown are two typical example views of reads aligned to the genome for an identical number of mapped reads of each method. (Tn. = Transposase-based, Son. = Sonication-based)

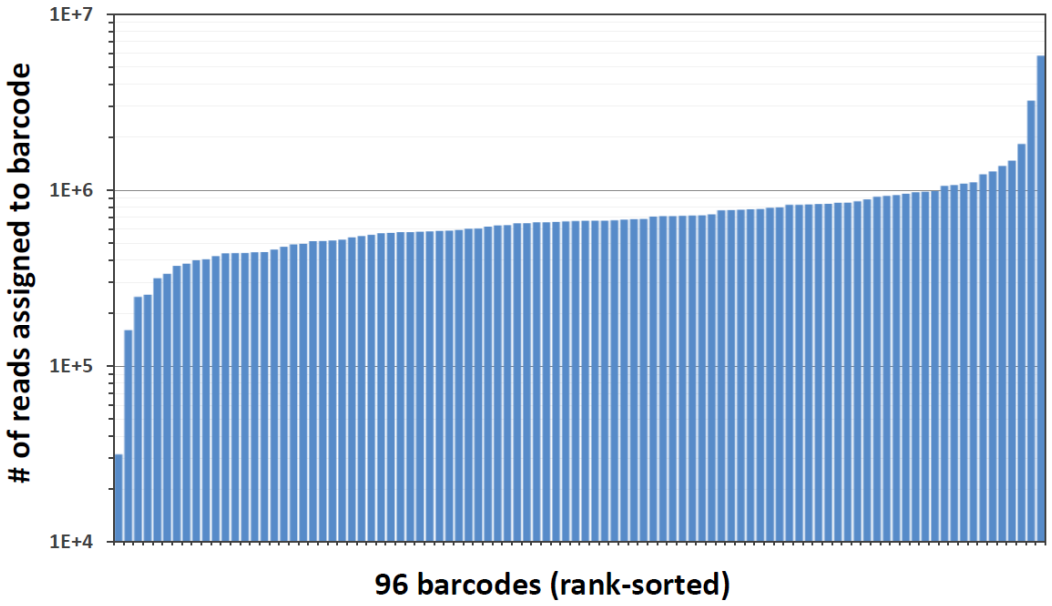


Figure A.3.7. Uniformity of 96-plex barcode indexing for transposase-mediated library construction.
Libraries were constructed from 96 DNAs using adaptors bearing 9 bp index sequences. Samples were quantified by Nanodrop prior to normalized pooling. The number of reads assigned to each index is plotted here. 90% (86 of 96) fall within a 4-fold range of relative abundance.

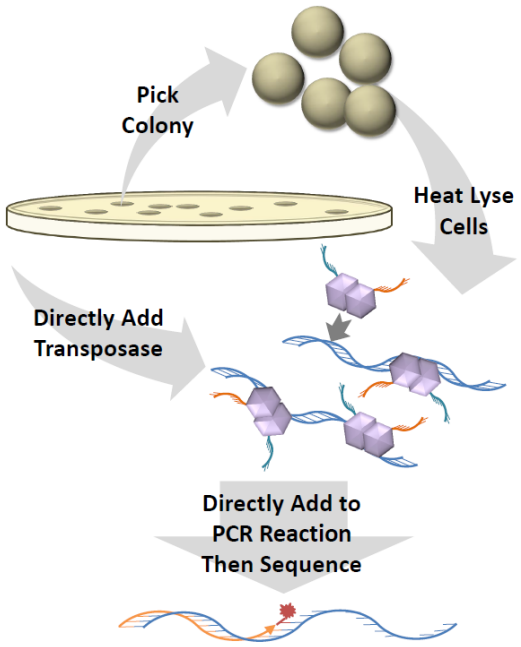


Figure A.3.8. Schematic for "Colony-PCR" based transposase-mediated library construction.
A colony is picked using a pipette tip followed by lysing the cells by boiling. The lysate is then used as the material for transposase-catalyzed adaptor insertion without clean-up then directly used in the PCR reaction, also without cleanup. After PCR a column purification is performed and the library is sequencer ready.

Appendix B: Supplementary material for Chapter 4

B.1 Supplementary methods for Chapter 4

Tn5mC-seq library construction and sequencing: Transposome complexes were generated by incubating 2.5 µL of 10 µM Tn5mC-A1 (Tn5mC-A1top: 5′-GAT [5mC] TA [5mC] A[5mC] G [5mC] [5mC] T [5mC] [5mC] [5mC] T [5mC] G [5mC] G [5mC] [5mC] AT [5mC] AGA GAT GTG TAT AAG AGA CAG-3′, IDT, annealed to Tn5mC-A1bot: 5′-[Phos]-CTG TCT CTT ATA CAC A-3′, IDT, by incubating 10 µL of each oligo at 100 µM and 80 µL of EB [QIAGen] for 2 min at 95°C and then cooling to room temperature at 0.1°C/sec) with 2.5 µL 100% glycerol and 5 µL Ez-Tn5 transposase (Epicentre – Illumina) for 20 min at room temperature.

Genomic DNA prepared from GM20847 cell lines was used at respective input quantities with 4 µL Nextera HMW Buffer (Epicentre-Illumina), nuclease-free water (Ambion) to 17.5 µL and 2.5 µL prepared Tn5mC transposomes (regardless of the quantity of DNA used). Reactions were incubated for 9 min at 55°C in a thermocycler followed by SPRI bead cleanup (AMPure) using 36 µL of beads and the recommended protocol with elution in 14 µL nuclease-free water (Ambion). Adaptor 2 annealing was then carried out by adding 2 µL of 10× Ampligase Reaction Buffer (Epicentre-Illumina), 2 µL 10× dNTPs (2.5 mM each; Invitrogen), and 2 µL 10 µM Tn5mC-A2top (5′-/5Phos/ CTG TCT CTT ATA CAC ATC T [5mC] TGA G [5mC] GGG [5mC] TGG [5mC] AAG G [5mC] AGA [5mC] [5mC] GAT [5mC]-3′; IDT) to each reaction and incubating for 2 min at 50°C followed by 10 min at 45°C and cooling at 0.1°C/sec to 37°C and subsequent incubation for 10 min. Gap-repair was then performed by adding 3 µL of Ampligase at 5U/µL (Epicentre-Illumina) and 1 µL of either T4 DNA Polymerase (Tn5mC libraries A-G, NEB) or Sulfolobus DNA Polymerase IV (Tn5mC libraries H-J, NEB) and additional incubation for 30 min at 37°C. Reactions were then cleaned up using SPRI beads (AMPure) according to recommended protocol using 36 µL beads and elution in 50 µL nuclease-free water (Ambion). Bisulfite treatment was performed using an EZ DNA Methylation Kit (Zymo) according to recommended protocols with a 14-h 50°C incubation and 10 µL elution. Eluate was then used as the template for PCR using 12.5 µL Kapa 2G Robust HotStart ReadyMix (Kapa Biosystems), 1 µL 10 µM Tn5mC-P1 (5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG CCT CCC TCG CGC CAT CAG-3′; IDT), 1 µL 10 µM barcoded P2 (from Adey et al. 2010[59]), 0.15 µL 100× SYBR Green (Invitrogen), and 0.35 µL nuclease-free water (Ambion). Thermocycling was carried out on a BioRad Opticon Mini real-time machine with the following parameters: 5 min at 95°C; (15 sec at 95°C; 15 sec at 62°C; 40 sec at 72°C; Plate Read; 10 sec at 72°C) × 99. Reactions were monitored and removed from thermocycler as soon as plateau was reached (12–15 cycles).

Sequencing was carried out using either a full or partial lane on an Illumina HiSeq2000 using custom sequencing primers: read 1, Tn5mC-R1 (5′-GCC TCC CTC GCG CCA TCA GAG ATG TGT ATA AGA GAT AG-3′; IDT); index read, Tn5mC-Ix (5′-TTG TTT TTT ATA TAT ATT TCT GAG CGG GCT GGC AAG GC-3′; IDT); and read 2, Tn5mC-R2 (5′-GCC TTG CCA GCC CGC TCA GAA ATA TAT ATA AAA AAC AA-3′; IDT). Read lengths were either single-read at 36 bp with a 9-bp index (SE36, libraries A and B, not included in Table 4.1) or 101 bp paired-end with a 9-bp index (PE101, libraries C–J). Libraries were only sequenced on runs that did not have lanes containing Nextera libraries as a precaution due to the similarity between sequencing primers.

Tn5mC-Seq 1.1 library preparation (Supplementary Fig. B.2.1d) was carried out as previously described with several modifications: (1) Transposase recognition sequence reverse compliment is 3′ blocked to prevent nonspecific extension in final PCR. (2) Replacement oligo is methylated through the region complementary to the transposase recognition sequence to maintain complexity during bisulfite conversion and allow the use of standard Nextera sequencing primers. (3) Replacement oligo is 3′ blocked to prevent degradation by 3′→5′ exonuclease activity of gap-repair polymerase (replacement oligo: Tn5mC1.1-A2top 5′-/5Phos/ [5mC] TGT [5mC] T [5mC] TTA TA [5mC] A [5mC] AT [5mC] T [5mC] TGA G [5mC] GGG [5mC] TGG [5mC] AAG G [5mC] AGA [5mC] [5mC] GA [inv dT]-3′, IDT; blocked transposase recognition sequence end: Tn5mC1.1-A1bot3block 5′-[Phos]-CTG TCT CTT ATA CA [ddC]-3′). Duplicate libraries were prepared from 100 ng, 10 ng, and 1 ng of starting material and were subject

to PCR amplification using either Kapa HiFi U+ Hot Start Ready Mix, or Kapa 2G Robust Hot Start Ready Mix (Kapa Biosystems) and were sequenced on a single-end 36-bp read plus a 9-bp index read run on an Illumina GAIIx. Library characterization can be found in Supplemental Fig. B.2.5.

Ligation chemistry WGBS library construction and sequencing: We subjected 1000, 100, and 10 ng of genomic DNA prepared from GM18507 cell lines to ligation chemistry–based library preparation according to methods described by Lister et al. (2009)[83] with several minor exceptions: (1) Bisulfite conversion was carried out using an EZ DNA Methylation Kit (Zymo), and (2) PCR was carried out using Kapa 2G Robust Hot Start Ready Mix (Kapa Biosystems). The change in PCR enzyme was due to several unpublished experiments demonstrating a much higher efficiency with Kapa Robust as opposed to PfuTurbo Cx used according to the method described by Lister et. al.(2009) [83]. Sequencing was performed on an Illumina MiSeq instrument using a single-end 100-bp sequence read run.

Read filtering and alignment: The GRCh37 reference genome was first bisulfite-converted in silico for both the top (C changed to T, C2T) and bottom (G changed to A, G2A) strands. Prior to alignment, reads were filtered based on the run metrics, as several libraries were run on lanes in which instrument valve failures resulted in poor quality or reads consisting primarily of N bases. Filtering was carried out by first calculating the base compositions as well as mean base quality scores at each position in the read. Many of the lanes had significantly reduced quality scores at the start and/or end of the read and were globally trimmed to remove any start or end positions that had a mean phed score of less than or equal to 10. The start and ends of the reads were additionally globally trimmed if a position within the first or last 25 bases of the read had a mean composition of 10% Ns, which generally corresponded to the quality-based trimming. Additionally, reads that contained three or more Ns were also removed. It is important to note that the reduced qualities in the runs were “flowcell-wide” regardless of the library that was run and not isolated to Tn5mC-seq libraries. Subsequent runs for the Tn5mC-seq 1.1 and polymerase testing experiments did not suffer instrument failures, and no trimming of the reads was necessary. Next, reads were aligned to both the C2T and G2A strands using BWA with default parameters. Reads that aligned to both strands were removed. Read pairs in which neither aligned to either strand were then pulled and trimmed to 76 bp (except for SE36 runs) and again aligned to both C2T and G2A strands. Duplicate reads (pairs sharing the same start positions for both reads 1 and 2) were removed and complexity determined. Reads with an alignment score less than 10 were then filtered out prior to secondary analysis. Total fold coverage was calculated using the total bases aligned from unique reads over the total alignable bases of the genome (slightly below 3 Gb per strand).

5mC calling: Methylated cytosines were called using a binomial distribution as in the method described by Lister et al. (2009)[83], whereby a probability mass function is calculated for each methylation context (CpG, CHG, CHH) using the number of reads covering the position as the number of trials and reads maintaining cytosine status as successes with a probability of success based on the total error rates that were determined by the combined nonconversion rate and sequencing error rate. The total error rate was initially determined by unmethylated lambda DNA spike-ins; however, we found that the error rate estimation from the gap-repair portion of reads (as described in the main text) gave a more comprehensive estimate, which was slightly higher than that of the lambda estimate. Therefore to be conservative, we used the highest determined error rate at 0.009. If the probability was below the value of M, where $M \times$ (number of total unmethylated CpG) $< 0.01 \times$ (number of total methylated CpG), the position was called as being methylated, thus enforcing that no more than 1% of positions would be due to the error rate.

B.2 Supplementary figures for Chapter 4

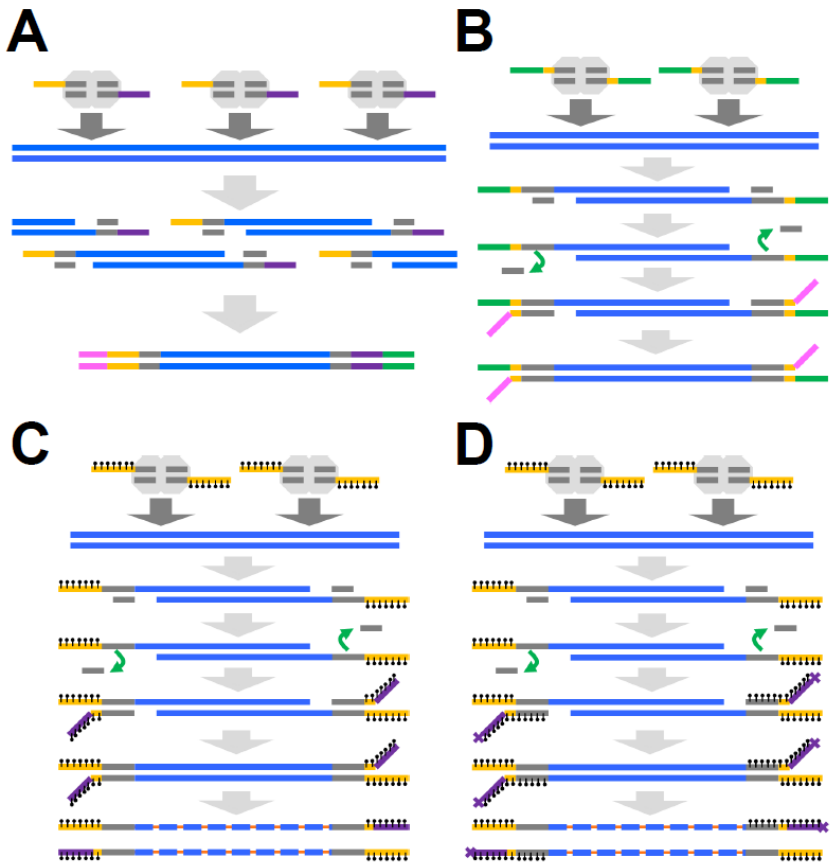


Figure B.2.1. Transposase-mediated library construction.
(A) Standard DNA-seq transposase based library construction as in Figure 1A. (B) Oligo replacement strategy. Loaded transposase attacks genomic DNA with a single adaptor (yellow/green) that also contains one of the flowcell primer sequences (green). An oligo-replacement approach then anneals a second adaptor (yellow/pink) that terminates in the second flowcell primer (pink) which is then subject to gap-repair. Resulting library can then be directly loaded onto the sequencer without a PCR step. (Gruenwald et. al. (2011)) (C) Tn5mC-seq 1.0 library preparation as in Figure 1B. (D) Tn5mC-seq 1.1 library preparation. Libraries are constructed as in Tn5mC-seq 1.0 with several modifications. 1) Replacement oligo is methylated through the region complementary to the transposase recognition sequence to maintain complexity during bisulfite conversion and allow the use of standard Nextera sequencing primers. 2) Replacement oligo is 3' blocked to prevent degradation by 3'->5' exonuclease activity of gap-repair polymerase.

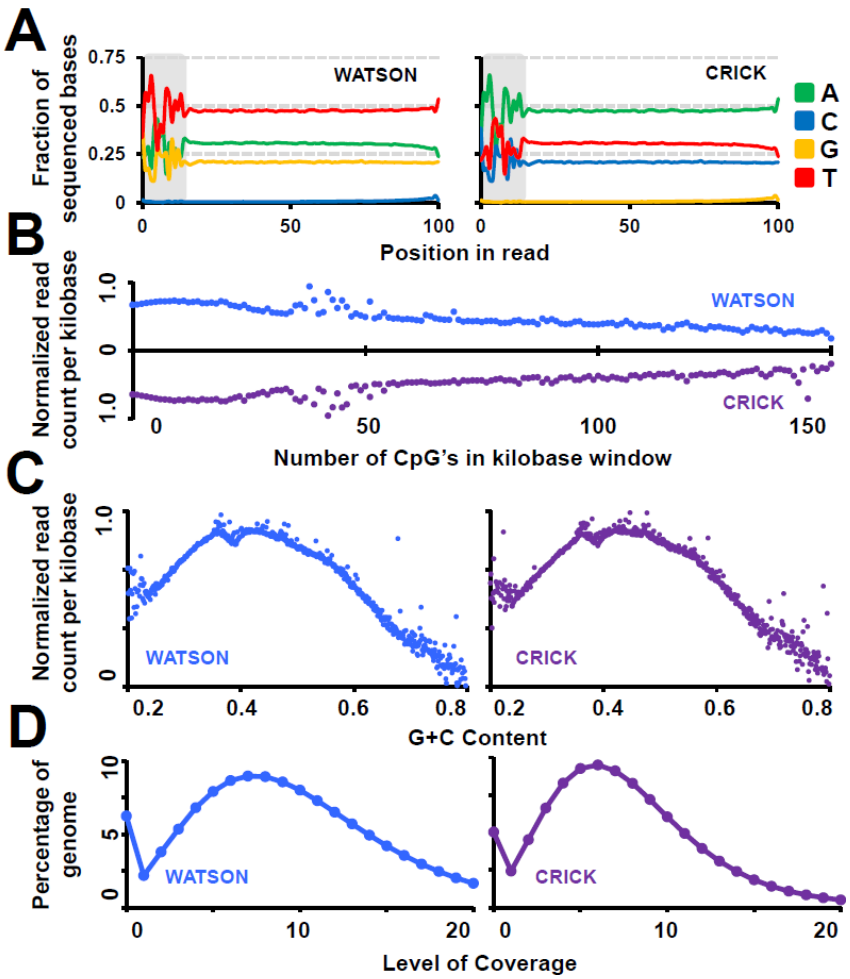


Figure B.2.2. Tn5mC-seq library characteristics.
(A) Base composition of aligned reads to each respective strand. Gray box represents transposon bias. (B) Read uniformity by CpG density. The human genome reference (hg19, GRC h37) was divided into kilobase windows and CpG's within each window tallied. The normalized number of reads aligning to each window were then tallied and averaged for each CpG density level and plotted. Ideal uniformity would be a horizontal line. Expansion in the 25-50 CpG density region occurs due to repetitive elements. (C) Read uniformity by G+C content. The human genome reference (hg19, GRC h37) was divided into kilobase windows and G+C content calculated and rounded to the thousandth. The normalized number of reads aligning to each window were then tallied and averaged for each G+C fraction and plotted. The typical PCR bias of reduced coverage at G+C extremes is observed. (D) Coverage uniformity across each strand. The percentage of the genome at corresponding levels of coverage is plotted. A slight decrease in coverage was observed in the 'Crick' (bottom) strand.

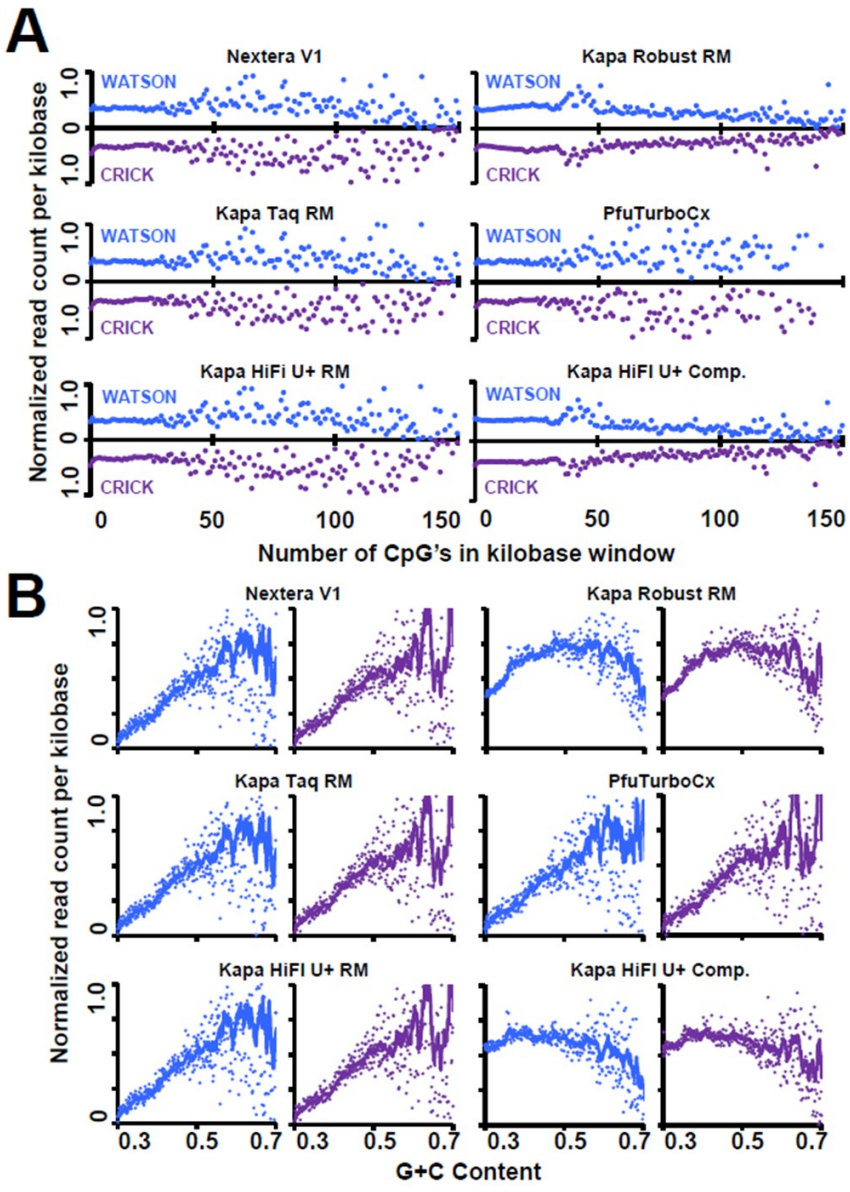


Figure B.2.3. PCR Comparisons.
Six Tn5mC-seq libraries were generated using 100 ng starting genomic DNA as in methods and pooled prior to PCR. One sixth of the pool was used as template in PCR using one of the following polymerases: Nextera V1 (Epicentre, Illumina), Kapa 2G Robust Hot Start Ready Mix, Kapa Taq Hot Start Ready Mix, PfuTurboCx (Stratagene), Kapa HiFi U+ Hot Start Ready Mix, and Kapa HiFi U+ Hot Start comp. (individual components). With the exception of Kapa Robust and Kapa HiFI U+ comp., library amplification was extremely poor and limited numbers of reads were aligned to the reference resulting in significant noise. (A) Coverage of each library by CpG density is represented as in Supplemental Figure S2. With the exception of Kapa Robust and Kapa HiFI U+ comp., CpG coverage uniformity is extremely noisy. Kapa Robust and Kapa HiFi U+ comp. perform equivalently well with respect to coverage uniformity across CpG density windows. (B) Coverage with respect to G+C content is represented as in Supplemental Figure S2. Solid lines represent the moving average across 10 windows. Again, with the exception of Kapa Robust and Kapa HiFI U+ comp., G+C bias profiles are extremely noisy. Kapa HiFI U+ comp. provided the most uniform coverage with respect to G+C content.

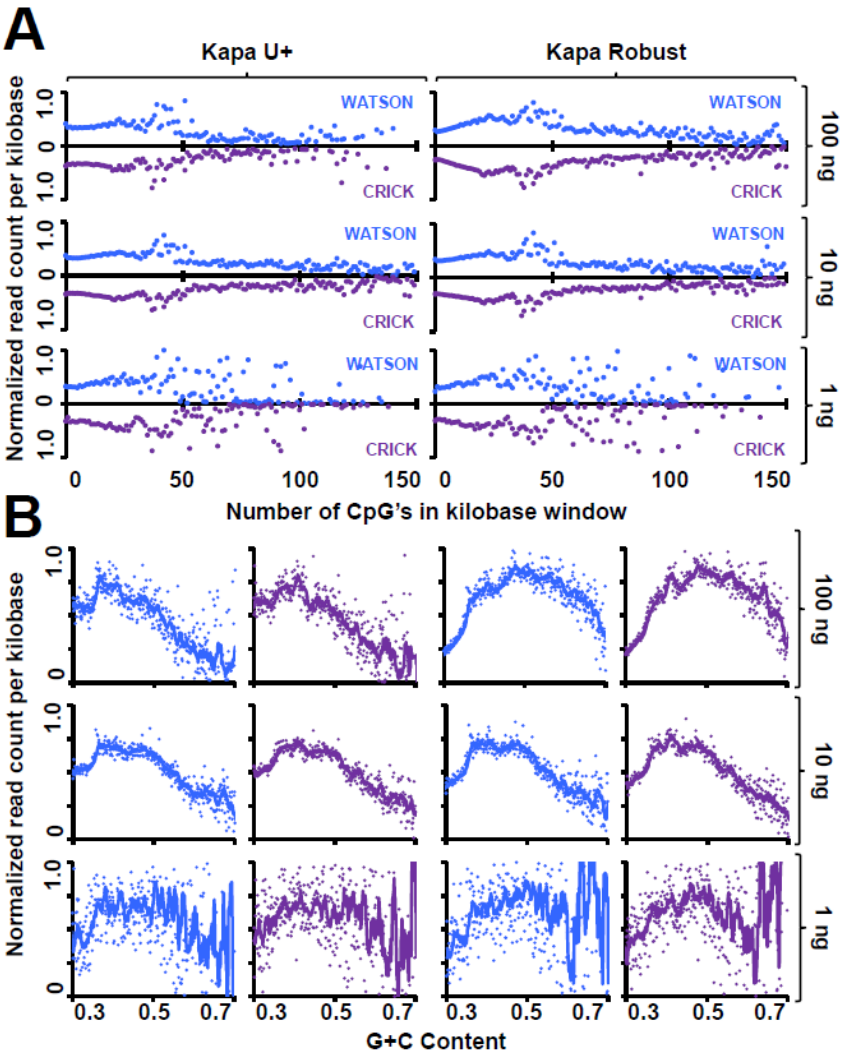


Figure B.2.4. Tn5mC-seq 1.1 library characteristics.
Duplicate libraries starting from 100 ng, 10 ng, and 1 ng were constructed using Tn5mC-seq 1.1 (described in methods and Supplemental Figure S1) and amplified using either Kapa HiFi U+ Hot Start Ready Mix, or Kapa 2G Robust Hot Start Ready Mix (Kapa Biosystems) with barcode adaptors and sequenced on a single end 36 bp read plus 9 bp index read run on an Illumina GAIIx. Reads were filtered and aligned as previously described. (A) Coverage of each library by CpG density is represented as in Supplemental Figure S2. Kapa HiFi U+ provides a slightly more uniform amplification. Interestingly, libraries constructed from 10ng of starting material exhibited the most uniform coverage with respect to CpG density. Libraries constructed from 1 ng of starting material showed greatly reduced uniformity. (B) Coverage with respect to G+C content is represented as in Supplemental Figure B.2.2. Solid lines represent the moving average across 10 windows. For libraries constructed from 100 ng, Kapa HiFi U+ exhibits greater coverage at extremely A+T rich regions of the genome. For libraries constructed from 10 ng and 1 ng amplification bias with respect to G+C content is comparable between the two PCR mixes tested.

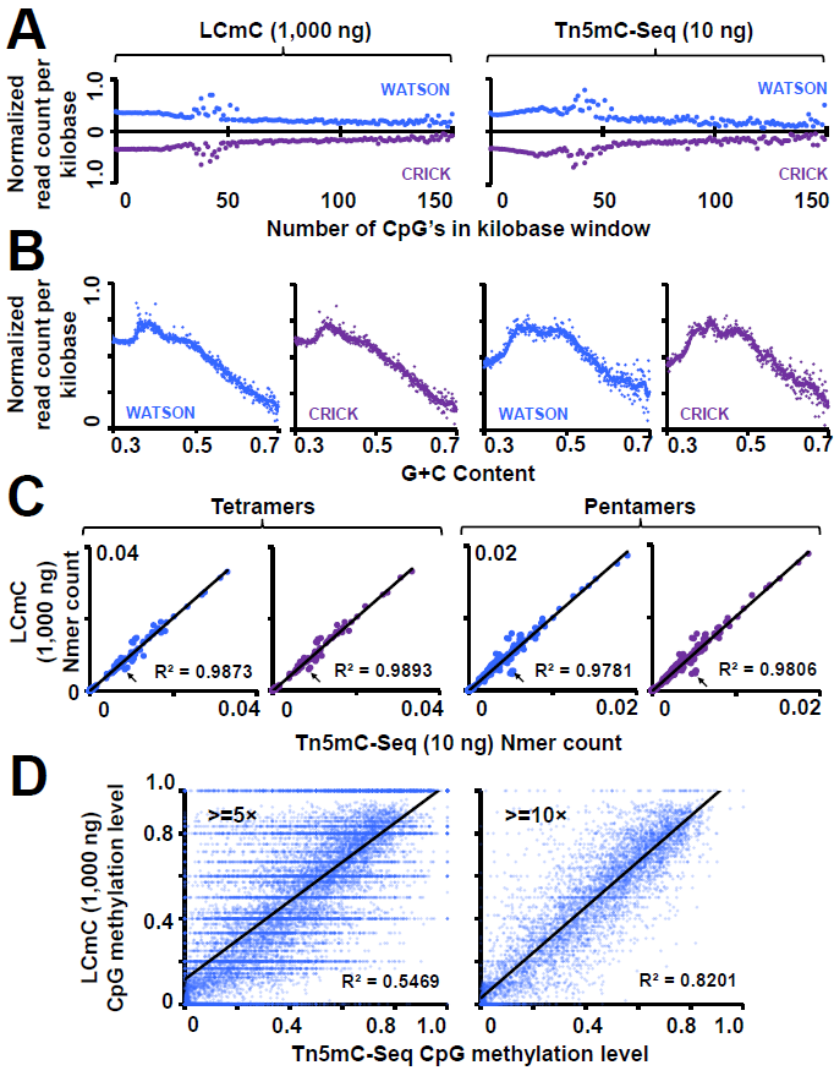


Figure B.2.5. Tn5mC-Seq comparison to ligation chemistry based WGBS library preparation (LCmC).

LCmC libraries were constructed as in Lister et. al. (2009) using 1,000 ng of starting genomic DNA (see methods) and sequenced on an Illumina MiSeq. After alignment the same number of aligned reads were randomly chosen from a Tn5mC-Seq 1.1 (Supplemental Figure B.2.1D) library constructed from 10 ng and used for comparison. (A) Coverage of each library by CpG density is represented as in Supplemental Figure B.2.2. (B) Coverage with respect to G+C content is represented as in Supplemental Figure B.2.2. Solid lines represent the moving average across 10 windows. As both PCRs were performed using Kapa 2G Robust Hot Start Ready Mix, the G+C bias profiles are very similar. (C) Tetramer (left) and pentamer (right) normalized sub-sequence counts of aligned reads for LCmC (Y-axis) and Tn5mC-Seq 1.1 (X-axis). Distribution of sequence across these local contexts are highly similar between the two methods with R2 values above 0.97. The outliers indicated by the black arrow are polyT with a single non-T for the Watson strand, and polyA with a single non-A for the Crick strand which are at slightly higher representations in the Tn5mC-Seq library. (D) Two dimensional scatterplot of methylation levels for CpG positions with a combined Watson and Crick strand depth of >=5× (left) and >=10× (right) in both LCmC (Y-axis) and Tn5mC-Seq (X-axis). The overall methylation levels are highly similar. Additional coverage for the LCmC would likely increase the R2 values based on the improved correlation observed when increasing the minimum coverage from >=5× to>=10× fold coverage.

Appendix C: Supplementary material for Chapter 5

C.1 Supplementary methods for Chapter 5

Library Construction: Genomic DNA was prepared using the QIAGEN Gentra Puregene Kit on 10 mg fresh liver (*mus musculus*, BL6) and 10 flies (*Drosophila melanogaster*, Canton S.) and analyzed on a pulsed field gel (Supplementary Fig. C.4.7). TC-Seq libraries for human were those used in Amini *et. al.* (2013, in review). For mouse and fly, TC-Seq libraries were prepared and sequenced using custom chemistry as described in Amini *et. al.* (2013, in review) with a slight modification to PCR template amounts for the fly sample to account for substantially decreased genome size (1, 2, or 3 pg inputs were used in each PCR reaction as opposed to 10 pg for Human and Mouse, Supplementary Fig. C.4.8). Fosmid libraries were previously generated for Adey *et. al.* (2013)[148]. Long fragment read (LFR) libraries were generated on CEU gDNA using methods outlined in Kaper *et. al.* (2013)[98].

Input Assemblies: The human, mouse, and fly input assemblies were the same input assemblies generated using *ALLPATHS-LG* as described in Burton *et. al.* (2013)[94]. For an additional human assembly of lower quality, we split the original assembly into individual contigs by splitting scaffolds at any N base (N50 437 → 47 kbp). Simulated input assemblies were generated by splitting the human reference sequence at regular intervals.

Contig Scaffolding: Reads were aligned using BWA[149] to references created from the input assemblies. The ends of contigs (1 kbp fosmid and LFR, 5-10 kbp TC-Seq) were then used as vertices and pools that had reads aligning to the vertices were called as hits. A histogram was then generated of the number of pools hitting each vertex and the highest and lowest 5% of vertices were removed (Supplementary Fig. C.4.9). Each vertex was then analyzed individually by calculating the fraction of pools that overlap with all other vertices (shared_pools / (total_vertex1_pools + total_vertex2_pools – shared_pools)) and subsequently calculating the mean and standard deviation for the vertex. *P*-value cutoff thresholds were then determined by finding the –log10(p-value) score that would result in a mean of 1 (conservative), 1.25 (standard), 1.5 (less-conservative), or 3.25 (lenient / short input contigs) outliers per vertex. Vertices (vertex 1) that have an outlier vertex (vertex 2) where vertex 1 is also an outlier in vertex 2 were then considered as potential reciprocated links with an edge weight of the combined $-\log_{10}$(p-value) score of each outlier, and stored as an included edge if that weight was at least 2 (for standard and conservative modes) or 2.5 (lenient modes) times the $-\log_{10}$(*p*-value) cutoff score previously set. Vertices that were at the ends of the same input contig were automatically given an edge score beyond the maximum score cutoff to prevent input contig splitting. Connected components in the graph were then determined followed by identification of the Minimum Spanning Tree (MST) by implementing Prim's algorithm. The longest path (trunk) was then determined by identifying every path between degree one vertices and taking the maximum length path. Non-trunk vertices (branches) were then placed by inserting the branch into the trunk at a position that allows for the maximum trunk path edge weight.

fragScaff Parameters: *fragScaff* was run with varying parameters based on input assembly size and the desired output contiguity. The primary variable parameters are (i) end-node size, (ii) mean passing links per vertex, and (iii) link reciprocation factor (Supplementary Table C.3.1d); however a number of other options can be modified, including the use of a repeat-masking bed file to exclude reads in identified repeats. For the end-node size, the primary determining factor was the N90 of the input scaffolds. The end-node size was set to a minimum of 5 kbp with an optimal size at approximately ½ of the input N90, up to a maximum of 10 kbp. A small N50 generally requires an increase to the mean passing links per vertex. The default is 1.25, though that was increased to 1.4 (mouse), 3.25 (fly), and 4.0 (human, N50 = 47 kbp). The reciprocation factor is the multiplier to the link score cutoff that the sum of the link and reciprocal link must meet to establish it as an edge. The default is 2, but 2.5 was used for smaller N50 inputs where the mean passing links per vertex was elevated in order to recover stringency (2.5 for human N50 = 47 kbp, and fly). For the successive decrease in stringency on the human (input N50 = 437 kbp), the mean passing links and reciprocation factors were set to 1 & 2, 1.25 & 2, 1.5 & 2.5, and 2 & 2.5 in order of decreasing stringency.

fragScaff Accuracy Measurement: Join accuracy was assessed only at locations where consecutive joined input scaffolds are properly mapped back to the reference assembly and were considered an accurate join if the scaffolds were within 5 of one another in the ordered rank. Orientation accuracy was also only assessed for consecutive, reference aligned scaffolds and considered accurate if the correct ends of the input scaffolds were joined. The fraction of base pairs properly placed was assessed by totaling the amount of correctly joined sequence in all scaffolds and then totaling the amount of sequence in all scaffolds that did not belong to the dominant locus of that scaffold. For example, if 3.5 Mb of a scaffold was properly joined in correct order, and was fused with another set of properly joined scaffolds that was 2 Mb in length, we would consider the 2 Mb set of sequence as improperly placed bases. Scaffolds were considered fused if more than 10% of the sequence the scaffold was considered improperly placed. A graphical representation of accuracy measurements can be found in Supplementary Fig. C.4.2.

Contig Anchoring: TC-Seq reads were aligned to the human reference assembly using *BWA*[149] (to GRCh36, as GRCh37 contains a number of the sequences we were anchoring). Fragments were then called in each pool using an alignment-distance cutoff of 15 kbp (Amini 2013, in review) on reads with a mapping quality ≥10. Unaligned reads were then aligned using BWA[149] to a reference of all of the unanchored contigs. Pools with at least one read aligned with a mapping quality ≥10 were then considered hits. Pools with fragments spanning windows of 1-5 kbp in the reference genome were then identified (Supplementary Fig. C.4.5). For each contig, the fraction of window pools shared with the contig pools and the fraction of contig pools shared with the window were calculated for each window, sorted and ranked. The window with the top combined rank was then assigned as the anchor position. Agreement with published loci was determined by checking if our anchoring position was within 1 Mb of the published position.

Misassembly Detection: TC-Seq reads were aligned to the resulting *fragScaff* assembly using *BWA*[149] to assign a pool set for every 5 kbp window (excluding N's) in the assembly. The shared pool fraction for immediately adjacent windows as well as for windows one apart was calculated. Window junctions that had shared pool fractions in the bottom 5th percentile for both immediately adjacent windows and windows one apart were flagged as suspicious.

C.2 Supplementary note for Chapter 5

fragPhase algorithm description:

1. Graph Definition:

   (a) The graph $G = (V, E, F, w)$ contains a set of vertices V defined as all heterozygous sites, a set of edges E where $e \in E$ which are direct links between any two vertices by fragments $K \in F$. Each fragment K has its own distinct graph $K = (V_K, E_K)$ where the vertices V_K are present in V, but V_K and E_K are distinct to each individual fragment where $k \in V_K$ and $l \in E_K$. Each vertex $k_i \in V_K$ has an allele call: $L_{k_i} \leftarrow 0$ if the call is the reference allele, or $L_{k_i} \leftarrow 1$ if the call is the alternate allele. The weight w of each edge $e \in E$ is defined as:

$$w(e) \leftarrow \{l: l \in\}$$

   Isolated vertices are removed from the graph and the haplotype H_i is set to "−" (undefined).

2. Block Determination:

   (a) Each block $I = (V_I, E_I, F_I) \subseteq G$ is a connected component subgraph of G. Each I is initiated at a vertex i and all other vertices j are added to V_I if a path exists between j and i. E_I is defined as all edges touching V_I. Each F_I are all fragment sets of edges in I where each fragment $K_I \in F_I$.

3. Block Phasing: for each block $I = (V_I, E_I, F_I) \subseteq G$:

   (a) Agree $A(i,j)$ and Disagree $D(i,j)$ scores are calculated for every edge $e = \langle i,j \rangle \in F_I$ where:

$$A(i,j) \leftarrow \left\{l: l = \langle k_i, k_j \rangle \in F_I \wedge L_{k_i} = L_{k_j}\right\}$$
$$D(i,j) \leftarrow \left\{l: l = \langle k_i, k_j \rangle \in F_I \wedge L_{k_i} \neq L_{k_j}\right\}$$

   (b) Each block I is initiated at a vertex $i \in V_I \wedge i \notin T_I \wedge c(i) = \max\big(c(V_I)\big)$ where $T_I \subseteq V_I$ that contains all vertices that have been used to initiate a previous iteration and $c(i)$ is the connectivity of i:

$$c(i) \leftarrow \left\{l: l = \langle k_i, k_j \rangle \in F_i\right\}$$

   (c) The haplotype H_i is set to 0, where 0 designates the reference allele. For all other vertices $j \in V_I$, $H_j = -$ (undefined).

   (d) Each vertex $j \in V_I \wedge H_j = -$ is then assigned a haplotype in genomic order of variants in the block $\{0 \ldots N\}$, where the initiating vertex position is n_i, assignment is carried out for vertices with order numbers $\big\{(n_i + 1) \ldots N, (n_i - 1) \ldots 0\big\}$. Haplotype is assigned to j by interrogating all connected vertices j' in subgraph $I_j = \big(V_j, E_j, F_j\big)$ where:

$$I_j \subseteq I \wedge \exists\, e = \langle j, j' \rangle \in E_I$$

   which includes all vertices, edges, and fragments directly associated with vertex j:

$$H_j = \begin{cases} 0, & \sum_{j' \in V_j} S_a(j,j') \geq \sum_{j' \in V_j} S_b(j,j') \\ 1, & \sum_{j' \in V_j} S_a(j,j') < \sum_{j' \in V_j} S_b(j,j') \end{cases}$$

Where:

$$S_a(j,j') = \begin{cases} A(j,j'), & H_{j'} = 0 \\ D(j,j'), & H_{j'} = 1 \\ 0, & H_{j'} = - \end{cases} \quad \text{and} \quad S_b(j,j') = \begin{cases} A(j,j'), & H_{j'} = 1 \\ D(j,j'), & H_{j'} = 0 \\ 0, & H_{j'} = - \end{cases}$$

Since positions beyond those previously interrogated are connected to the current vertex, the haplotype $H_{j'} = -$ and therefore the score is 0.

(e) After initialization a first phase of greedy improvement is performed where each vertex $i \in V_I$ has its fragment disagreement score $S(I_i)$ calculated for subgraph $I_i = (V_i, E_i, F_i)$ where:

$$I_i \subseteq I \wedge \exists\, e = \langle i,j \rangle \in E_I$$

which includes all vertices, edges, and fragments directly associated with vertex i:

$$S(I_i) = \sum_{K \in F_i} s(K) \text{ and } s(K) = \begin{cases} \frac{P_2}{(P_1+P_2)}, & P_1 \geq P_2 \\ \frac{P_1}{(P_1+P_2)}, & P_1 < P_2 \end{cases}$$

$$P_1 \leftarrow \{k : k \in V_K \wedge L_k = H_k\} \text{ and } P_2 \leftarrow \{k : k \in V_K \wedge L_k \neq H_k\}$$

The haplotype is then reassigned for vertices $j \in V_i$:

$$H'_j = \begin{cases} 0, & (H_j = 0 \wedge j \neq i) \vee (H_j = 1 \wedge j = i) \\ 1, & (H_j = 1 \wedge j \neq i) \vee (H_j = 0 \wedge j = i) \end{cases}$$

Which effectively switches the haplotype of vertex i. The new fragment disagreement score $S'(I_i)$ is then calculated as above for $S(I_i)$, with the following exception:

$$P_1 \leftarrow \{k : k \in V_K \wedge L_k = H'_k\}$$
$$P_2 \leftarrow \{k : k \in V_K \wedge L_k \neq H'_k\}$$

If $S'(I_i) < S(I_i)$ then the switch is made and incorporated into H_I. This process is iterated until no switches are made during an iteration.

(f) The haplotype assignment H_I after the first greedy improvement phase is then compared to previous iterations of initiation events. If H_I or $\overline{H}_I$ has previously been observed where for each vertex $i \in V_I$ the haplotype $\overline{H}_i$ is defined as:

$$\overline{H}_i = \begin{cases} 0, & H_i = 1 \\ 1, & H_i = 0 \end{cases}$$

Then assign the final haplotype of the block as H_I from the previously observed iteration and skip the second phase of improvement.

(g) The next phase of greedy improvement is then carried out on vertices $i \in V_I \wedge n_i > 1 \wedge n_i < (N-1)$ which examines long range switches. The fragment disagreement score $S(I_i)$ as described above is then calculated for subgraph $I_i = (V_i, E_i, F_i)$ where:

$$I_i \subseteq I \wedge \exists \left(K_i \in F_I \wedge \left(\min(n_j) \leq n_i \wedge \max(n_j) \geq (n_i + 1)\right) \in V_I\right)$$

Which includes all vertices and edges associated with all fragments that contain vertices at or before and at or after the vertex i being interrogated. The haplotype is then reassigned for vertices $j \in V_i$:

$$H'_j = \begin{cases} 0, & \left(H_j = 0 \wedge n_j > n_i\right) \vee \left(H_j = 1 \wedge n_j \leq n_i\right) \\ 1, & \left(H_j = 1 \wedge n_j > n_i\right) \vee \left(H_j = 0 \wedge n_j \leq n_i\right) \end{cases}$$

Which effectively flips the haplotype assignment of all vertices to the right of vertex i and the new fragment disagreement score $S'(I_i)$ is calculated. If $S'(I_i) < S(I_i)$ then haplotypes calls H_I are reassigned as for all $j \in V_I$:

$$H_j = \begin{cases} 0, & \left(H_j = 0 \wedge n_j > n_i\right) \vee \left(H_j = 1 \wedge n_j \leq n_i\right) \\ 1, & \left(H_j = 1 \wedge n_j > n_i\right) \vee \left(H_j = 0 \wedge n_j \leq n_i\right) \end{cases}$$

The next part of the second phase of greedy improvement involves the same approach as the first greedy improvement except instead of interrogating a single vertex, sets of sequential vertices from $\{1 \dots M\}$ are iterated through where M is the max number of sequential vertices in a set. For each set size in $\{1 \dots M\}$, each vertex $i \in V_I$ is assigned to a set C, where:

$$C = \left\{ j \in V_I \wedge \left(n_j \geq n_i \wedge n_j \leq (n_i + M)\right) \right\} \in V_I$$

and the fragment disagreement score $S(I_C)$ for subgraph $I_C = (V_C, E_C, F_C)$ where:

$$I_C \subseteq I \wedge \exists\, e = \langle \{C\}, j \rangle$$

Which includes all vertices, edges, and fragments directly associated with any vertex contained in C. The haplotype is then reassigned for vertices $j \in V_C$:

$$H'_j = \begin{cases} 0, & \left(H_j = 0 \wedge j \notin C\right) \vee \left(H_j = 1 \wedge j \in C\right) \\ 1, & \left(H_j = 1 \wedge j \notin C\right) \vee \left(H_j = 0 \wedge j \in C\right) \end{cases}$$

Which effectively flips the haplotype assignment of all vertices contained in C and the new fragment disagreement score $S'(I_C)$ is calculated. If $S'(I_C) < S(I_C)$ then haplotypes calls are flipped for all $j \in C$. Both parts of the second improvement phase are iterated until no block switches or vertex set switches are made in an iteration.

(h) After convergence is found, a similar version of the block switching improvement step is made where for each vertex $i \in V_I \wedge n_i > 1 \wedge n_i < (N-1)$, the fragment disagreement score $S(I_i)$ is then calculated for subgraph $I_i = (V_i, E_i, F_i)$ where:

$$I_i \subseteq I \wedge \exists \left(K_i \in F_I \wedge \left(\min\left(n_j\right) \leq n_i \wedge \max\left(n_j\right) \geq (n_i + 1)\right) \in V_I\right)$$

Followed by haplotype reassignment:

$$H'_j = \begin{cases} 0, & \left(H_j = 0 \wedge n_j > n_i\right) \vee \left(H_j = 1 \wedge n_j \leq n_i\right) \\ 1, & \left(H_j = 1 \wedge n_j > n_i\right) \vee \left(H_j = 0 \wedge n_j \leq n_i\right) \end{cases}$$

And then calculation of the new fragment disagreement score $S'(I_i)$. If $S'(I_i) = S(I_i)$ then the long range switch is equivalent and the block is split into two new blocks where each vertex $j \in V_I$ is assigned to its new block along with all associated edges and fragments by:

$$I_{new} = \begin{cases} I_{left}, & n_j \leq n_i \\ I_{right}, & n_j > n_i \end{cases}$$

(i) For each final block I print out the final haplotype assignment H_I as well as the second haplotype $\overline{H}_I$ where for each vertex $i \in V_I$ the haplotype $\overline{H}_i$ is defined as:

$$\overline{H}_i = \begin{cases} 0, & H_i = 1 \\ 1, & H_i = 0 \end{cases}$$

C.3 *Supplementary tables for Chapter 5*

a

		Input				
Organism	**Name**	**Sequence used**	**N10 (Kb)**	**N50 (Kb)**	**N90 (Kb)**	**Scaffold count**
Human	H, Short	S*	133	47	11	127,088
Human	H, Long, A	S + 3Kb	1,235	437	102	18,921
Human	H, Long, B	S + 3Kb	1,235	437	102	18,921
Human	H, Long, C	S + 3Kb	1,235	437	102	18,921
Human	H, Long, D	S + 3Kb	1,235	437	102	18,921
Human	H, Long, F	S + 3Kb	1,235	437	102	18,921
Human	H, Long, L	S + 3Kb	1,235	437	102	18,921
Human	H, Ref, 15	R, 15Kb	15	15	15	191,312
Human	H, Ref, 25	R, 25Kb	25	25	25	115,162
Human	H, Ref, 50	R, 50Kb	50	50	50	57,586
Human	H, Ref, 75	R, 75Kb	75	75	75	38,394
Human	H, Ref, 100	R, 100Kb	100	100	100	28,817
Human	H, Ref, 150	R, 150Kb	150	150	150	19,212
Human	H, Ref, 200	R, 200Kb	200	200	200	14,412
Human	H, Ref, 300	R, 300Kb	300	300	300	9,612
Mouse	Mouse	S + 3Kb	912	224	44	25,964
Fly	Fly	S	207	68	8	7,109

b

	fragScaff					
Name	**Method**	**N10 (Kb)**	**N50 (Kb)**	**N90 (Kb)**	**N50 Imp.**	**Scaffold count**
H, Short	TC-Seq	4,447	1,615	255	34×	27,619
H, Long, A	TC-Seq	17,011	5,469	1,336	13×	7,145
H, Long, B	TC-Seq	33,075	10,147	2,100	23×	6,245
H, Long, C	TC-Seq	42,377	13,918	2,753	32×	5,832
H, Long, D	TC-Seq	75,564	27,440	4,914	63×	5,357
H, Long, F	Fosmid	1,633	567	133	1.3×	15,303
H, Long, L	LFR	2,355	668	151	1.5×	14,476
H, Ref, 15	TC-Seq	1,069	361	15	24×	36,790
H, Ref, 25	TC-Seq	7,281	1,751	100	70×	18,127
H, Ref, 50	TC-Seq	14,161	3,753	250	75×	8,464
H, Ref, 75	TC-Seq	18,915	4,754	300	63×	5,660
H, Ref, 100	TC-Seq	24,605	6,601	400	66×	4,223
H, Ref, 150	TC-Seq	31,813	7,503	600	50×	2,838
H, Ref, 200	TC-Seq	34,366	10,202	800	51×	2,134
H, Ref, 300	TC-Seq	38,708	12,302	1,200	41×	1,451
Mouse	TC-Seq	3,926	1,414	332	6.3×	6,196
Fly	TC-Seq	1,190	459	22	6.8×	4,024

c

Name	fragScaff					
	Included Sequence	Join accuracy	Orientation accuracy	Bp properly placed	Longest Perfect (Kb)	Fused scaffold (%)
H, Short	97.69	98.46	67.99	99.38	4,869	1.2
H, Long, A	98.20	97.27	90.46	98.69	23,954	3.4
H, Long, B	98.77	96.65	88.81	98.43	36,642	4.84
H, Long, C	99.00	96.01	87.26	98.24	36,670	5.82
H, Long, D	99.16	94.63	85.49	97.52	48,788	10.95
H, Long, F	38.90	71.57	74.66	87.95	3,497	6.06
H, Long, L	46.58	34.00	62.78	60.53	4,570	20.75
H, Ref, 15	*77.95*	99.53	96.02	99.53	2,304	0.04
H, Ref, 25	*88.66*	99.60	99.52	99.60	16,313	0.04
H, Ref, 50	*89.87*	99.31	99.39	99.31	31,025	0.08
H, Ref, 75	*90.23*	99.39	99.58	99.39	35,954	0.63
H, Ref, 100	*90.47*	99.29	99.58	99.29	30,706	0.47
H, Ref, 150	*90.64*	99.36	99.66	99.36	39,916	0.29
H, Ref, 200	*90.78*	99.43	99.59	99.43	52,410	0.37
H, Ref, 300	*90.79*	99.62	99.72	99.62	54,311	0.71
Mouse	94.88	98.31	84.19	98.75	6,550	2.32
Fly	87.90	99.36	70.84	99.89	469	0

d

Name	Parameters		
	End node size (-E, Kb)	Mean Links Per Vertex (-l)	Reciprocation Factor (-u)
H, Short	6	4	2.5
H, Long, A	10	1	2
H, Long, B	10	1.25	2
H, Long, C	10	1.5	2.5
H, Long, D	10	2	2.5
H, Long, F	1	1.25	2
H, Long, L	5	1.25	2
H, Ref, 15	5	1.25	2
H, Ref, 25	7.5	1.25	2
H, Ref, 50	10	1.25	2
H, Ref, 75	10	1.25	2
H, Ref, 100	10	1.25	2
H, Ref, 150	10	1.25	2
H, Ref, 200	10	1.25	2
H, Ref, 300	10	1.25	2
Mouse	10	1.4	2
Fly	5	3.25	2.5

Table C.3.1. *fragScaff* assembly results

a. Input assemblies prior to *fragScaff* assembly. Sequence used indicated the types of libraries included. S = shotgun, 3 kbp = 3 kbp mate pair sequencing, R = reference genome segmented into following contig sizes. (*indicates shotgun assembly that was created by fragmenting the larger human assembly) **b.** fragScaff assembly output and sequencing method used to perform scaffolding. Scaffold count includes scaffolds that were not joined to any other scaffolds during the *fragScaff* process. **c.** *fragScaff* assembly performance. Numbers in italics for included sequence indicates that a max of 100% is not possible due to the initial 10% node filtering that *fragScaff* implements, which translates to 10% of the assembly when all contigs are identical lengths. Also in simulated contigs a number are comprised solely of N bases in scaffolded regions of the human reference. Longest perfect refers to the longest scaffold that is not fused for which each successive input scaffold is properly placed in order. Additional accuracy description can be found in Figure C.4.1. **d.** Variable *fragScaff* parameters used. All other parameters were program defaults.

a

Organism	Input: Initial Assembly	Input: N50 (Kb)	CPM Only: Percent Clustered	CPM Only: Percent Ordered	CPM Only: Clustering Accuracy	CPM Only: Ordering Accuracy	CPM Only: Orienting Accuracy
Human	S*	47	89.7	41.5	99.4	95.2	89.3
Human	S + 3Kb	437	98.2	94.4	99.9	99.5	98.8
Human	R, 15Kb	15	35.9	0.2	13.5	96.8	97.0
Human	R, 25Kb	25	28.9	0.4	15.0	97.5	97.5
Human	R, 50Kb	50	70.8	16.5	95.6	86.0	74.2
Mouse	S + 3Kb	224	98.0	86.7	99.8	99.5	98.1
Fly**	S	68	81.2	82.0	n/a	n/a	n/a

b

Organism	Input: Initial Assembly	Input: N50 (Kb)	fragScaff + CPM: Percent Clustered	fragScaff + CPM: Percent Ordered	fragScaff + CPM: Clustering Accuracy	fragScaff + CPM: Ordering Accuracy	fragScaff + CPM: Orienting Accuracy
Human	S*	47	99.4	99.1	97.5	95.9	96.3
Human	S + 3Kb	437	98.8	96.0	99.7	98.9	98.5
Human	R, 15Kb	15	91.9	88.2	99.9	96.0	96.4
Human	R, 25Kb	25	92.5	93.1	99.9	96.5	97.6
Human	R, 50Kb	50	93.6	95.5	96.9	97.4	97.2
Mouse	S + 3Kb	224	99.8	98.6	98.0	99.2	98.5
Fly**	S	68	96.2	93.0	n/a	n/a	n/a

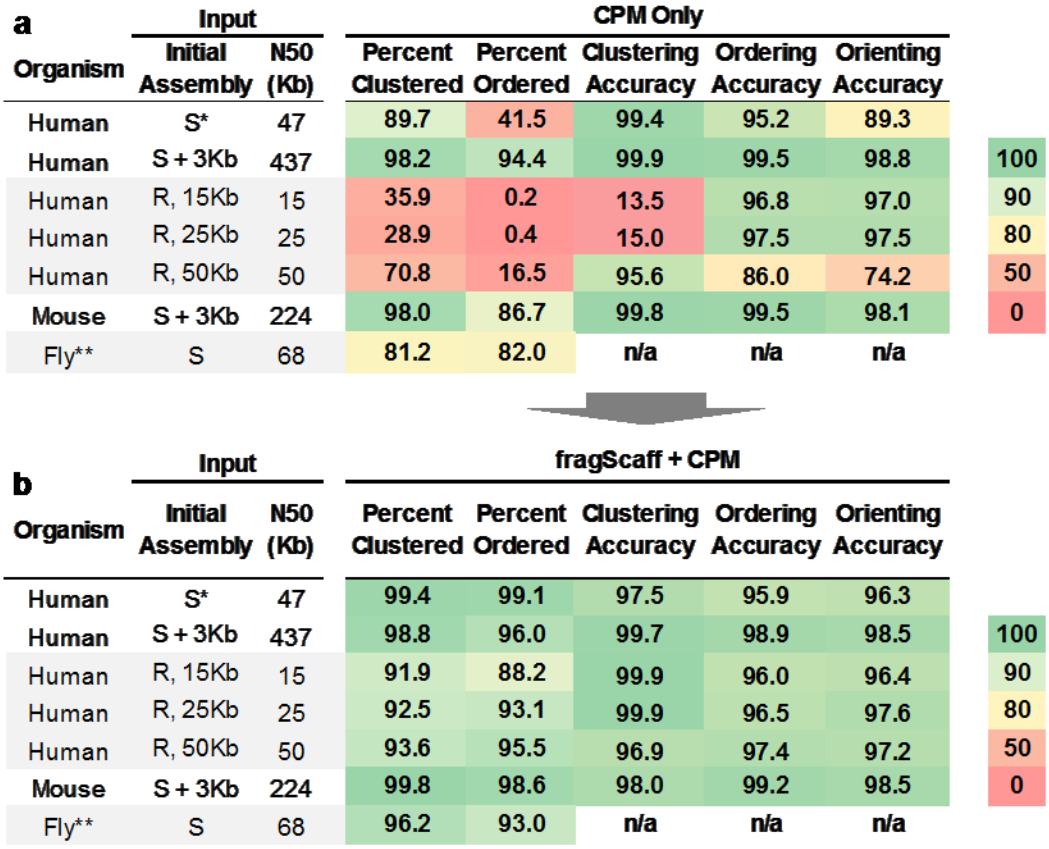


Table C.3.2. CPM improvements when first using *fragScaff*

a. Completeness and accuracy measurements for CPM scaffolding on various input assemblies. **b.** Completeness and accuracy measurements for CPM scaffolding after first scaffolding using TC-Seq and *fragScaff* on the same input assemblies.

* Shotgun assembly generated by fragmenting the shotgun + 3 kbp mate-pair assembly at any N-base.

** Fly accuracy measurements not possible due to the inability to uniquely align a large proportion of input contigs to the fly reference genome.

C.4 Supplementary figures for Chapter 5

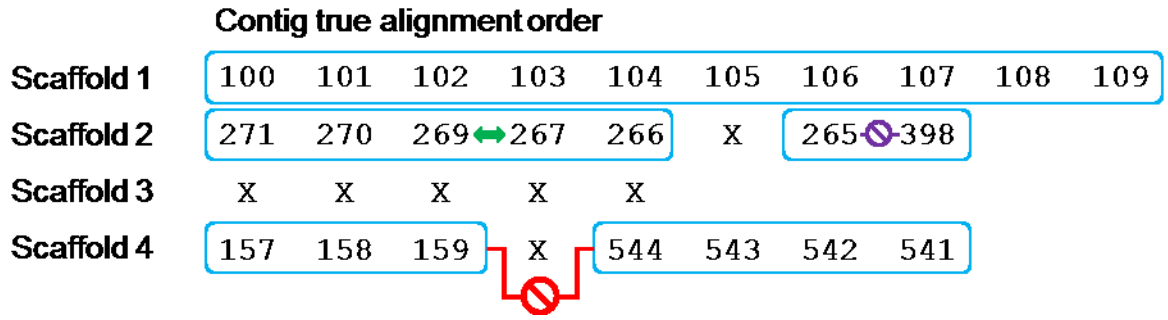


Figure C.4.1. Accuracy determination schematic
The true alignment order of each scaffold is determined via BLAST alignments of the contigs to the reference genome. Alignments are retained if they contain at least 80% of the contig sequence, otherwise the order is not determined and denoted as an X. Join accuracy is only determined for sub-segments of scaffolds in which all successive contigs have a true alignment order (blue boxes) where the ordering is within 5 of the actual ordering (eg. Green double-arrow link is considered accurate, despite a jump of 2 in the ordering as opposed to 1) and the jump between 265-398 is considered an incorrect join (purple crossed link). A fused scaffold that is bridged by an unmapped contig is still included as a fused scaffold in the accuracy statistics (red crossed link). For the fused scaffold all bases belonging to the shorter sub-segment of the scaffold would be considered improperly placed.

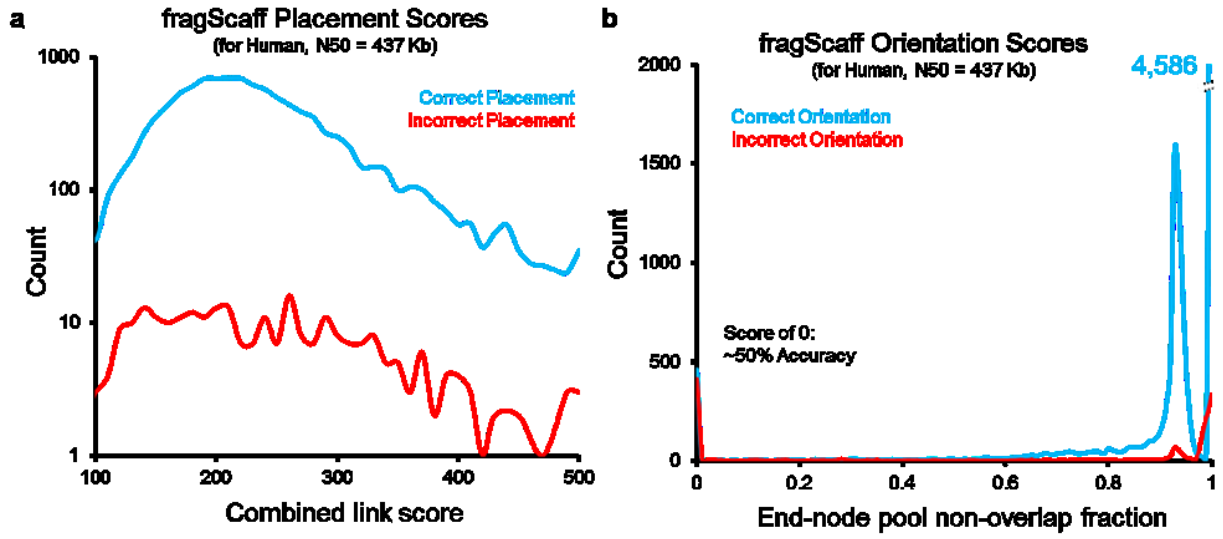


Figure C.4.2. Link placement and contig orientation scores
a. Histogram of link quality scores for properly placed (blue) contigs and improperly placed (red) contigs for the stringent human assembly (input N50 = 437 kbp). Due to extensive filtering and pruning during the fragScaff, there is little discrimination between accurate and inaccurate links based on the score. **b.** Orientation quality score distribution for correctly oriented (blue) and incorrectly oriented (red) contigs. The score is 1 – (pool overlap fraction), resulting in completely overlapping end nodes sharing 100% of pools and thus an orientation score of 0 resulting in a 50/50 chance of proper assignment.

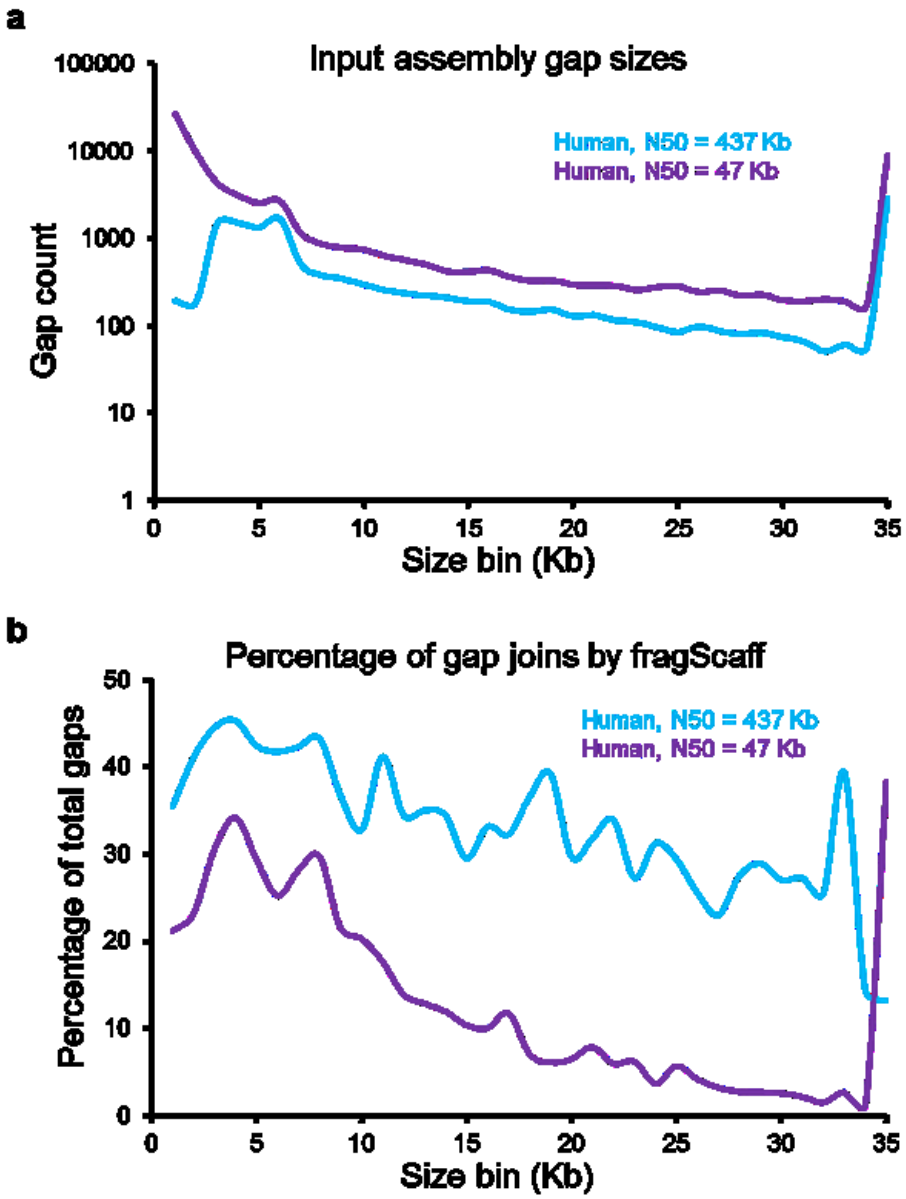


Figure C.4.3. Input assembly gap distribution and joining percentages
a. Histogram of the number of gaps present at increasing sizes up to 35+ kbp for both human input assemblies. **b.** Percentage of joins made at each input assembly gap size for the stringent human assembly and the small input N50 human assembly.

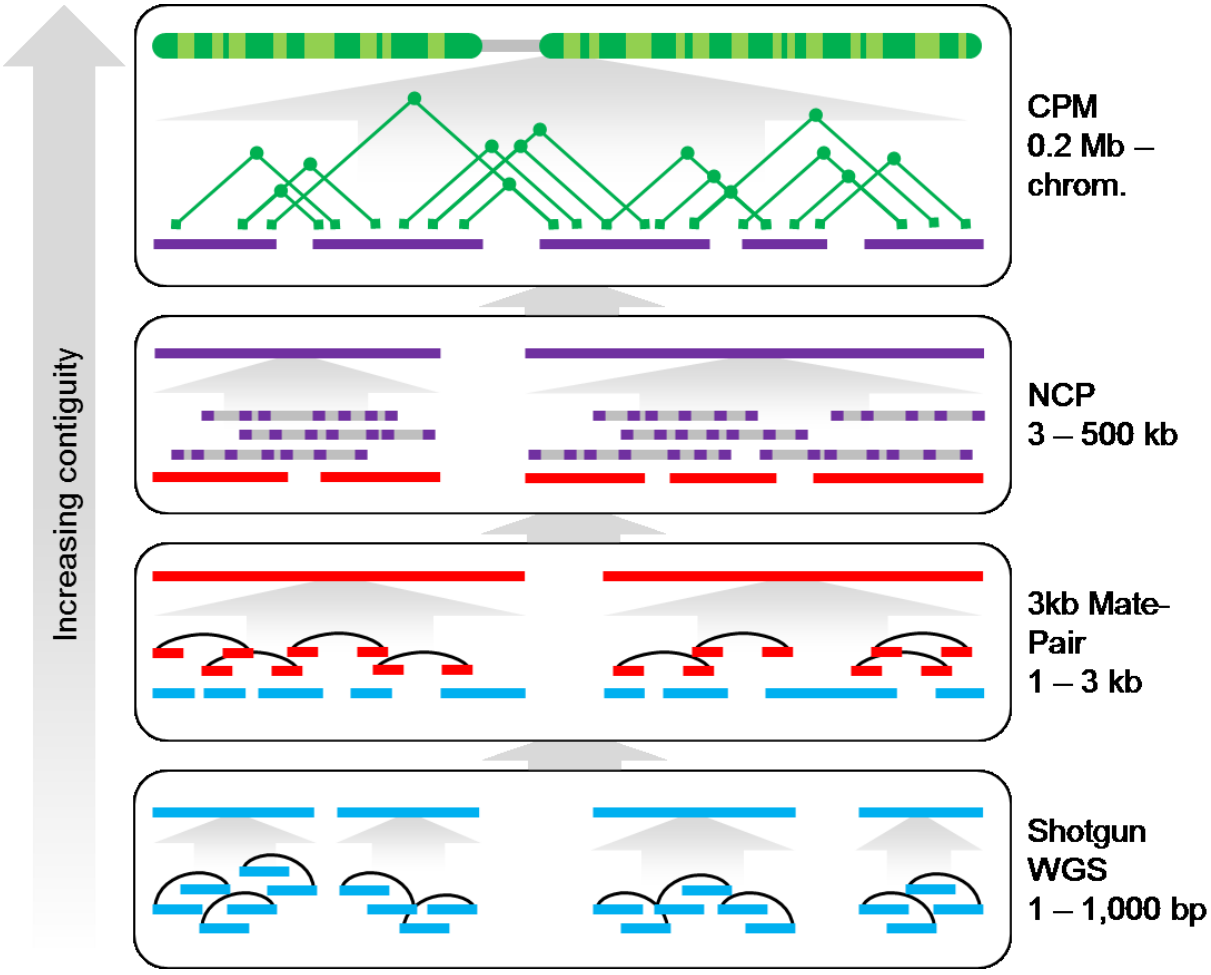


Figure C.4.4. *in vitro, de novo* assembly of genomes
First, shotgun paired-end reads are generated to high depth (blue pairs) and assembled into local sequence contigs (long blue lines). Next, 3 kbp mate pair libraries (red pairs) can be used to generate mid-size scaffolds (long red lines). NCP can then be used (gray and purple lines) to increase scaffold sizes large enough so that the last phase (CPM) can be used. Finally CPM will cluster scaffolds into chromosomes and assemble into a complete reference.

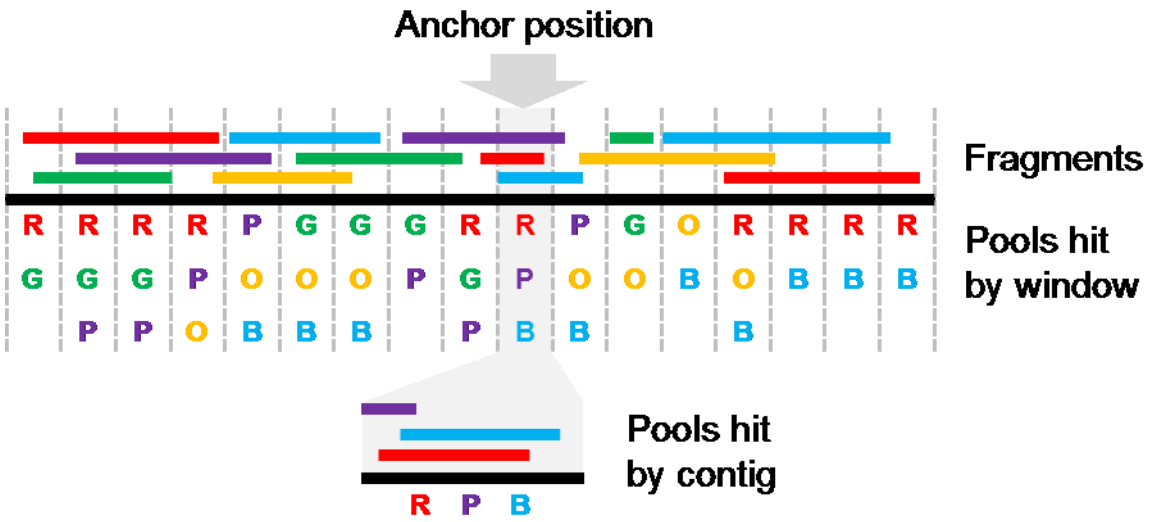


Figure C.4.5. Contig anchoring scheme
To anchor novel contigs to a reference genome (black line), we first aligned all NCP reads to the genome and called fragments (red, green, purple, blue, and orange lines). Next, the pools that contain fragments present in windows (gray dashed lines) tiling across the genome are identified (R, G, P, B, and O). Unaligned reads are then aligned to the contigs to identify which pools hit each contig. The window in the reference genome that has the largest fraction of shared pools (gray shading) is then identified as the most likely anchoring position.

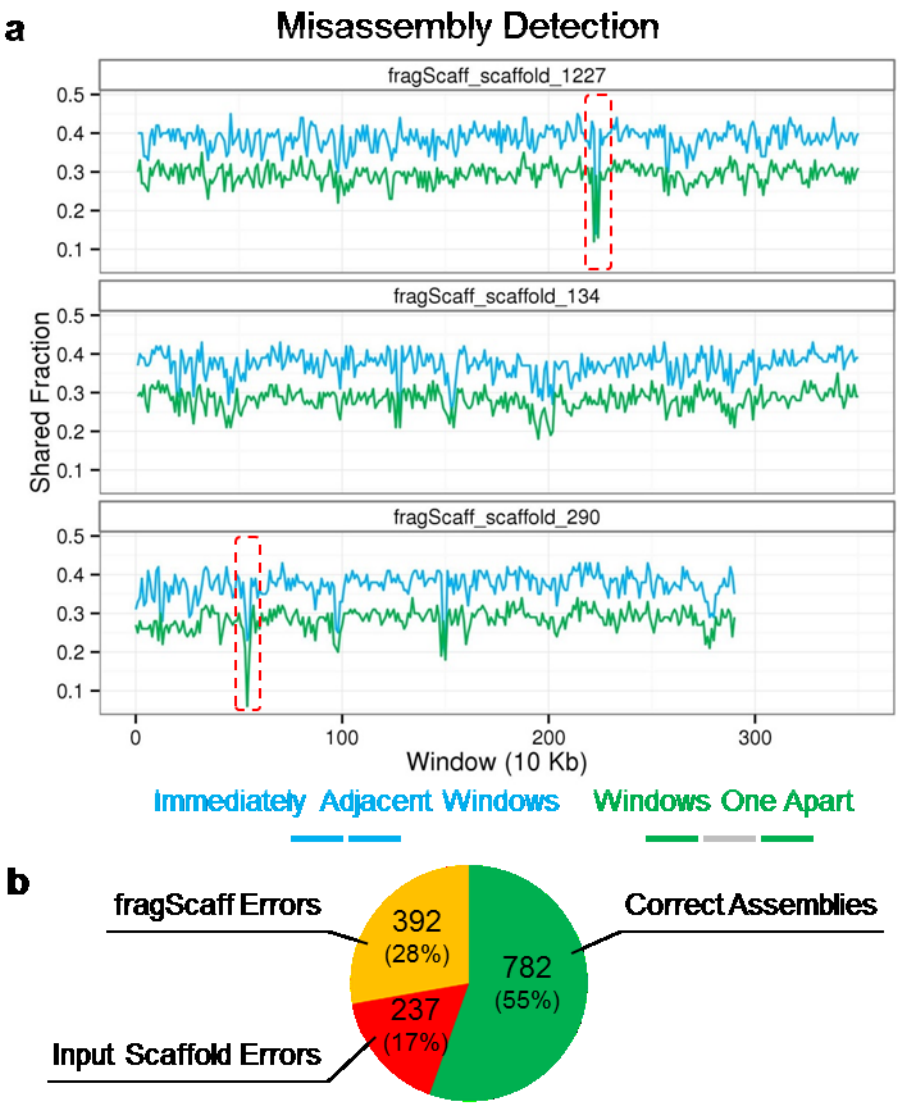


Figure C.4.6. Misassembly detection with TC-Seq
a. Misassemblies are detected using a sliding window approach and calculating the fraction of coinciding pools between immediately adjacent windows (blue) and windows one apart (green). Dips in both scores are called as misassembly candidates. **b.** Breakdown of candidate misassemblies. Just under 50% are accurately identified.

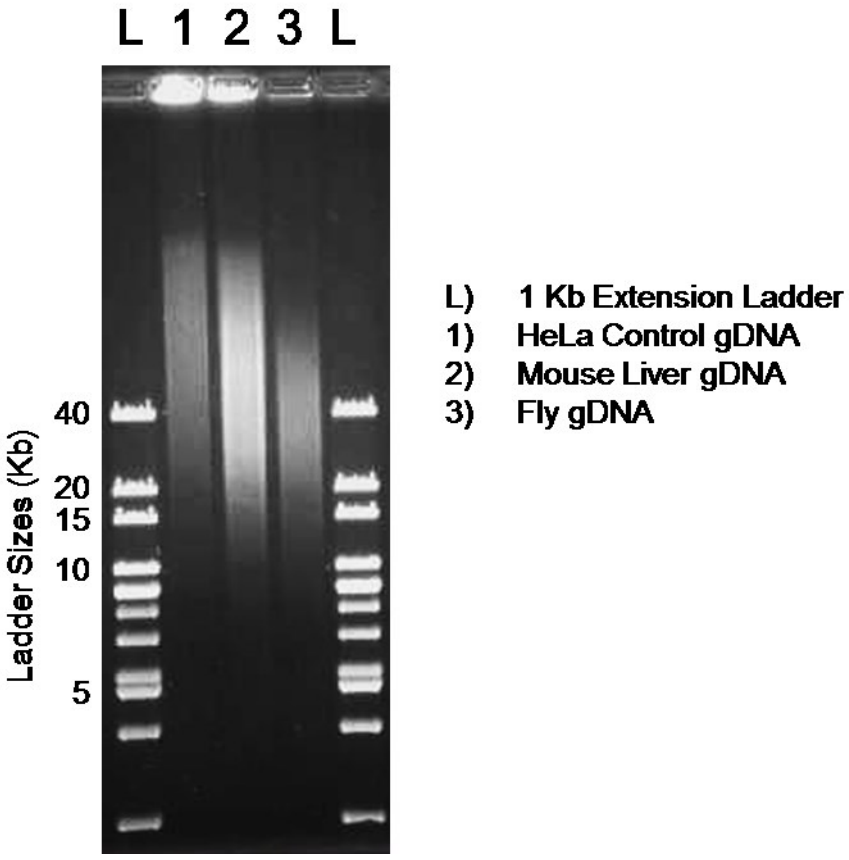


Figure C.4.7. Pulsed-field gel electrophoresis of TC-Seq input gDNA samples
Pulsed field gel electrophoresis was run on a BioRad Pulsed Field Gel Electrophoresis system using a 1% agarose TBE gel run at 170 V for 16 hours at 14°C with a switch time from 1 to 6 seconds.

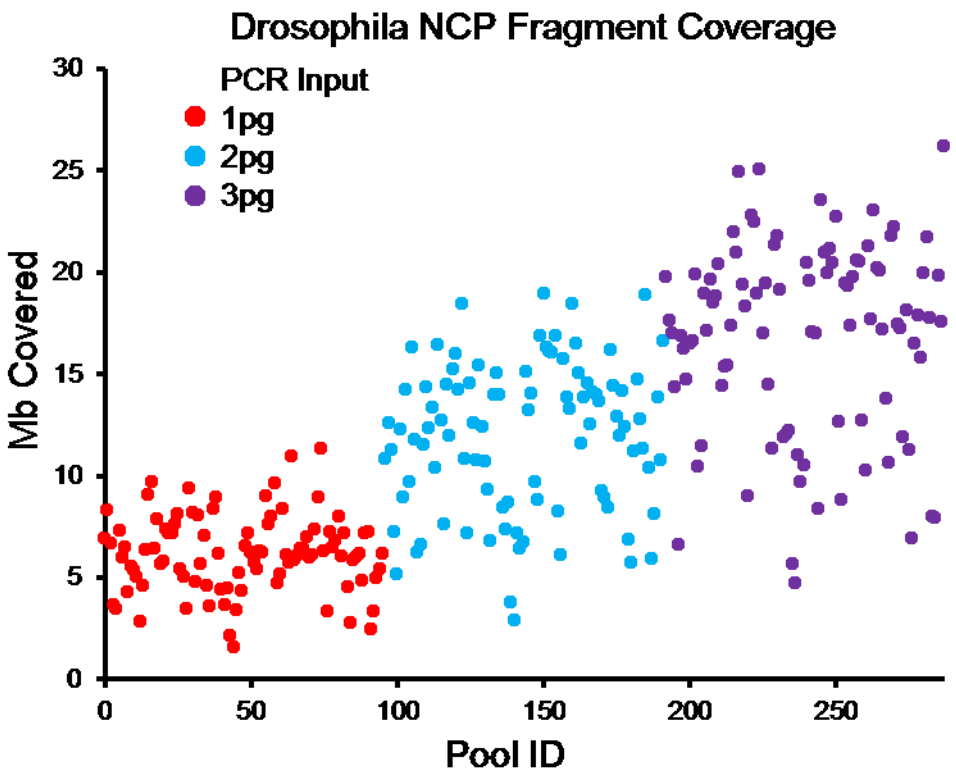


Figure C.4.8. *Drosophila melanogaster* TC-Seq called fragment coverage
96 pools for each input (PCR) mass are shown along with the total megabases of the genome that are covered by called fragments. The amount covered tracks well with the input mass, thus allowing method tailoring to reflect the genome size of the organism being assembled.

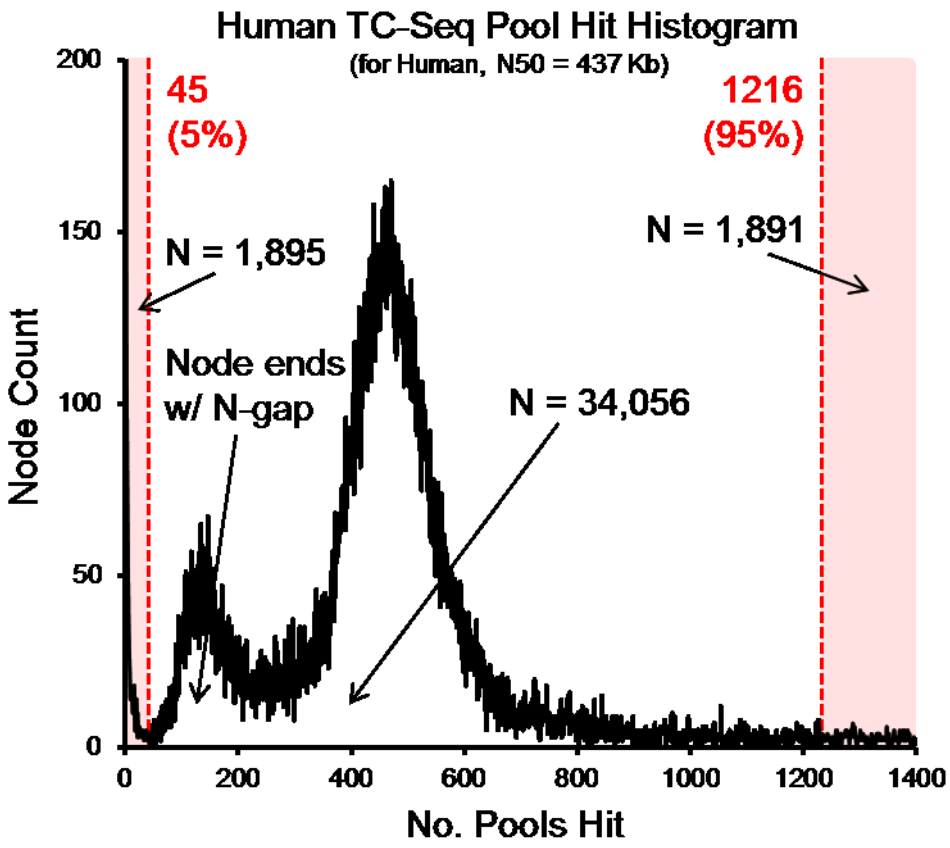


Figure C.4.9. End-node pool hit histogram
Histogram of the pool hit count for contig ends for the human (N50 = 437 kbp) input assembly. The top and bottom 10% of node ends are automatically filtered out. The lower peak is due to node ends that contain a segment of N's in the node end resulting in less alignable space.

Appendix D: Supplementary material for Chapter 6

D.1 Supplementary methods for Chapter 6

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, GRCh37) using BWA[76] followed by duplicate removal, quality score recalibration and local indel realignment using GATK[150]. SNVs were called using samtools[151], indel variants were called using GATK[150] and short tandem repeats (STRs) were called using LobSTR[152] (Supplementary Note D.2.1). Indel detection as a function of coverage was investigated further as described in Supplementary Note D.2.2. Gene ontology term analysis was carried out using DAVID[153]. Data sets used for each analysis are depicted as a flow chart in Supplementary Fig. D.4.48.

Read depth copy number analysis: Shotgun reads for HeLa and Human Genome Diversity Project (HGDP) control genomes[106] along with a similarly prepared control library with a matched G + C profile were aligned using mrsFAST[154], processed as described previously[27] to generate read depth-based copy number predictions within non-overlapping windows of singly unique nucleotide k-mers (SUNK windows; Supplementary Note D.2.3). Copy-number calling in HeLa was carried out at high (approximately 1.5 kbp) and low (approximately 77 kbp) resolution using an HMM (Supplementary Note D.2.4), and a recalibration process was then used to account for widespread aneuploidy (Supplementary Note D.2.5). Short amplifications and deletions were identified using a sliding-window approach (Supplementary Note D.2.6). 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 (Supplementary Note D.2.7). 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 (Supplementary Note D.2.8).

Mate-pair library construction, sequencing and analysis: Library construction for 40 kbp 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[92] (Supplementary Note D.2.9). Libraries of approximately 3-kbp inserts were constructed following protocols described previously[155] (Supplementary Note D.2.9). 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 (Supplementary Note D.2.10).

Fosmid pool construction, sequencing and haplotype phasing: Three replicate fosmid libraries were prepared as described previously[61], and then partitioned by limited dilution into 96 sub-libraries. This was followed by outgrowth, barcoded transposase-based library preparation[59], sequencing and alignment (Supplementary Note D.2.11). Clone boundaries were inferred as described previously[61], 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[100] (Supplementary Note D.2.12). 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, Supplementary Note D.2.12). 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 (Supplementary Note D.2.13). Haplotype scaffolds were analysed for variant population frequencies to investigate the ancestral origin of phased blocks (Supplementary Note D.2.14). Finally, overall copy numbers were compared among all HeLa strains sequenced in this study (Supplementary Note D.2.15).

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 kbp. Single-molecule real-time sequencing libraries for the Pacific Biosciences RS sequencer were prepared using the Pacific Biosciences DNA Template Prep Kit (3–10 kbp), 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 (Supplementary Note D.2.16).

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 (Supplementary Note D.2.17).

ENCODE and RNA-seq phasing: Directional, PolyA+ RNA-seq data generated in-house on HeLa S3 (Supplementary Note D.2.18) were analysed 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 (ref. [105]) (Supplementary Note D.2.19). RNA-seq reads were aligned using TopHat[156] and transcript quantification was carried out using Cufflinks[157]. 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 (Supplementary Note D.2.20). Reference bias was investigated in all tracks and removed in a subset to identify its impact on outlier calling (Supplementary Note D.2.21). Haplotype-specific peaks were then identified in all data tracks (Supplementary Note D.2.22). Finally, a meta-analysis of all data tracks was used to identify large regions of haplotype imbalance (Supplementary Note D.2.23).

D.2 Supplementary notes for Chapter 6

D.2.1 Shotgun sequencing and variant calling

Aim: To construct shotgun sequencing libraries for all HeLa strains for use in variant calling, and copy number analysis.

Input: Cells from HeLa: ATCC, CCL-2 (lab stock); HeLa S3 (lab stock): ATCC, CCL-2.2 (lab 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 and AV-3: ATCC, CCL-21.

All shotgun libraries were constructed using standard ligation chemistry methods from 1 µg of isolated genomic DNA (QIAGen Gentra PureGene kit). HeLa CCL-2 and HeLa S3 libraries were generated in duplicate at two different size ranges (2x 150-250 bp, and 2x 250-500 bp). All other HeLa isolate shotgun libraries were generated at a single size range (100-250 bp). All libraries were sequenced on an Illumina HiSeq 2000 with 13 lanes of paired-end 101 bp (PE101) reads for HeLa CCL-2 (six lanes required trimming read 2 to 40 bp due to an instrument solenoid valve failure on cycle 41 of read 2), 2 lanes of PE101 plus 2 lanes of PE51 for HeLa S3, and ½ lane of PE51 for all other HeLa isolates. For variant calling, reads were aligned to the human reference genome (hg19, GRCh37) using BWA (v0.5.9)[76] with default parameters. Alignments of each library were merged and filtered to remove PCR duplicates followed by quality score recalibration and indel realignment using the Genome Analysis ToolKit (GATK v1.6)[150]. SNVs were called using Samtools (v0.0.18)[151] due to its increased sensitivity for variants at low allele balances, and indels were called using GATK with a filter requiring a non-reference frequency of 0.1, coverage of at least 5, and quality score of at least 500.

D.2.2 Indel calling with respect to coverage

Aim: To investigate the effects of copy number and sequencing depth on the ability to call indels in order to assess how comparable the number of indels in HeLa is to controls.

Data Sources: Indel calls in HeLa CCL-2 (~88X coverage), HeLa CCL-2 (downsampled to ~35X coverage), HeLa S3 (~26X coverage), and HGDP controls[106]; ~30-45X coverage), and bam files used to generate them.

Indel calling in HeLa resulted in markedly higher numbers than that of the HGDP controls (HeLa ≈ 4.2 x 10^5, HGDP average ≈ 3.3 x 10^5). We initially thought this increased number of calls was simply due to increased coverage, so we downsampled the HeLa alignment to a comparable fold coverage (with respect to GRCh37) as the HGDP individuals (~35X) and re-called indels. This resulted in a markedly lower number of final calls (n ≈ 2.1 x10^5), which was surprising until the aneuploidy of HeLa was taken into consideration. To investigate this further, we tallied indel and read depth in windows across the genome and plotted called indel counts with respect to sequencing depth, revealing comparable trends for HeLa and controls (Supplementary Fig. D.4.1).

D.2.3 "SUNK" read depth determination

Aim: To establish coverage scores for windows in the genome of uniquely mappable positions termed "SUNK" windows (uingly unique nucleotide k-mer) as in Sudmant *et. al.* (2010)[27].

Data Sources: Sequence reads in fastq format (Samples and controls)

Reads were aligned using the mrsFAST[154] read aligner which reports all possible alignments to a repeat-masked reference genome. Alignments were then processed by retaining unique alignments to non-overlapping windows each containing 50 uniquely mappable positions, or SUNK windows (singly unique nucleotide k-mers) as previously described in Sudmant *et. al.* (2010) [27]. These unique read counts per SUNK window were then used for read depth copy number analysis. (high-resolution; ~1.5 kbp; 50

uniquely mappable positions) as well as for merged windows of 50 SUNK windows (low-resolution; ~77 kbp; 2,500 uniquely mappable positions).

D.2.4 Copy number state HMM

Aim: To establish large-scale copy number states. Absolute copy number identification as well as identification of outlying copy number regions to be determined in later steps.

Data sources: SUNK read depth (Samples and controls)

SUNK window scores for HeLa CCL-2, HeLa S3, the 11 HGDP controls[106] and a GC-matched control library were first normalized to a control constant to account for total read count differences. Scores for all samples were compared to one another to first identify windows of no coverage that may be population-specific deletions. These windows in HeLa were set aside from HMM copy number calling. Window ratio scores for HeLa strains over the GC-matched diploid normal control were then generated and served as the input observations to a basic HMM. These ratios were plotted as a histogram to identify approximate copy number ratio scores and increments between copy number states (Supplementary Fig. D.4.5). Initiating ratios for copy number states were as follows: State ID = Ratio; 1 = 0.33, 2 = 0.66, 3 = 1.00, 4 = 1.33, 5 = 1.66, 6 = 2.00, 7 = 2.33, 8 = 2.66, 9 = 3.00, 10 = 3.33. These scores fit the histogram and also fit a triploid numerator, as might be expected given that the majority of the HeLa genome is at copy number three based on previous karyotypes in Macville *et. al.* (1999))[104]. A maximum copy number of 10 was used, as nearly all of the HeLa genome falls below that copy number with the exception of small outliers that are determined via later steps. Emission scores for the HMM were determined by calculating the Gaussian probability for each of the state means using an initiating standard deviation of 0.1 for HeLa CCL-2 and 0.25 for HeLa S3:

$$E(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where x = the ratio score for the window, μ = the mean ratio of each state, and σ = the standard deviation. Transition probabilities were initiated to 0.99999 within-state and 0.000001/N out of state (where N = the total number of other possible states). The low out-of-state transition probability was set so as to prevent overfitting of regions of fluctuating GC content. Additionally, each chromosome arm was segmented individually. One Viterbi iteration was deemed enough for convergence (as the initiating assumption of a triploid baseline, which was used in initializing the means, is extremely close to the aggregate ploidy of HeLa).

D.2.5 Copy number recalibration and integer assignment

Aim: To take into account aneuploidy to recalibrate copy number ration and assign absolute copy number states.

Data sources: SUNK ratio scores segmented into states using an HMM.

To account for the effects of different sequencing depths, SUNK window scores were normalized to an equal value for each sample. This results in the assumption that the case and control samples have equivalent masses of DNA in each cell, which is not true in the case of widespread aneuploidy such as in HeLa in comparison to control genomes from healthy individuals. To account for this, a recalibration process was implemented that exhaustively assigns integer copy numbers to the states previously segmented by the HMM, without any preconceived assumptions about the relation between SUNK score and absolute copy number. These integers are then summed and divided by 2N (where N is the total number of windows) to represent a diploid denominator resulting in a “Genetic Material Ratio” or GMR. The GMR is then applied as a recalibration constant to the SUNK windows and new state means are determined followed by comparison to the theoretical ratio scores expected under each copy number assignment hypothesis over a diploid control. The best-fitting hypothesis is then used and the states are assigned their respective absolute copy number values. This process is visually represented in Supplementary Fig. D.4.6.

D.2.6 Copy number outlier determination

Aim: To identify outlier sets of windows that are not called during the HMM process due to the HMM's stringent initial transition probabilities.

Data sources: Recalibrated SUNK ratio scores with copy number calls.

The HMM used to segment copy number states is designed to prevent overfitting that might occur due to local GC-content biases. As such, short regions of altered copy number are not segmented out and must be determined post-HMM. The outlier calling process utilized a sliding window approach that required 3 of 5 SUNK window ratio scores to be a minimum of 3 standard deviations away from the state mean in the same direction. Window ratio scores in opposite directions within a set of 5 count against one another to mitigate effects of noise. Consecutive outlier windows were then merged, and a mean score for the outlying span was determined and assigned to an integer copy number based on theoretical ratio scores determined in the previous recalibration process (to a maximum copy number of 100). After outliers were called, standard deviations of each span were determined and if the standard deviation was $> 0.5\mu$ (μ = span mean), the window was split at various positions in which each new segment contained at least 5 SUNK windows. The split windows then had their mean and standard deviations calculated and were reassigned to a copy number. If the split provided a more optimum fit to copy number states over that of the original combined window then the split was retained. The splitting process was iterated until no splits were retained.

D.2.7 Copy number comparison of 8 additional HeLa strains

Aim: To characterize the copy number profile of 8 additional HeLa strains and compare them to CCL-2 and S3.

Data Sources: Fastq files for shotgun sequencing libraries of 8 additional strains (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).

Reads were aligned and processed in two methods: 1) alignment followed by copy number calling as in Supplementary Notes D.2.4 and D.2.5, or D.2.2) alignment with BWA (v0.5.9)[76], followed by genotyping against all variants in HeLa CCL-2. For copy number profiling, integer copy number calls were for all 8 strains (Supplementary Fig. D.4.25). The raw SUNK window scores (for low-resolution, 50 merged SUNK windows, ~77 kbp) were also compared in an all by all basis (Supplementary Fig. D.4.26). These comparisons revealed that while all strains have HeLa copy number characteristics, they all have their own unique copy number profiles. Due to the differences in library preparation, particularly size selection (which we have found to have significant effects on coverage profiles - unpublished data), between HeLa CCL-2 (Agarose size selection, 2 size ranges, 2 replicates each), HeLa S3 (PAGE size selection, 2 size ranges, 2 replicates each), and the additional 8 strains (PAGE size selection, 1 size range, 1 replicate), the direct comparisons between HeLa CCL-2 or HeLa S3 are notably noisier than between the additional 8 strains. In order to determine potential lineage information, large-scale windows were utilized (600 target SUNK windows, mean = 955,176 bp) and clustered using "pvclust" in R (Supplementary Fig. D.4.27). This method utilized bootstrapping (we performed 1000 iterations) to assign confidence to the clustering dendrogram.

All 8 additional strains as well as S3 (DNA-seq and RNA-Seq) were also aligned using BWA (v0.5.9) [76] and genotyped at HeLa CCL-2 sites. Positions with a coverage of at least 8X in both CCL-2 and the other strain being compared and were not within SegDups were analyzed for the presence of the CCL-2 variant (Supplementary Table D.3.15). The majority of strains and S3 RNA-Seq had 90-97% of interrogated positions showing evidence of the CCL-2 variant with the exception being protein-altering variants shared between CCL-2 and S3. To investigate this discrepancy, variants were split into bins of alternate allele coverage in CCL-2 and plotted for fraction of concordance. (Supplementary Fig. D.4.23). In general a lower fraction of HeLa CCL-2 sites have the alternate allele in HeLa S3 for bins of lower HeLa CCL-2 alternate allele coverage. This trend is consistent between all private variants and private protein-altering variants with both approaching 1.0 for alternate allele coverage in HeLa CCL-2 > 70X. Vertical bars

represent a 95% binomial confidence interval. The decrease in concordance for protein-altering variants is likely caused by statistical noise or sequencing artifacts; however the possibility of a true excess of strain specific protein-altering variants cannot be conclusively ruled out.

D.2.8 Loss-Of-Heterozygosity (LOH) region calling

Aim: To determine regions in the HeLa genome that have undergone loss of heterozygosity.

Data sources: SUNK window intervals; HeLa (CCL-2) shotgun SNV calls.

Loss of heterozygosity (LOH) was called using the local density of heterozygous markers by designing a hidden Markov model (HMM) with two states: presence and absence of LOH. At each interval used for copy number estimation (SUNK windows), the model emitted counts of homozygous and heterozygous variants called within that interval. To avoid falsely rejecting LOH due to the presence of spurious variants or somatic mutations, this analysis was restricted to SNVs from the shotgun data with VCF quality scores of at least 50 that overlapped with variants from the 1000 Genomes Project. Indels were excluded because their allele frequencies tended to be affected by reference mapping bias.

In each window, the quantity of homozygous and heterozygous SNVs was tallied, and the likelihood of observing this number of homozygous and heterozygous SNVs, given LOH or no LOH, was calculated (Supplementary Fig. D.4.7). Each SNV was marked as being in or not in a repeat; SNVs were considered to be in repeat regions if they fell in any of the following UCSC Genome Browser tracks: "segdups.bed", "repeat_masked.bed", "interrupted_repeats.bed", "simple_repeats.bed", "microsat_repeats.bed", and regions with wgEncodeCrgMapabilityAlign100mer ≤ 0.05. It was assumed that SNVs in repeat regions would have a higher observed rate of heterozygosity due to mismapped reads. Specifically, in regions with LOH, the heterozygosity rate was assumed to be 0.5% of variants not in repeats and 10% of variants in repeats; while in regions without LOH, the heterozygosity rate was assumed to be 50% of variants not in repeats and 70% of variants in repeat. These round numbers were estimated from surveys of regions on chromosome 2 with clear signatures of LOH/non-LOH. The likelihood of each window's set of observations (the number of homozygous/heterozygous variants in/not in repeats) was calculated as the binomial probability of those observations assuming the region was in LOH or not:

$$P\left(N_{het,r}, N_{hmz,r}, N_{het,nr}, N_{het,nr} \middle| LOH\right) = (0.005)^{N_{het,nr}} \times (0.995)^{N_{hmz,nr}} \times (0.1)^{N_{het,r}} \times (0.9)^{N_{hmz,r}}$$

$$P\left(N_{het,r}, N_{hmz,r}, N_{het,nr}, N_{het,nr} \middle| NO\ LOH\right) = (0.5)^{N_{het,nr}} \times (0.5)^{N_{hmz,nr}} \times (0.7)^{N_{het,r}} \times (0.3)^{N_{hmz,r}}$$

To guard against the risk of specious results in individual windows arising from phenomena such as mapping artifacts and sudden small-scale changes in copy number, the minimum possible value of P, for any SUNK window and LOH state, was set at 10^{-6}. The HMM was initialized with transition probabilities of 10^{-8} between the two states and equal initiation probabilities of the two states.

The HMM was run through a single iteration of Viterbi training, which was observed to be sufficient for convergence. The Viterbi training yielded a best path through the model, which indicated a prediction of the LOH state of each interval. Adjacent windows with LOH were merged, and the state of the most distal window on each chromosome arm was assumed to extend to the telomere. The resulting LOH calls are shown in Supplementary Table D.3.9.

D.2.9 Mate-pair library construction

Aim: To construct mate-pair sequencing libraries of 3 kbp and 40 kbp.

Input: Genomic DNA (3 kbp libraries), Pooled fosmid libraries (40 kbp)

Library construction for 40-kilobase mate pair libraries was carried out similar to previously described methods in Gnerre *et. al.* (2011)[92] starting with fosmid clone DNA pooled within each original fosmid preparation. Nicks introduced during clone DNA isolation were first repaired by incubating 10 µg pooled fosmid clone DNA with 10 units E. coli DNA Ligase I (NEB) in E. Coli DNA Ligase Reaction Buffer at 16ºC

for 60 min in a 50 µL reaction volume followed by denaturation at 65ºC for 20 min and AMPure SPRI-bead cleanup (Agencourt). Introduction of nicks flanking the cloning site was performed by incubation at 37ºC for 60 min with 25 units nicking restriction endonuclease Nb.BbvCI (NEB) in NEBuffer 2 and a 50 µL reaction volume and then denaturation at 80ºC for 20 min followed by placing reactions on ice. Nicks were translated into the insert by addition of 10 units E. coli DNA Polymerase I (NEB) directly to the previous reaction followed by incubation for 45 min on ice and immediate denaturation at 80ºC for 20 min then SPRI cleanup. Translated nicks were converted to double-stranded breaks by addition of T7 Exonuclease (NEB) in NEBuffer 4 and a 50 µL volume then incubation at 25ºC for 2 hours, followed by SPRI cleanup and subsequent incubation with 0.2 µL S1 Nuclease (Fermentas) in S1 Nuclease Buffer at 25ºC for 30 mins in a 30 µL reaction volume followed by addition of 2 µL 0.5M EDTA and heating to 80ºC for 20 min. Fragments were end-repaired (NEB End-Repair Module) and 100 ng was then treated with 5 µL T4 DNA Ligase in a 500 µL reaction volume overnight at 25ºC to promote intramolecular circularization. The circularized mate-pair fragments were amplified and converted to Illumina sequencing libraries by PCR with primers complementary to the vector backbone, followed by gel size selection. Libraries were pooled for sequencing with paired 100 bp reads on an Illumina HiSeq2000. Libraries of ~3 kilobase inserts were constructed following protocols described in Talkowski et. al. (2011)[155].

Reads were trimmed to 50 bp for the 40 kbp libraries (to reduce the amount of reads which extend through the nick translation portion, through the ligation junction, and into the opposite segment) and to 25 bp for the 3 kbp libraries (as the restriction enzyme utilized in the protocol cuts 25 bp away from the center ligation segment, thus only allowing 25 bp of genomic DNA), and then aligned to the human reference genome (hg19, GRCh37) using BWA (v0.5.9)[76], filtered for phred-scale mapping quality >= 10 , and filtered to remove PCR duplicates. For 40 kbp libraries, a subset of read pairs with very short insert sizes (possibly resulting from translation of one but not both nicks) were suppressed, as were clusters of read pairs with nearly equal outer mapping coordinates. Insert size distributions can be found in Supplementary Fig. D.4.8.

D.2.10 Identification of structural rearrangements

Aim: Identification of deletions, inversions, and translocations from mate-pair data.

Data Sources: Aligned sequence reads from 3 kbp and 40 kbp mate-pair libraries.

Each alignment file was processed to sort potential rearrangement-spanning read pairs into one of five classifications: 1) concordant (expected read pair orientation, insert size: 3 kbp: 1,200 bp <= X < 5,000 bp; 40 kbp: 20,000 bp <= X < 60,000 bp), 2) short-pair (expected read pair orientation, insert size: 3 kbp: X <= 1,000 bp; 40 kbp: X <= 1,500 bp), 3) deletion (expected read pair orientation, insert size: 3 kbp: X >= 5,000 bp; 40 kbp: X >= 80,000 bp), 4) inversions (same chromosome, opposite read pair orientations, any insert size), and 5) translocations (different chromosome, any orientation).

Deletion, inversion, and translocation classifications for each library type were then subjected to a sliding window approach of 1 kbp windows by 500 bp and the numbers of reads with start coordinates within each window were tallied. Windows with read counts in the top 5% had the respective read pairs investigated as to whether or not they fell into other top 5% ranking windows and the fraction of reads in the window falling into each of the other top 5% windows. For deletions and inversions a cutoff was set to make a call where at least 80% of the read pairs within the window link to the same alternate window, whereas translocations were set to require a 50% cutoff. Resulting calls were then merged to account for the overlapping nature of the sliding window approach followed by trimming the edges of the windows down to the first read pair identified that spans the link. Example calls can be found in Supplementary Figs. D.4.9-11.

D.2.11 Fosmid construction and sequencing

Aim: To construct and align fosmid dilution pools for purposes of haplotype phasing.

Input: HeLa (CCL-2) genomic DNA.

Three replicate fosmid libraries were prepared using the Epicentre CopyControl Fosmid Library Production Kit as previously described in Kitzman et. al. (2011)[61] except for the use of a vector (GenBank Accession: EU140751.1) that was modified to contain Nb.BbvCI restriction enzyme sites at the ends of the vector to allow nicking to facilitate fosmid jumping library construction. Each fosmid library was then partitioned by limiting dilution into 96 sub-libraries which were then outgrown and converted into barcoded DNA-seq libraries by transposase-mediated tagging and fragmentation[59] and pooled for sequencing on a single lane of PE101 on a HiSeq 2000 for each fosmid set. Reads were aligned using BWA (v0.5.9)[76] with default parameters and filtering for a mapping quality phred score of 10.

D.2.12 Haplotype phasing

Aim: To phase germline variants ascertained by shotgun sequencing onto haplotypes using fosmid clone pool sequencing and allele ratios.

Data sources: HeLa (CCL-2) whole-genome shotgun variants (SNV and indel calls); HeLa (CCL-2) fosmid clone dilution pool shotgun reads.

Deep whole-genome shotgun sequencing was used for discovery of SNVs and indels. Sub-genomic pools of long insert clones were sequenced in order to determine the haplotype phase for inherited (germline) variants. Because each pool sampled only a small fraction of the genome (median = 1697 clones/pool, or ~2.0% of the haploid genome, given a median insert size of 33 kbp), overlaps between clones within a given pool were expected to be rare, and the reads corresponding to a given clone could be assumed to derive entirely from one germline haplotype or another.

Clone inserts were mapped by a sliding window read depth approach, essentially as previously described in Kitzman *et. al.* (2011)[61]. For each pool, reads with mapping quality ≥ 20 were counted within 1 kbp non-overlapping windows across the genome. Windows with low "mappability", defined as those having fewer than 300 SUNKs (30mers unique within the genome), were excluded. A candidate clone location was recorded where, within a run of 20 to 45 consecutive mappable windows, at least 60% of the windows had read depth above the background level (defined as the 95th percentile of windowed read depths for the equivalent number of read mapping positions drawn at random from the genome). To map the boundaries of each clone insert, overlapping candidates were grouped, and the candidate was selected that maximized the score:

$$\begin{aligned} candScore = \; & (\#mappable\ windows\ in\ candidate\ interval\ with\ read\ depth > background) \\ & - 2 \times (\#flanking\ windows\ with\ read\ depth > background\ within\ \pm 5\ kbp) \\ & + 10 \times (\%\ of\ mappable\ windows\ in\ candidate\ with\ read\ depth > background) \end{aligned}$$

Regions with predicted LOH are haplotype-resolved with respect to germline variants by virtue of their hemizygosity, and were excluded from further analysis. Clones within non-LOH regions were intersected with heterozygous single-base and indel variants from the whole-genome shotgun data, limiting to likely germline variants (those found among individuals in the 1000 Genomes Project). Within each pool, variants at which clone-derived reads had discordant genotypes (indicative of overlapping clones or sequencing errors) were excluded, as were variants called in only a single read on the clone.

To merge individual clones into longer haplotype blocks, we used a custom implementation of the ReFHap algorithm[100], which determines consensus haplotypes from the genotypes of overlapping haploid fragments (clones). Briefly, this algorithm creates a graph in which the nodes represent individual clones and the edges connecting them represent overlaps between the clones. Two clones c_a, c_b are considered to overlap if there are one or more variants covered by both clones. The edge representing this overlap is assigned a weight as follows:

$$W(c_a, c_b) = \sum_{\substack{variants \\ called \\ by\ c_a, c_b}} \begin{cases} 1\ if\ calls\ are\ equal \\ -1\ if\ calls\ are\ unequal \end{cases}$$

An estimated minimum cut is then calculated for the graph using iterations of the "GreedyInit" and "GreedyImprovement" steps, as in Duitama *et. al.* (2012)[100]. The minimum cut determines the set of edges with the lowest possible total weight that can be removed in order to divide the graph into two disjoint subgraphs. The subgraphs represent the two germline haplotypes, and each clone was assigned to a haplotype based upon which subgraph it was in, although determination of the major ("A") and minor ("B") haplotypes was not made until the scaffolding process, described below. Within each haplotype-representing subgraph, the clone phases were converted to variant phases. For each variant – including

variants not present in 1000 Genomes – the set of all calls made by all clones at the variant wasconsidered, taking the clones' phases into account. If a majority of calls supported one phase over the other, that phase was taken as the correct phase of the variant; if there was an exact tie, the variant was not phased. For the vast majority of variants, the calls unanimously implied the same phase (Supplementary Table D.3.12).

Next, haplotype blocks were further combined into longer "haplotype scaffolds" in regions of uneven copy number (*i.e.*, where one germline haplotype was present at greater copy than the other Supplementary Fig. D.4.13). The principle is that, if a large genomic region has a consistent copy number such as 2:1 (*i.e.*, haplotype A is duplicated and haplotype B is not) then the variants from each haplotype should have distinct allele frequencies among the whole-genome shotgun reads – in this example, alleles on haplotype A should have shotgun read frequencies centered on 2/3 and alleles on haplotype B should have frequencies centered on 1/3. By convention, A and B are the haplotypes with more and fewer copies, respectively.

An HMM was used to combine all haplotype blocks within each contiguous interval of consistent, uneven copy number. The model contained three states: (1) haplotype A, (2) haplotype B, and (3) gaps between haplotype blocks. Each observation was a single variant that had been phased into haplotype blocks; each gap between adjacent haplotype blocks was also considered an observation. The transition probability between states A and B was initialized to 10^{-8}, and transitions from the gap state into states A and B were equally likely (reflecting the lack of *a priori* knowledge of the relative phase of the haplotype blocks). At each variant, the HMM emitted the observed counts of whole-genome shotgun reads matching the alleles phased on each haplotype. The emission probability for each variant was then calculated as the likelihood of these read counts under a binomial distribution, parameterized by the total number of shotgun reads at each site and the predicted copy number of each haplotype:

$$P(phase\ A|read\ counts, CNs) = \left(\frac{CN_A}{CN_A + CN_B}\right)^{counts_A} \times \left(\frac{CN_B}{CN_A + CN_B}\right)^{counts_B}$$

The HMM was run through a single iteration of Viterbi training which was observed to be sufficient for convergence. The Viterbi training yielded a best path through the model, which indicated a prediction of the relative phase of all variants.

In addition to connecting adjacent haplotype blocks into longer scaffolds, these results assigned the "A" and "B" labels, indicating which haplotype was of higher copy. In addition, this model introduced switches within blocks between 0.19% of adjacent phased sites. These switches reflect likely errors on the part of fosmid-based phasing which were corrected by the signal of allelic imbalance among whole-genome shotgun reads.

D.2.13 Calling Haplotype-Resolved Copy Numbers (HRCNs)

Aim: To create a genome-wide profile of HRCN – that is, the distinct copy numbers of each haplotype.

Data sources: HeLa (CCL-2) genome-wide copy-number and LOH calls; HeLa (CCL-2) haplotype scaffolds of phased variants.

HRCNs were called at all haplotype scaffolds, including unscaffolded blocks. First, the haplotype blocks/scaffolds were split at sites where the total copy number is predicted to change, and for each resulting interval, the most likely HRCN was determined according to the following process:

Haplotype-resolved copy numbers (HRCNs) are written here as [total copies]:[copies haplotype A]:[copies haplotype B]. For instance, regions at normal autosomal copy without LOH would be 2:1:1. Intervals coinciding with LOH regions were assigned an HRCN of CN_{total}:CN_{total}:0. Blocks/scaffolds in triploid regions without LOH are labeled as 3:2:1. Blocks/scaffolds in non-LOH regions of copy number 4 or more required special attention because more than one HRCN was possible; for instance, 4:3:1 and 4:2:2. For these cases, the alternate allele frequencies (AAFs) of all SNVs in the block/scaffold were tallied up, and each SNV's AAF was rounded to the nearest 1/CN_{total}. For N in the range 1…(CN_{total}-1), each AAF rounded to N/CN_{total} was counted as evidence in favor of an CN_{total}:N:CN_{total}-N split (or CN_{total}:CN_{total}-N:N, if N < CN_{total}/2) split. A value of N was called as the "correct" value if at least 10 SNVs, totaling at least 2/3 of the total number of SNVs, support it; otherwise, the evidence was considered inconclusive and no HRCN call was made. The total set of HRCN calls is shown in Supplementary Table D.3.8. To interrogate potential selection on private protein-altering variants in HeLa CCL-2 we investigated the number of protein-altering variants that occur on the haplotype at a greater copy number than the wildtype. This resulted in 50.77% of private protein-altering SNVs and 43.64% of private protein-altering indels occurring on the amplified haplotype suggesting, at least globally, that there is no such correlation.

The breakpoints of HRCN spans were plotted as positions on chromosome ideograms to provide positional information that was used to generate blocks of contiguous haplotype of appropriate phase and copy number. This was then used along with marker chromosome descriptions from Macville *et. al.* (1999)[104] as well as mate-pair structural calls to identify marker chromosomes and large-scale rearrangements likely present within the HeLa CCL-2 strain resulting in Figure 6.1a.

D.2.14 Ancestry-based analysis of HeLa haplotype phasing

Aim: To analyze our scaffolds of phased variants by comparing them to expectations arising from the assumption of mixed European and West African ancestry in the HeLa genome.

Data sources: HeLa (CCL-2) haplotype scaffolds of phased variants; 1000 Genomes Project (CEU & YRI) variants with population frequencies.

The set of 60 CEU individuals and 59 YRI individuals from the 1000 Genomes Project was used as a reference panel. Each haplotype scaffold was partitioned into fixed windows of 1000 SNVs present in the reference panel, resulting in 1,161 windows across the genome. For both haplotypes on each window, a score S_{CEU}^{H} (net similarity of a haplotype to CEU) score was calculated, as the relative log-likelihood of the variants on that haplotype occurring in a CEU individual compared to a YRI individual:

$$S_{CEU}^{H} = log_{10}\left[\prod_{i=1}^{1000} \frac{f_{CEU}(v_i)}{f_{YRI}(v_i)}\right]$$

where $f_{POP}(v_i)$ is the frequency of variant v_i in population POP. For variants appearing in one of the CEU, YRI populations but not the other, the frequency of the variant in the other population was set to a pseudocount value of 1/120 (*i.e.*, the equivalent of one occurrence in one haplotype among that population's reference panel.)

The values of S_{CEU}^{H} across all haplotypes in all windows form a clearly bimodal Gaussian distribution (Supplementary Fig. D.4.18a). There is also a consistent negative correlation between S_{CEU}^{H} and the number of non-1000 Genomes Project variants present on a haplotype, consistent with a commonly observed enrichment of previously unknown variants on African haplotypes (Supplementary Fig. D.4.18b). The haplotype blocks were called as either "CEU-like" or "YRI-like" based on thresholds of $S_{CEU}^{H} \geq 0.1$ and $S_{CEU}^{H} \leq -0.2$, respectively. These calls allowed both haplotypes of the entire HeLa genome, outside of LOH regions, to be "painted" as either CEU-like or YRI-like (Supplementary Fig. D.4.19).

D.2.15 Haplotype analysis of 8 additional HeLa strains

Aim: To explore the haplotype patterns of the 8 HeLa strains sequenced to low coverage by comparing their haplotypes with that of HeLa CCL-2.

Data sources: HeLa (CCL-2) haplotype scaffolds of phased variants; HeLa (CCL-5,6,13,17,21,23,25,62) shotgun SNV calls with low read coverage.

The read coverage on the 8 HeLa strains is too low for *ab initio* LOH analysis, but it can be compared with the result of the variant phasing on HeLa strain CCL-2. The following analysis was performed on all 8 of the HeLa strains but is illustrated for CCL-13 only in Supplementary Figs. D.4.28 and D.4.29.

The genome was divided into large SUNK windows of ~800 kbp each. In each window, the set of all variants phased in CCL-2 was inspected, and the coverage of these variant positions in reads from CCL-13 was tabulated. Each CCL-13 read covering one of these variant positions was phased as “A” or “B” according to the CCL-2 haplotype containing the allele seen in the CCL-13 read; then a total fraction of “A coverage” / “B coverage”, or “A/B ratio”, was calculated for the window.

Supplementary Fig. D.4.28 shows the A/B ratios for the entire genome (excluding windows in which HeLa CCL-2 is in LOH and thus contains no phased variants.) The A/B ratio is expected to be close to the allele balance of the CCL-2 haplotypes in CCL-13.Hence, in regions where the A/B ratio is close to 1 or 0, one of the CCL-2 haplotypes is likely absent in CCL-13. This can be seen in chr4q, chr9p, and chr18p. Also note regions in which the A/B ratio fluctuates rapidly between high and low numbers, such as chr4. These are areas in which CCL-2 has a balanced copy number, thus the haplotype blocks could not be combined into large-scale scaffolds, and the concept of haplotypes “A” and “B” is not expected to be consistent across the chromosome. Supplementary Fig. D.4.29 shows the A/B ratio and the copy number profile across chromosome 9 of HeLa CCL-13. CCL-13 has a similar copy number profile on chromosome 9 as CCL-2 (Figure 6.1b), except that the p arm is diploid rather than triploid. Notably, the A/B ratio is very close to 1 on all of chromosome 9p, implying that CCL-2 haplotype B is not present in CCL-13. These two lines of evidence strongly suggest that CCL-13 has lost its sole copy of the B haplotype of chromosome 9q, changing from a triploid 3:2:1 state to a diploid 2:2:0 state with LOH.

D.2.16 Identification of putative post-aneuploidy mutations

Aim: To identify somatic mutations in HeLa that are very likely to have arisen after duplications in their region.

Data sources: HeLa (CCL-2) genome-wide LOH and haplotype-resolved copy number (HRCN) profiles; HeLa (CCL-2) haplotype scaffolds of phased variants.

We searched for candidate somatic mutations, starting from the set of all biallelic single-nucleotide variants ascertained by whole-genome shotgun sequencing (n=3,994,385), followed by removal of homozygous sites (41.4% of sites with zero high-quality reads matching the reference allele). Of the remaining sites (n=2,339,608), those present among the 1000 Genomes Project (86.7%) were taken as likely to be inherited rather than somatic variants, and were discarded. The remaining sites (n=311,847) included true somatic mutations, rare or private inherited variants not observed within the 1000 Genomes Project dataset, and spurious calls corresponding to sequencing or mapping errors. To stringently exclude false positive mutations at sites of systematic error (e.g. artifacts occurring near repeat tracts or due to unannotated paralogs), we removed sites at which any one of the 11 control genomes’ alignments contained the mutant allele (at a frequency of at least 10%), or where at least one of the control genomes had missing coverage (fewer than 10 reads). Additionally, we removed sites within annotated segmental duplications. After application of these filters, 66,829 sites remained.

We then selected sites that were polymorphic between duplicated copies of the same germline haplotype and are therefore *de facto* somatic mutations occurring after the haplotypes’ duplication (Supplementary Fig. D.4.20). To find such sites, we searched dilution pools for clones derived from the same germline haplotype but with differing genotypes at the candidate site. Clones derived from the opposite germline

haplotype were required to match the reference allele (except for in LOH regions, where only one germline haplotype remains). We excluded sites at which haplotype-resolved copy number was ambiguous (either uncalled, or where calls based upon low- and high-resolution windows were discordant), as well as sites lacking coverage from any phased clones. In sum, this confirmed 8,165 somatic mutations (Supplementary Table D.3.14), throughout the genome (Supplementary Fig. D.4.22)

Requiring observation of both the mutant and wild-type alleles on distinct clones from the same germline haplotype is a stringent filter, but it rejects true mutations in cases where not enough clones are sampled to observe both the mutant and wild-type alleles on the same germline haplotype. To estimate the proportion of true somatic mutation sites lost by undersampling, we considered the expected number of cases in which our method would fail to sample enough clones to observe both alleles among sites containing true variants (sites shared between the HeLa genome and the 1000 Genomes Project). We started with 1,203,938 phased, heterozygous sites present within HRCN=3:2:1 regions. At the 21.5% of sites covered by zero or one clones from the duplicated haplotype, it would have been impossible to observe both alleles in separate clones. For sites with at least two clones from the duplicated haplotype, we computed the expected number of sites for which both wild-type and mutant clones would have been observed under the assumption that each are sampled with equal likelihood:

$$n_j = [\#sites\ with\ j\ clones\ derived\ from\ duplicated\ haplotype]$$

The expected number of sites with *j* clones derived from duplicated haplotype, where at least one wild-type and one mutant clone are observed is:

$$x_j = n_j(1 - 0.5^{j-1})$$

The overall expected sensitivity is:

$$\frac{\sum_{j=2}^{\infty} x_j}{\sum_{j=0}^{\infty} n_j}$$

By this measure, the expected sensitivity for validation of somatic sites discovered by shotgun sequencing is 0.611. Extrapolating from HRCN=3:2:1 regions provides a conservative (low) estimate for sensitivity, because clone sampling depth is greater within more heavily duplicated regions (that is, those for which the duplicated haplotype is present at more than two copies), increasing the likelihood of sampling enough clones to observing both alleles. Therefore, the expected true number of somatic mutations is no more than $[8{,}165/(0.611)]$ = 13,364.

For a large majority of the clone-confirmed somatic mutations, the shotgun allele frequencies (Supplementary Fig. D.4.21) are consistent with the idea that the mutations occurred after all duplications have taken place and the cell line has reached stable copy, although these allele frequencies were not used as a criterion for selecting the mutations (other than to exclude invariant sites - those with allele frequency equal to zero or one). Among regions with HRCN=3:2:1, 3,919 sites (94.9%) and among 3:3:0 regions, 1,334 sites (92.8%) had shotgun allele frequencies less than 50%, over which range presence on only one copy of duplicated haplotype A is more likely than presence among both. Among HRCN 4:2:2, 4:3:1, and 4:4:0 regions, the respective counts and proportions of sites consistent with mutation strictly after duplication, and therefore presence on only one copy (that is, sites with allele frequency <= 37.5%), were 231 (80.5%), 153 (70.8%), and 809 (90.8%). The 3% of the genome with higher copy number is omitted from this analysis, because of the potential for bias against the variants at low shotgun allele frequencies.

D.2.17 HPV-18 integration site assembly

Aim: To assemble the complex repetitive structure of the HPV-18 integration site on chromosome 8q24.21.

Data sources: Fosmid clone dilution pool reads; Final copy number calls; Raw shotgun reads and read depth; PCR assays.

In order to first identify potential locations of HPV-18 integration in the genome, the whole-genome shotgun and fosmid clone dilution pool reads were aligned to a modified reference that contained the HPV-18 genome sequence as well as the sequence of the fosmid vector backbone (for clone end determination). Clone pools were then sorted based on coverage with respect to the HPV-18 genome and read pairs flagged as interchromosomal between HPV-18 and a human chromosome were used to determine the site of integration in the human reference sequence. This region was constrained exclusively to chromosome 8q24.21 and therefore all pools with coverage spanning chromosome 8q24.21 were also pulled for further analysis. Potential breakpoints were then determined using the breakpoint-spanning read pairs in the fosmid pools as well as the shotgun read pairs. These breakpoints were then confirmed by breakpoint PCR of all possible primer pair combinations for each breakpoint, followed by the construction of shotgun libraries of the amplicons using transposase-based library preparation[59] and then sequencing on a MiSeq (Supplementary Fig. D.4.31). Coverage profiles were then generated for clone pools with coverage of HPV-18 and/or chromosome 8q24.21 (Supplementary Fig. D.4.30). These coverage profiles of a fixed expected length were then used in conjunction with shotgun read depth as well as copy number calls to determine the exact repetitive structure of the integration locus (Figure 6.2).

D.2.18 Directional PolyA$^+$ RNA-Seq library construction

Aim: To generate high-depth in-house directional, PolyA RNA-Seq libraries in order to better assess effects of copy number and transcription.

Input: HeLa S3 cell culture.

HeLa S3 was chosen over HeLa CCL-2 due to the clonality of the strain. However, copy number heterogeneity has been observed in HeLa S3, it is notably less than that of HeLa CCL-2. Total RNA was isolated using the RNeasy mini kit (QIAGEN) followed by quiantification using a NanoDrop 8000 spectrometer. 1 µg of total RNA was then used for mRNA isolation using the NEBNext Poly(A) mRNA Magentic Isolation Module (NEB) followed by directional RNA-Seq library preparation using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB). All kit protocols were carried out according to manufacturer instructions. The library was then sequenced on one lane of HiSeq 2000 using paired-end 51 bp reads.

D.2.19 RNA-Seq in-depth computational analysis

Aim: To provide additional RNA-Seq depth of the HeLa S3 strain.

Data Sources: HeLa S3 RNA-Seq fastq files (from Nagaraj *et. al.* (2011)[105], ENCODE CSHL Long PolyA Cell, and in-house directional RNA-Seq).

Reads were aligned to GRCh37 as well as haplotype-specific HeLa reference for haplotype A and haplotype B (Supplementary Note D.2.21) with TopHat (v2.0.6)[156] using the RefGene "gtf" file downloaded from the UCSC genome browser. The GRCh37 alignment was then used for transcript quantification using Cufflinks (v2.0.2)[157]. For the in-house RNA-Seq, quantification was also performed on each haplotype alignment resulting in minimal differences (Supplementary Note D.2.21, Supplementary Fig. D.4.37). Transcripts with RPKM scores greater than or equal to 1 were used for further comparisons to mitigate noise associated with inactive transcripts. Correlations were performed using the "cor.test" function in R for both Pearson and Spearman tests. Additionally, a correlation between in-house RNA-Seq and ENCODE (CSHL, Cell, Long, PolyA) RNA-Seq was also performed (Spearman = 0.646, Pearson, 0.363; Supplementary Fig. D.4.33). Lastly, reads unaligned to the human genome were aligned to the HPV-18 reference to investigate transcription levels iof the integrated viral genome (Supplementary Fig. D.4.32).

Global transcription levels by copy number were assessed by using genes in regions of constant copy number between CCL-2 and S3 and split by underlying copy number. This resulted in an increasing trend with a *p*-value of 0.075 according to a permutation analysis by which copy number identities are shuffled

at each iteration (to a total of 100,000 iterations), yet retaining the total number of genes in each copy number bin (Figure 6.3a). Scores were then normalized to underlying copy number and the test performed again which resulted in a *p*-value of 0.485 according to the previously described permutation analysis (Figure 6.3b). While the increasing trend is not significant enough to definitively claim increased expression by copy number, the comparison to the copy number normalized *p*-value which is extremely near the null hypothesis is convincing nonetheless.

D.2.20 ENCODE epigenome and RNA-Seq phasing

Aim: To assign haplotype phase to transcripts and epigenomic data tracks generated on HeLa.

Data Sources: In-house generated directional, PolyA RNA-Seq on HeLa (CCL-2) and HeLa (S3) as aligned bam files, downloaded HeLa (CCL-2) RNA-Seq from Nagaraj *et. al.* (2011)[105] as aligned bam files, and downloaded ENCODE data sets on HeLa (S3) comprised of: DNase (UW), FAIRE-Seq (UNC, 2 tracks), Histone modifications (Broad, 13 tracks), Histone modifications (UW, 3 tracks), Repli-Seq (UW, 6 tracks), RNA-Seq (CSHL, 9 tracks), RNA-Seq (CalTech, 4 tracks), Transcription factor binding (Hudson-Alpha, 4 tracks), Transcription factor binding (Stanford, Yale, Duke, Harvard, 48 tracks) as aligned bam files and with called peaks where appropriate.

For all RNA-Seq, RPKM (Reads Per Kilobase per Million) scores for global levels of transcription were generated by tallying the number of reads per kilobase window of the genome. RPKM scores were also generated by a gene-model based approach using Cufflinks (v2.0.2)[157]. All bam files were then genotyped for all phased variants, and the fractional contribution of each haplotype to each RPKM or peak score was calculated. Copy number normalization was then performed by dividing the haplotype-specific score by the underlying copy number of that haplotype to find the haplotype contribution per copy. A global view of these tracks can be found in Supplementary Fig. D.4.34.

D.2.21 Reference bias assessment and removal

Aim: To identify the contribution of reference bias in alignment and subsequent allele balance calculation.

Data Sources: Raw fastq sequence read files, homozygous and phased heterozygous SNVs in HeLa, peaks called for corresponding sequence tracks.

Reference bias in allele balance at informative sites was determined by calculating the haplotype A / (haplotype A + haplotype B) ratio at each HRCN classification split by sites where the reference allele is either haplotype A or haplotype B. This resulted in globally lower ratios for positions where haplotype B is the reference allele and is summarized for all regulatory peaks in HRCN 3:2:1 regions in Supplementary Figs. D.4.36 and D.4.38. To remove this bias, two new reference genomes were generated with all homozygous SNVs as well as either haplotype A or haplotype B heterozygous SNVs (referred to as HAPREF) followed by alignment of raw reads for a subset of the ENCODE data sets as well as in-house RNA-Seq to each reference and tallying counts for respective heterozygous calls for each haplotype. This process effectively removed the reference bias, as shown in Supplementary Fig. D.4.38. We next assessed the effect of reference bias on haplotype-specific peak calling (Supplementary Note D.2.22) by comparing results from the GRCh37 and HAPREF data sets revealing only a very minor shift in the number of called outliers (Supplementary Fig. D.4.44). For the in-house RNA-Seq of HeLa S3, gene transcript quantifications were made on the GRCh37 alignment as well as for each individual haplotype reference and compared (Supplementary Fig. D.4.37) resulting in 0.54% of transcripts with a difference in RPKM score ≥10% between haplotype A and haplotype B references with 0.63% and 0.64% of transcripts showing a ≥10% difference for haplotype A and haplotype B references respectively when compared to GRCh37.

D.2.22 Identifying haplotype-specific peaks in ENCODE data

Aim: To identify peaks originating almost exclusively from a single haplotype.

Data Sources: Phase-resolved peaks from epigenetic data tracks.

In order to quantitatively asses the number of haplotype specific peaks for ENCODE data tracks, a scoring metric was derived that takes into account both the statistical significance of the peak allele balance differing from the null hypothesis of the HRCN theoretical allele balance as well as quantifying is the bias of the allele balance away from the null hypothesis of the HRCN mean allele balance. The first score was calculated using the following probability mass function:

$$f(t; a; c_A; c_B) = \left(\frac{t!}{a!\,(t-a)!}\right)\left(\frac{c_A}{c_B}\right)^a\left(1-\frac{c_A}{c_B}\right)^{t-a}$$

where t = the total coverage at the position, a = the number of bases supporting the haplotype A allele, and $c_{A/B}$ = the copy number for haplotype A or haplotype B at the position. The resulting score is a *p*-value corresponding to the significance that the alleles observed at the peak are different from the null hypothesis. This metric only provides a *p*-value against the null hypothesis. In order to quantify the difference in allele balance, a second normalized Gaussian score was applied:

$$f(x) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where x = the haplotype A / (haplotype A + haplotype B) ratio at the position, σ = the standard deviation of haplotype A / (haplotype A + haplotype B) ratios for the entire HRCN state, and μ = the mean of haplotype A / (haplotype A + haplotype B) ratios for the entire HRCN state. This score (shown for CTCF binding peaks in Supplementary Fig. D.4.41) is then normalized across all HRCN classifications. These two scores can then be used with a cutoff to identify peaks of extreme haplotype imbalance (Supplementary Figs. D.4.42 and D.4.43).

D.2.23 Haplotype imbalance identification

Aim: To identify regions of excessive haplotype imbalance.

Data Sources: Phased epigenomic and transcriptomic data sets.

Regions of excessive haplotype imbalance were identified using a sliding window approach (1.5 Mb sliding by 0.5 Mb) which took into account the ranking of the peak weight within each respective data set with a cap set at the top 50th percentile (i.e. smallest score of 0.5 and decreasing out to 1.0 as the smallest peak, so as not to over-weight excessively high peaks) as well as the haplotype imbalance score for the dominating haplotype (positive score for A, negative score for B) based on coverage at haplotype-resolved heterozygous variants and normalized for their underlying haplotype copy number which was set to a maximum ratio of 10 for haplotype A and -10 for haplotype B in order to minimize noise. The score for each track was calculated by dividing the haplotype imbalance score by the peak weighting score to produce maximums of 20 for haplotype A and -20 for haplotype B. The absolute values for all tracks within the window were then summed to produce the final window score. Several iterations of the capping values were implemented to highly similar results; these final values were used as they mitigated noise caused by single, dominating peaks and tended to favor windows containing multiple haplotype imbalanced data sources. Windows were then ranked and filtered to remove regions of LOH followed by combining like-windows in the top set of hits and rescoring merged windows (Supplementary Fig. D.4.45).

D.3 Supplementary tables for Chapter 6

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 D.3.1. Sequence 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 3 kbp jumping libraries as well as fosmid based 40 kbp 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**

*Assuming 2.8 Gbp alignable reference

Table D.3.2. HGDP control genomes.
Sequencing data summary for 11 HGDP control individuals from Rohland, N. and Reich, D (2012).

	African	European		European		African
	DNK02, Dinka	HGDP00521, French	HGDP00542, Papuan	HGDP00665, Sardinian	HGDP00778, Han	HGDP00927, Yoruban
Number of SNVs	**4,490,051**	**3,787,833**	**3,722,767**	**3,777,671**	**3,833,823**	**4,588,782**
Number of indels	**359,565**	**326,732**	**301,328**	**314,531**	**324,555**	**378,632**
Number of 1kG SNVs	**4,014,503**	**3,403,243**	**3,192,306**	**3,383,120**	**3,420,873**	**4,128,798**
Number of non-1kG SNVs	**475,548**	**384,590**	**530,461**	**394,551**	**412,950**	**459,984**
Number of 1kG indels	**213,292**	**183,321**	**170,851**	**179,872**	**183,241**	**221,893**
Number of non-1kG indels	**146,273**	**143,411**	**130,477**	**134,659**	**141,314**	**156,739**
% SNVs that are homozygous	35.38%	39.33%	49.19%	39.57%	42.36%	35.06%
Ti/Tv for SNVs in 1kG	2.15	2.15	2.15	2.15	2.14	2.15
Ti/Tv for SNVs not in 1kG	1.62	1.59	1.70	1.61	1.58	1.59
Private Protein-Altering (PPA) SNVs	**304**	**117**	**678**	**199**	**249**	**143**
PPA SNVs in COSMIC	**3**	**0**	**7**	**1**	**2**	**1**
PPA SNVs in Cancer Genes	**10**	**5**	**17**	**3**	**4**	**4**
PPA indels	8	8	16	13	19	19
PPA indels in COSMIC	0	0	0	0	0	0
PPA indels in Cancer Genes	0	0	0	0	0	0
Total bases in homozygous tracts	**30,173,695**	**42,933,724**	**124,265,489**	**62,502,510**	**50,148,051**	**62,624,662**

		African	African		African
	HGDP00998, Karitiana	HGDP01029, San	HGDP01284, Madenka	HGDP01307, Dai	HGDP0456, Mbuti
Number of SNVs	**3,570,099**	**5,015,502**	**4,621,804**	**3,774,107**	**4,783,268**
Number of indels	**303,842**	**363,202**	**352,080**	**321,657**	**337,610**
Number of 1kG SNVs	**3,156,213**	**4,009,840**	**4,108,845**	**3,380,387**	**4,100,827**
Number of non-1kG SNVs	**413,886**	**1,005,662**	**512,959**	**393,720**	**682,441**
Number of 1kG indels	**169,091**	**206,717**	**214,807**	**180,726**	**204,927**
Number of non-1kG indels	**134,751**	**156,485**	**137,273**	**140,931**	**132,683**
% SNVs that are homozygous	51.21%	37.33%	34.56%	42.46%	38.21%
Ti/Tv for SNVs in 1kG	2.14	2.15	2.15	2.14	2.15
Ti/Tv for SNVs not in 1kG	1.62	1.81	1.65	1.59	1.76
Private Protein-Altering (PPA) SNVs	**276**	**1258**	**212**	**238**	**626**
PPA SNVs in COSMIC	**1**	**7**	**3**	**2**	**2**
PPA SNVs in Cancer Genes	**10**	**18**	**5**	**5**	**15**
PPA indels	7	48	14	6	35
PPA indels in COSMIC	0	0	0	0	0
PPA indels in Cancer Genes	1	0	0	0	0
Total bases in homozygous tracts	**327,170,977**	**51,559,155**	**25,223,209**	**65,259,057**	**77,220,484**

	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 D.3.3. 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.

Gene	Class
ADAM18	stop-gained
AKAP2,PALM2-AKAP2	stop-gained
ANKRD35	stop-gained
ASXL3	stop-gained
BHLHE22	stop-gained
C10orf68	frameshift
C1orf162	frameshift
CELF2	isplice-5
COL5A2	stop-gained
COL9A2	isplice-3
CTBP2	frameshift
ELMO1	frameshift
EPPK1	stop-gained
FAM20C	frameshift
FREM3	stop-gained
GBA2	stop-gained
GSDMD	frameshift
IL1RAP	frameshift
ING1	frameshift
KRT4	frameshift
LRP2	stop-gained
MAGEC3	frameshift
MMRN1	stop-gained
MYOM2	frameshift
OR1Q1	frameshift
QTRT1	stop-gained
RPS29	stop-gained
SAMD4B	isplice-5
SEMA4G,MRPL43	splice-5
STK36	stop-gained
SULT2B1	frameshift
TBXAS1	stop-gained
TINAG	stop-gained
USP28	isplice-5
XRN2	frameshift

Table D.3.4 Genes altered by private, protein-altering SNVs in HeLa

Positon	Mutation	ID	Gene	Class
chr1:40768483	T>TGGAG	dbSNP_107	COL9A2	splice-3
chr1:112020370	A>AG	none	C1orf162	frameshift
chr3:190326934	TC>T	none	IL1RAP	frameshift
chr7:286468	G>GC	none	FAM20C	frameshift
chr7:36895197	TG>T	none	ELMO1	frameshift
chr8:2054156	TGA>T	none	MYOM2	frameshift
chr8:144644488	TC>T	none	GSDMD	frameshift
chr9:125377175	AC>A	dbSNP_132	OR1Q1	frameshift
chr10:33136818	TAA>T	none	C10orf68	frameshift
chr10:126727614	CA>C	dbSNP_134	CTBP2	frameshift
chr12:53207583	C>CCACCAAAGCCACCAGTGCCGAAACC	dbSNP_120	KRT4	frameshift
chr13:111368115	T>TC	none	ING1	frameshift
chr19:49079289	AC>A	none	SULT2B1	frameshift
chr20:21319715	AAG>A	none	XRN2	frameshift
chrX:140967154	GCTACACCCTTTCCCTTC>G	dbSNP_130	MAGEC3	frameshift
chr10:11374574	CTCCAGTCTTCCTCTTCGGCATGCCCTGGAAGCTTCT>C	dbSNP_134	CELF2	splice-5
chr11:113675391	TTAC>T	none	USP28	splice-5
chr19:39843014	G>GGTA	none	SAMD4B	splice-5

Table D.3.5 Genes altered by private-protein altering Indels in HeLa

Position	Variant (ref>alt)	dbSNP	Het / Hom	Gene	Amino Acid No.	Class	Alteration	PolyPhen	Grantham	PhastCons	GERP	COSMIC Match	SangerCGP Somatic	SangerCGP Type
chr1:3347438	**C>T**	**none**	**het**	**PRDM16**	**1096**	**miss.**	**ALA,VAL**	**unknown**	**64**	**0.768**	**4.03**	**gene**	**gene**	**translocation**
chr1:228476552	T>G	none	het	OBSCN	3434	miss.	CYS,TRP	unknown	215	0.981	-1.98	exact, COSM210235	none	none
chr4:1808852	**G>A**	**none**	**het**	**FGFR3**	**761**	**miss.**	**ASP,ASN**	**unknown**	**23**	**1**	**4.53**	**gene**	**gene**	**missense, translocation**
chr9:139418189	C>T	none	hom	NOTCH1	128	miss.	ARG,HIS	unknown	29	0.996	2.29	gene	gene	translocation, missense, other
chr14:99641176	**G>A**	**none**	**het**	**BCL11B**	**667**	**miss.**	**PRO,LEU**	**unknown**	**98**	**0.88**	**3.95**	**gene**	**gene**	**translocation**

Position	Variant (ref>alt)	dbSNP	HGDP	Gene	Amino Acid No.	Class	PhastCons	GERP	COSMIC Match Type	SangerCGP Somatic	SangerCGP Type
chr22:41565536	**TTCA>T**	**none**	**No**	**EP300**	**1401**	**coding**	**1**	**5.55**	**gene**	**gene**	**translocation, frameshift, missense, nonsense, other**

Table D.3.6. HeLa CCL-2 private protein-altering SNVs/indels overlapping COSMIC or SCGC.
Variant alleles listed were called in HeLa CCL-2 and found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Also shown for each gene is overlap with the Sanger Cancer Gene Census (SCGC).

Chr	Start	End	CN
chr1	10,485	126,219	9
chr1	132,314	252,387	3
chr1	527,487	2,582,492	3
chr1	2,693,799	3,849,920	3
chr1	3,994,920	13,219,912	3
chr1	13,351,155	13,386,377	3
chr1	13,514,395	13,519,395	3
chr1	13,735,227	24,256,637	3
chr1	24,281,002	24,313,498	9
chr1	24,338,498	25,590,373	3
chr1	25,590,373	25,621,145	7
chr1	25,621,147	29,878,082	3
chr1	30,028,082	55,092,300	3
chr1	55,103,399	93,613,021	3
chr1	93,613,021	93,620,247	7
chr1	93,620,247	103,785,126	3
chr1	103,913,906	108,296,689	3
chr1	108,296,689	119,010,719	4
chr1	119,010,719	120,936,685	5
chr1	121,086,698	121,349,829	5
chr1	121,349,829	121,351,595	2
chr1	142,543,754	142,837,872	2
chr1	142,840,597	142,893,361	1
chr1	142,955,514	142,966,274	2
chr1	143,117,763	143,544,525	4
chr1	143,660,151	143,770,922	4
chr1	143,871,002	144,191,705	4
chr1	144,328,314	144,338,314	4
chr1	144,452,606	144,708,795	4
chr1	144,811,669	145,833,117	4
chr1	145,987,067	146,017,396	4
chr1	146,243,767	147,424,817	4
chr1	147,528,801	147,551,330	4
chr1	147,706,477	148,026,034	4
chr1	148,176,038	148,361,358	4
chr1	148,511,358	148,754,909	4
chr1	148,785,523	148,824,939	4
chr1	148,845,858	149,032,493	4
chr1	149,038,041	149,221,211	100
chr1	149,221,211	149,226,886	4
chr1	149,230,930	149,263,449	28
chr1	149,271,468	149,335,570	4
chr1	149,335,570	149,368,049	21
chr1	149,378,542	149,433,401	100
chr1	149,516,996	152,555,480	4
chr1	152,555,480	152,590,169	100
chr1	152,590,169	155,182,045	4
chr1	155,182,045	155,190,755	15
chr1	155,190,755	156,514,292	4
chr1	156,514,292	156,674,478	5
chr1	156,674,478	158,868,760	4
chr1	158,869,586	159,648,906	4
chr1	159,649,530	174,218,685	4
chr1	174,218,685	174,548,204	3
chr1	174,548,204	205,926,632	4
chr1	206,072,707	206,336,056	4
chr1	206,482,221	214,775,748	4
chr1	214,775,748	222,653,660	3
chr1	222,656,865	236,158,340	3
chr1	236,158,340	248,912,526	4
chr1	249,062,526	249,225,340	4
chr2	13,454	40,766,700	3
chr2	40,767,297	41,187,544	3
chr2	41,187,544	41,851,511	2
chr2	41,851,511	76,774,665	3
chr2	76,775,282	87,177,338	3
chr2	87,212,338	87,252,412	3
chr2	87,268,655	87,608,185	3
chr2	87,740,602	89,625,088	3
chr2	89,880,436	89,937,593	3
chr2	89,942,593	90,265,886	3
chr2	90,407,327	90,475,028	3
chr2	91,635,768	92,267,489	3
chr2	95,326,171	106,760,712	3
chr2	106,760,712	106,880,397	2
chr2	106,881,595	108,275,317	2
chr2	108,280,317	110,113,857	2
chr2	110,248,857	110,542,218	2
chr2	110,733,137	111,008,704	2
chr2	111,023,704	111,151,966	2
chr2	111,344,425	117,771,913	2
chr2	117,771,913	117,777,090	4
chr2	117,777,090	117,783,384	5
chr2	117,783,384	131,221,523	2
chr2	131,233,997	131,288,039	2
chr2	131,328,039	138,602,269	2
chr2	138,602,269	138,612,448	4
chr2	138,612,448	146,855,624	2
chr2	146,855,624	146,872,752	5
chr2	146,872,752	162,135,049	2
chr2	162,135,384	162,144,182	4
chr2	162,149,182	177,268,772	2
chr2	177,271,565	183,090,568	2
chr2	183,091,068	184,849,287	2
chr2	184,849,787	208,352,967	2
chr2	208,352,967	208,359,328	5
chr2	208,359,328	222,365,502	2
chr2	222,365,502	222,672,340	1
chr2	222,672,340	227,165,726	2
chr2	227,169,258	228,771,277	2
chr2	228,771,277	228,783,063	5
chr2	228,783,063	243,062,019	2
chr2	243,062,039	243,085,389	5
chr2	243,085,389	243,187,975	2
chr3	60,000	9,702,436	3
chr3	9,702,436	10,013,210	4
chr3	10,013,210	48,626,439	3
chr3	48,626,439	48,631,230	5
chr3	48,631,230	60,033,311	3
chr3	60,033,311	60,290,678	2
chr3	60,290,678	60,556,563	1
chr3	60,561,018	60,563,081	1
chr3	60,565,698	60,567,678	1
chr3	60,570,256	60,577,461	1
chr3	60,579,850	60,582,223	1
chr3	60,582,892	60,583,768	1
chr3	60,584,863	60,586,026	1
chr3	60,590,257	60,591,140	1
chr3	60,592,458	60,593,618	1
chr3	60,594,353	60,595,624	1
chr3	60,598,849	60,602,894	1
chr3	60,604,033	60,615,050	1
chr3	60,616,152	60,616,855	1
chr3	60,618,453	60,620,938	1
chr3	60,624,599	60,629,561	1
chr3	60,632,813	60,633,705	1
chr3	60,634,523	60,635,979	1
chr3	60,637,445	60,637,945	1
chr3	60,655,623	61,125,403	1
chr3	61,125,403	65,531,331	2
chr3	65,531,331	65,894,483	1
chr3	65,894,483	65,901,495	3
chr3	65,901,495	66,025,389	1
chr3	66,025,389	68,736,251	2
chr3	68,741,251	68,747,683	5
chr3	68,747,683	90,124,968	2
chr3	90,124,968	90,311,949	3
chr3	93,527,460	95,000,138	3
chr3	95,000,138	98,947,133	2
chr3	98,948,732	99,228,718	2
chr3	99,228,718	112,643,760	3
chr3	112,643,760	112,650,208	5
chr3	112,650,208	120,162,538	3
chr3	120,164,301	129,076,376	3
chr3	129,077,560	129,079,153	3
chr3	129,084,153	136,603,244	3
chr3	136,603,244	136,632,377	7
chr3	136,632,377	145,506,305	3
chr3	145,506,305	162,514,458	4
chr3	162,569,400	162,570,101	4
chr3	162,584,403	162,585,138	4
chr3	162,602,315	162,603,291	4
chr3	162,623,850	192,875,787	4
chr3	192,883,457	192,883,957	4
chr3	192,884,937	197,910,311	4
chr4	27,175	9,170,304	2
chr4	9,173,177	9,191,016	2
chr4	9,191,016	9,197,772	5
chr4	9,199,164	9,217,084	5
chr4	9,232,561	9,251,799	2
chr4	9,370,793	10,217,781	2
chr4	10,217,781	10,227,327	5
chr4	10,227,327	10,229,619	2
chr4	10,229,619	10,236,065	5
chr4	10,236,065	49,323,812	2
chr4	49,488,941	49,515,469	1
chr4	49,516,028	49,631,009	2
chr4	52,687,650	104,741,167	2
chr4	104,741,167	104,746,728	4
chr4	104,746,728	104,761,899	6
chr4	104,761,899	128,072,735	2
chr4	128,072,735	128,468,369	1
chr4	128,468,369	137,830,665	2
chr4	137,830,665	137,842,008	4
chr4	137,842,008	158,519,508	2
chr4	158,519,508	161,335,829	1
chr4	161,335,829	166,003,594	2
chr4	166,004,460	173,340,822	2
chr4	173,340,822	173,348,822	4
chr4	173,348,822	173,696,311	2
chr4	173,696,311	173,701,836	6
chr4	173,701,836	186,442,146	2
chr4	186,443,952	190,966,833	2
chr4	190,977,588	191,006,231	2
chr5	17,963	34,102,549	6
chr5	34,102,549	36,235,436	5
chr5	36,235,436	36,240,624	9
chr5	36,240,624	37,670,470	5
chr5	37,670,470	37,692,616	14
chr5	37,692,616	38,267,285	5
chr5	38,267,285	38,282,099	9
chr5	38,282,099	40,338,544	5
chr5	40,347,334	45,272,759	5
chr5	45,272,759	45,913,294	4
chr5	46,093,294	46,123,423	4
chr5	49,558,423	49,562,042	4
chr5	49,562,042	57,324,465	3
chr5	57,330,903	58,449,735	3
chr5	58,449,735	58,849,699	1
chr5	58,849,699	68,928,876	3
chr5	69,237,988	69,488,612	3
chr5	69,715,594	69,957,192	3
chr5	70,150,381	70,174,418	3
chr5	70,189,418	70,194,418	3
chr5	70,305,412	70,311,051	3
chr5	70,351,051	70,494,904	3
chr5	70,612,810	83,952,451	3
chr5	83,954,544	92,062,043	3
chr5	92,062,043	92,066,124	7
chr5	92,066,124	102,971,957	3
chr5	102,971,957	102,977,687	7
chr5	102,977,687	133,557,726	3
chr5	133,557,726	133,560,891	7
chr5	133,560,891	134,259,302	3
chr5	134,259,302	134,264,208	29
chr5	134,264,208	135,816,902	3
chr5	135,816,902	135,828,136	5
chr5	135,828,136	147,553,050	3
chr5	147,553,908	167,076,615	3
chr5	167,076,615	167,082,610	8
chr5	167,082,610	180,854,202	3
chr6	143,476	381,695	2
chr6	381,695	9,273,587	3
chr6	9,273,587	9,284,284	7
chr6	9,284,284	28,710,761	3
chr6	28,750,761	28,859,124	3
chr6	28,889,124	28,894,124	3
chr6	28,909,124	28,914,124	3
chr6	28,948,243	28,973,243	3
chr6	29,068,414	29,070,560	3
chr6	29,188,248	29,367,739	3
chr6	29,465,465	29,731,197	3
chr6	29,792,349	29,802,349	1
chr6	29,805,772	29,815,772	1
chr6	29,820,772	29,825,772	1
chr6	29,825,772	29,945,461	3
chr6	29,960,461	29,976,030	3
chr6	30,024,161	30,183,107	3
chr6	30,208,107	30,213,107	3
chr6	30,368,107	30,438,309	3
chr6	30,465,210	30,470,210	3
chr6	30,575,210	31,054,523	3
chr6	31,059,523	31,064,523	3
chr6	31,074,523	31,145,812	3
chr6	31,165,812	31,170,812	3
chr6	31,190,812	31,247,042	3
chr6	31,248,091	31,282,661	10
chr6	31,318,726	31,443,086	22
chr6	31,443,086	31,482,846	3
chr6	31,577,846	31,592,846	1

Chr	Start	End	CN
chr6	31,602,846	31,622,846	1
chr6	31,652,846	31,662,846	1
chr6	31,707,846	31,879,275	3
chr6	31,963,308	32,319,271	3
chr6	32,329,271	32,369,271	3
chr6	32,369,271	32,413,039	14
chr6	32,438,039	32,455,726	3
chr6	32,489,181	32,494,834	3
chr6	32,494,834	32,665,421	63
chr6	32,665,421	32,674,736	3
chr6	32,679,736	32,797,656	3
chr6	32,797,656	32,807,656	8
chr6	32,817,656	32,907,794	8
chr6	32,937,794	33,101,311	16
chr6	33,106,311	33,359,926	3
chr6	33,394,926	33,399,926	3
chr6	33,454,926	48,930,534	3
chr6	48,930,534	48,936,904	8
chr6	48,936,904	50,576,822	3
chr6	50,576,822	50,599,734	8
chr6	50,599,734	53,929,151	3
chr6	53,929,151	53,932,712	7
chr6	53,932,712	54,663,780	3
chr6	54,664,967	58,299,919	3
chr6	58,299,919	58,312,890	1
chr6	58,312,890	58,720,004	3
chr6	61,966,206	62,127,801	3
chr6	62,233,589	64,449,989	3
chr6	64,449,989	69,111,189	2
chr6	69,116,189	70,015,233	3
chr6	70,016,983	74,592,036	3
chr6	74,602,228	78,967,832	3
chr6	78,973,065	78,975,079	1
chr6	78,976,958	78,977,580	1
chr6	79,000,337	79,001,223	1
chr6	79,011,277	79,011,777	1
chr6	79,023,451	79,024,370	1
chr6	79,026,987	79,027,770	1
chr6	79,036,566	95,677,643	3
chr6	95,827,643	116,641,244	3
chr6	116,641,244	119,165,135	2
chr6	119,172,662	140,769,716	2
chr6	140,769,716	140,781,735	5
chr6	140,781,735	165,725,505	2
chr6	165,731,532	167,697,216	2
chr6	167,697,216	167,777,910	4
chr6	167,778,823	171,031,016	2
chr7	29,843	24,038,252	3
chr7	24,038,979	30,044,531	3
chr7	30,044,531	30,199,075	2
chr7	30,199,075	47,947,924	3
chr7	47,947,924	47,964,220	7
chr7	47,964,220	51,595,412	3
chr7	51,603,031	57,938,791	3
chr7	61,058,203	61,078,949	3
chr7	61,660,866	61,845,441	3
chr7	62,451,118	62,948,071	3
chr7	62,955,220	63,074,649	3
chr7	63,079,649	65,094,801	3
chr7	65,118,348	74,423,631	3
chr7	74,463,678	74,597,751	3
chr7	74,775,724	74,852,765	3
chr7	75,022,539	81,804,585	3
chr7	81,804,585	81,808,937	5
chr7	81,808,937	126,046,428	3
chr7	126,049,319	130,158,693	3
chr7	130,268,693	133,783,665	3
chr7	133,783,665	133,802,276	7
chr7	133,802,276	143,309,234	3
chr7	143,309,234	143,402,897	15
chr7	143,415,538	159,128,494	3
chr8	72,629	125,779	3
chr8	151,012	594,921	3
chr8	599,434	7,721,227	3
chr8	7,721,227	7,741,227	9
chr8	7,755,263	7,785,263	11
chr8	7,785,263	7,793,474	12
chr8	7,866,880	7,874,612	12
chr8	7,876,899	8,023,399	10
chr8	8,043,399	8,072,168	3
chr8	8,072,168	8,084,514	8
chr8	8,084,514	11,492,404	3
chr8	11,495,966	12,088,995	3
chr8	12,152,658	12,177,658	3
chr8	12,223,642	12,255,900	12
chr8	12,274,593	12,304,395	3
chr8	12,331,488	25,066,854	3
chr8	25,071,167	32,680,262	3
chr8	32,692,511	38,206,531	3
chr8	38,206,531	38,321,980	4
chr8	38,321,980	39,233,407	3
chr8	39,233,407	39,379,739	7
chr8	39,379,739	43,403,452	3
chr8	46,934,439	46,959,041	3
chr8	47,454,041	86,575,363	3
chr8	86,726,451	86,737,385	3
chr8	86,841,882	106,041,203	3
chr8	106,041,203	106,051,176	7
chr8	106,051,176	116,992,460	3
chr8	116,992,460	119,710,442	4
chr8	119,710,442	120,397,104	5
chr8	120,397,104	128,231,154	4
chr8	128,231,154	128,235,278	44
chr8	128,235,278	128,237,479	15
chr8	128,237,479	137,677,191	4
chr8	137,677,191	137,848,286	3
chr8	137,848,286	141,728,183	4
chr8	141,728,183	142,077,922	5
chr8	142,077,922	144,634,910	4
chr8	144,634,910	145,329,913	3
chr8	145,501,700	146,304,078	3
chr9	10,843	24,507,719	3
chr9	24,507,719	24,518,225	7
chr9	24,518,225	38,907,678	3
chr9	38,915,426	39,164,181	3
chr9	39,164,181	39,184,384	2
chr9	39,356,883	39,361,659	2
chr9	39,377,664	39,561,123	2
chr9	39,579,901	39,584,901	2
chr9	39,731,589	39,890,910	2
chr9	40,024,796	40,034,796	2
chr9	40,099,934	40,104,934	2
chr9	40,140,289	40,165,289	2
chr9	40,180,289	40,185,289	2
chr9	40,288,041	40,925,501	2
chr9	41,015,341	41,093,211	12
chr9	41,223,882	41,305,253	12
chr9	41,517,630	41,522,630	12
chr9	41,701,141	41,724,925	12
chr9	41,724,925	41,823,718	2
chr9	41,833,718	41,900,482	13
chr9	41,917,223	41,988,151	2
chr9	42,146,000	42,214,271	2
chr9	42,224,271	42,378,487	2
chr9	42,491,393	42,493,312	2
chr9	42,503,751	42,506,786	2
chr9	42,536,787	42,564,348	2
chr9	42,871,530	43,181,705	2
chr9	43,313,924	43,316,657	1
chr9	43,381,945	43,548,498	1
chr9	43,588,530	43,593,530	1
chr9	43,598,530	43,651,423	1
chr9	43,652,136	44,245,969	3
chr9	44,390,520	44,421,880	3
chr9	44,431,880	44,441,880	2
chr9	44,726,646	45,041,143	2
chr9	45,102,926	45,169,767	2
chr9	45,352,101	45,781,568	2
chr9	46,035,746	46,040,746	2
chr9	46,049,344	46,206,836	2
chr9	46,271,430	46,459,196	2
chr9	46,749,358	46,878,778	2
chr9	46,878,778	46,959,731	9
chr9	47,160,133	47,219,375	2
chr9	65,467,679	65,639,748	2
chr9	65,761,382	65,784,924	2
chr9	65,877,402	65,882,402	2
chr9	65,976,781	66,016,781	2
chr9	66,020,384	66,114,296	9
chr9	66,242,215	66,345,936	2
chr9	66,454,656	66,460,633	2
chr9	66,460,633	66,766,907	3
chr9	66,962,860	66,966,590	3
chr9	67,024,074	67,029,074	3
chr9	67,034,517	67,076,000	3
chr9	67,207,834	67,366,296	3
chr9	67,529,039	67,534,039	3
chr9	67,665,090	67,711,340	3
chr9	67,762,392	67,920,387	3
chr9	67,931,943	67,985,532	3
chr9	68,137,998	68,509,093	3
chr9	68,668,140	68,838,631	3
chr9	69,007,171	69,582,432	3
chr9	69,626,365	69,945,242	3
chr9	70,116,188	70,186,360	3
chr9	70,483,010	70,735,468	3
chr9	70,851,061	71,739,510	3
chr9	71,741,062	92,531,856	3
chr9	92,676,856	121,292,653	3
chr9	121,292,653	133,072,313	5
chr9	133,222,313	138,214,462	5
chr9	138,217,126	141,112,247	5
chr10	62,715	17,857,711	2
chr10	17,862,711	18,044,675	2
chr10	18,079,675	18,094,675	2
chr10	18,099,675	18,214,675	5
chr10	18,219,675	20,848,488	2
chr10	20,848,488	20,862,334	5
chr10	20,862,334	38,410,409	2
chr10	38,410,409	38,689,644	5
chr10	38,689,644	38,772,341	3
chr10	38,890,539	39,038,232	3
chr10	39,041,289	39,076,503	3
chr10	42,602,109	46,324,300	3
chr10	46,481,964	46,486,964	3
chr10	46,491,964	46,496,964	3
chr10	46,506,964	46,655,617	3
chr10	46,675,617	46,694,416	3
chr10	46,717,815	46,752,536	3
chr10	46,769,974	46,773,819	3
chr10	46,845,039	46,850,019	3
chr10	46,910,463	47,140,917	3
chr10	47,149,455	47,260,068	2
chr10	47,323,674	47,422,551	2
chr10	47,541,184	47,759,903	2
chr10	47,904,867	47,914,867	2
chr10	47,925,966	47,930,966	2
chr10	47,940,966	47,949,526	2
chr10	47,960,906	48,010,403	2
chr10	48,025,365	48,124,820	2
chr10	48,310,804	48,315,804	2
chr10	48,315,804	49,010,904	3
chr10	49,196,945	51,635,544	3
chr10	51,724,659	51,836,338	3
chr10	51,841,338	51,884,824	9
chr10	51,884,824	59,209,994	3
chr10	59,211,610	69,475,959	3
chr10	69,475,959	69,492,422	5
chr10	69,492,422	86,701,254	3
chr10	86,701,254	86,708,083	7
chr10	86,708,083	86,711,294	3
chr10	86,711,294	86,715,755	10
chr10	86,715,755	93,254,382	3
chr10	93,254,382	93,405,848	4
chr10	93,405,848	109,751,521	3
chr10	109,751,521	109,935,993	1
chr10	109,935,993	128,620,635	3
chr10	128,765,635	135,503,712	3
chr11	157,083	1,815,933	2
chr11	1,815,933	2,175,051	1
chr11	2,175,051	2,901,420	2
chr11	2,901,420	10,997,629	4
chr11	10,997,629	12,002,289	6
chr11	12,002,289	13,217,970	4
chr11	13,222,970	20,423,186	3
chr11	20,423,186	22,028,785	4
chr11	22,028,785	22,418,209	3
chr11	22,418,209	23,246,982	2
chr11	23,246,982	24,156,231	3
chr11	24,156,231	24,547,118	2
chr11	24,547,118	25,294,221	3
chr11	25,294,221	25,626,256	4
chr11	25,631,256	27,568,114	6
chr11	27,568,114	28,539,662	4
chr11	28,539,662	28,618,609	2
chr11	28,618,609	28,962,119	4
chr11	28,962,119	29,623,757	6
chr11	29,623,757	30,164,333	4
chr11	30,164,333	30,307,270	2
chr11	30,307,270	36,343,082	4

Chr	Start	End	CN
chr11	36,343,082	44,991,102	3
chr11	44,991,102	45,166,360	4
chr11	45,166,360	46,281,464	5
chr11	46,281,464	46,941,683	4
chr11	46,941,683	48,666,590	6
chr11	48,900,993	49,839,783	6
chr11	49,839,783	50,305,370	3
chr11	51,376,934	51,552,637	3
chr11	55,027,625	59,622,914	3
chr11	59,622,914	85,468,991	4
chr11	85,468,991	88,253,501	3
chr11	88,253,501	89,571,498	4
chr11	89,597,021	89,657,016	4
chr11	89,690,490	89,694,948	4
chr11	89,746,689	90,569,036	4
chr11	90,570,247	91,726,778	4
chr11	91,726,778	92,740,375	3
chr11	92,740,375	93,751,080	6
chr11	93,751,080	96,292,140	4
chr11	96,437,140	102,205,410	4
chr11	102,205,410	134,949,041	2
chr12	65,839	8,561,878	3
chr12	8,568,416	8,581,371	3
chr12	8,585,033	9,632,683	3
chr12	9,632,686	9,731,892	7
chr12	9,731,986	11,221,143	3
chr12	11,221,143	11,229,478	1
chr12	11,230,168	11,234,566	1
chr12	11,235,995	11,250,754	1
chr12	11,250,754	19,849,669	3
chr12	19,849,669	20,410,925	4
chr12	20,410,925	34,500,433	3
chr12	37,856,694	37,862,274	3
chr12	37,862,274	37,989,886	4
chr12	38,439,886	54,320,318	4
chr12	54,320,318	54,454,754	3
chr12	54,454,754	56,409,812	4
chr12	56,409,812	71,728,949	3
chr12	71,728,949	71,737,800	7
chr12	71,737,800	109,012,693	3
chr12	109,012,693	109,228,584	4
chr12	109,228,584	133,527,305	3
chr12	133,682,662	133,841,615	3
chr13	19,020,002	19,070,464	2
chr13	19,083,803	19,101,010	5
chr13	19,108,305	25,167,107	2
chr13	25,170,638	32,532,651	2
chr13	32,534,181	32,535,598	2
chr13	32,538,570	51,070,579	2
chr13	51,075,579	55,878,640	2
chr13	55,878,640	57,756,758	3
chr13	57,763,553	57,764,292	3
chr13	57,766,389	57,766,889	3
chr13	57,788,704	64,295,924	3
chr13	64,301,510	66,894,249	3
chr13	66,895,341	66,895,930	3
chr13	66,898,123	66,898,837	3
chr13	66,899,936	67,337,005	3
chr13	67,340,896	69,244,813	3
chr13	69,244,813	69,257,846	8
chr13	69,257,846	74,261,317	3
chr13	74,261,317	74,264,835	5
chr13	74,264,835	86,759,356	3
chr13	86,909,356	91,902,681	3
chr13	91,903,285	92,255,004	3
chr13	92,255,004	92,579,447	2
chr13	92,579,447	93,012,447	3
chr13	93,018,637	112,358,581	3
chr13	112,503,581	115,108,157	3
chr14	19,055,798	19,193,496	2
chr14	19,320,546	19,896,464	2
chr14	19,927,018	20,190,601	2
chr14	20,194,548	86,493,588	3
chr14	86,493,588	86,501,929	5
chr14	86,501,929	106,524,047	3
chr14	106,524,047	106,785,439	6
chr14	106,785,439	107,293,124	3
chr15	20,043,907	20,528,672	3
chr15	20,528,718	20,545,323	7
chr15	20,549,357	20,580,214	3
chr15	20,581,643	20,600,430	7
chr15	20,600,430	21,121,668	3
chr15	21,126,668	21,383,778	3
chr15	21,885,008	22,179,687	3
chr15	22,294,541	22,297,005	3
chr15	22,297,005	22,589,897	6
chr15	22,646,193	23,273,189	3
chr15	23,278,506	23,310,978	10
chr15	23,310,978	25,423,147	3
chr15	25,423,147	25,430,316	5
chr15	25,430,316	30,663,443	3
chr15	30,663,443	30,739,275	8
chr15	30,755,479	32,570,014	3
chr15	32,575,014	32,666,075	8
chr15	32,671,075	32,738,143	3
chr15	32,754,420	32,779,002	8
chr15	32,807,002	38,763,020	3
chr15	38,763,020	38,769,880	5
chr15	38,769,880	40,821,538	3
chr15	41,456,454	82,692,988	4
chr15	82,823,547	82,828,547	4
chr15	82,917,374	83,034,206	4
chr15	83,166,543	96,326,412	4
chr15	96,326,412	96,616,168	3
chr15	96,616,168	102,510,300	4
chr15	102,512,772	102,517,772	4
chr16	60,426	68,081	3
chr16	77,756	4,132,040	3
chr16	4,132,040	4,140,117	8
chr16	4,140,117	16,617,721	3
chr16	16,624,333	28,626,197	3
chr16	28,627,711	28,654,501	3
chr16	28,654,501	28,783,680	1
chr16	28,813,177	32,193,124	3
chr16	32,197,913	32,493,016	2
chr16	32,603,554	33,126,024	2
chr16	33,241,271	33,637,299	2
chr16	33,644,226	33,968,322	3
chr16	34,173,229	35,153,974	3
chr16	46,457,240	55,796,438	3
chr16	55,796,438	55,815,743	8
chr16	55,815,743	55,825,704	3
chr16	55,841,102	78,424,741	3
chr16	78,424,741	78,888,976	2
chr16	78,888,976	90,292,969	3
chr17	1	18,337,557	3
chr17	18,340,989	18,387,211	7
chr17	18,387,211	19,014,646	3
chr17	19,044,188	20,611,059	3
chr17	20,614,816	21,974,502	3
chr17	21,979,502	22,154,342	2
chr17	22,154,348	22,228,332	3
chr17	25,268,076	30,846,752	3
chr17	30,846,752	30,925,853	5
chr17	30,925,853	34,490,917	3
chr17	34,499,426	34,591,732	43
chr17	34,752,486	34,764,199	43
chr17	34,795,032	36,249,846	3
chr17	36,284,266	36,350,121	7
chr17	36,350,425	43,388,992	3
chr17	43,502,292	46,603,748	3
chr17	46,603,748	46,725,175	2
chr17	46,725,175	57,943,912	3
chr17	57,943,912	57,968,249	5
chr17	57,968,249	81,163,132	3
chr18	11,169	15,383,596	3
chr18	18,520,339	38,861,825	3
chr18	38,871,825	52,063,709	3
chr18	52,208,709	64,219,986	3
chr18	64,220,948	78,018,950	3
chr19	62,471	1,199,074	3
chr19	1,199,074	1,203,555	1
chr19	1,205,355	1,207,912	1
chr19	1,210,087	1,213,925	1
chr19	1,214,925	1,219,312	1
chr19	1,219,312	5,228,874	3
chr19	5,228,874	6,612,274	2
chr19	6,612,274	7,228,072	3
chr19	7,228,072	7,514,009	2
chr19	7,514,009	8,407,470	3
chr19	8,407,470	9,902,558	2
chr19	9,902,558	12,607,542	3
chr19	12,608,153	12,907,645	3
chr19	12,907,645	13,261,843	4
chr19	13,261,843	14,028,959	3
chr19	14,028,959	18,195,873	4
chr19	18,195,873	21,798,854	3
chr19	21,798,854	22,053,437	2
chr19	22,053,437	22,090,850	6
chr19	22,117,374	23,941,562	2
chr19	23,941,562	24,378,492	3
chr19	28,233,492	33,444,465	3
chr19	33,444,465	33,878,231	4
chr19	33,878,231	34,210,529	3
chr19	34,210,529	34,884,421	4
chr19	34,884,421	35,023,411	2
chr19	35,023,411	38,616,496	3
chr19	38,616,496	39,357,183	2
chr19	39,357,183	40,373,053	3
chr19	40,373,053	40,394,312	8
chr19	40,394,312	41,437,212	3
chr19	41,437,212	48,614,256	2
chr19	48,614,256	50,542,392	3
chr19	50,542,392	51,528,955	4
chr19	51,528,955	52,034,471	3
chr19	52,034,471	54,558,395	2
chr19	54,560,049	55,129,138	2
chr19	55,137,661	55,227,624	3
chr19	55,227,624	55,281,243	1
chr19	55,283,022	55,366,966	1
chr19	55,372,376	55,517,809	3
chr19	55,517,809	56,029,315	4
chr19	56,029,315	56,876,825	3
chr19	56,876,825	58,796,898	2
chr19	58,796,898	59,103,179	3
chr20	60,001	1,563,459	2
chr20	1,569,118	1,570,126	1
chr20	1,572,363	1,573,915	1
chr20	1,592,293	14,781,434	2
chr20	14,781,434	14,827,627	4
chr20	14,827,627	26,231,885	2
chr20	26,231,885	26,258,753	3
chr20	29,419,647	29,653,908	3
chr20	29,828,908	29,833,908	3
chr20	29,843,908	62,962,837	4
chr21	9,411,973	9,466,989	2
chr21	9,489,005	9,504,005	2
chr21	9,514,005	9,539,005	2
chr21	9,555,452	9,676,667	2
chr21	9,690,713	9,736,077	2
chr21	9,761,898	9,926,607	2
chr21	9,929,817	9,955,891	2
chr21	9,956,313	10,013,819	2
chr21	10,118,127	10,192,940	2
chr21	10,201,932	10,215,975	5
chr21	10,372,616	10,377,616	5
chr21	10,383,006	10,650,979	2
chr21	10,860,732	11,181,667	2
chr21	14,369,463	14,452,203	2
chr21	14,452,203	14,455,544	1
chr21	14,460,544	14,529,115	1
chr21	14,534,748	14,577,739	2
chr21	14,577,758	15,092,781	3
chr21	15,184,122	21,617,853	3
chr21	21,619,270	48,105,087	3
chr22	16,050,000	16,488,635	2
chr22	16,489,911	16,504,911	2
chr22	16,629,053	16,656,311	2
chr22	16,665,261	16,670,261	2
chr22	16,864,184	18,657,081	2
chr22	18,669,119	18,852,915	2
chr22	18,854,603	18,875,132	7
chr22	18,875,132	20,395,639	2
chr22	20,396,142	20,427,107	1
chr22	20,442,107	20,467,107	1
chr22	20,473,205	20,503,548	2
chr22	20,630,265	21,477,873	2
chr22	21,489,307	21,537,480	7
chr22	21,555,011	21,613,746	2
chr22	21,618,746	21,649,830	2
chr22	21,695,820	21,726,238	2
chr22	21,772,109	24,343,586	2
chr22	24,343,586	24,397,837	76
chr22	24,397,837	32,686,554	2
chr22	32,701,554	33,070,700	2
chr22	33,075,700	36,572,443	2
chr22	36,573,121	36,573,621	2
chr22	36,574,227	37,144,444	2
chr22	37,145,695	51,231,477	2
chrX	2,699,520	3,762,385	2

Chr	Start	End	CN
chrX	116,378,951	116,391,258	5
chrX	116,391,258	117,103,274	2
chrX	117,103,274	117,107,142	6
chrX	117,107,142	117,317,471	2
chrX	117,317,471	117,330,752	6
chrX	117,330,752	118,874,398	2
chrX	118,884,398	119,318,145	2
chrX	119,332,013	120,010,235	2
chrX	120,120,871	131,519,545	2
chrX	131,519,545	131,522,697	5
chrX	131,522,697	140,856,804	2
chrX	140,856,804	140,862,579	6
chrX	140,862,579	140,870,605	7
chrX	140,870,605	142,055,334	2
chrX	142,055,334	142,076,287	7
chrX	142,076,287	142,517,740	2
chrX	142,517,740	142,530,619	4
chrX	142,530,619	150,707,743	2
chrX	150,712,743	151,560,649	2
chrX	151,560,649	151,565,336	7
chrX	151,565,336	154,565,794	2
chrX	154,626,376	154,754,505	2
chrX	154,759,505	154,792,514	6
chrX	154,821,667	154,930,289	2

Chr	Start	End	CN
chrX	3,850,145	4,075,067	2
chrX	4,075,067	20,766,856	3
chrX	20,766,856	20,784,287	10
chrX	20,784,287	21,441,735	3
chrX	21,441,735	21,444,561	7
chrX	21,444,561	21,449,770	5
chrX	21,449,770	24,554,604	3
chrX	24,554,604	24,563,068	7
chrX	24,563,068	26,363,504	3
chrX	26,365,643	29,596,819	3
chrX	29,596,819	29,607,022	7
chrX	29,607,022	32,069,870	3
chrX	32,069,870	32,077,877	7
chrX	32,077,877	35,475,209	3
chrX	35,475,209	35,482,359	8
chrX	35,482,359	37,930,118	3
chrX	37,930,118	37,938,486	15
chrX	37,938,486	38,584,404	3
chrX	38,589,404	40,813,488	3
chrX	40,813,488	40,844,676	7
chrX	40,854,676	44,300,092	3
chrX	44,302,776	44,876,236	3
chrX	44,876,236	44,929,798	6
chrX	44,929,798	49,218,379	3
chrX	49,352,341	51,995,562	3
chrX	51,995,562	52,004,675	7
chrX	52,004,675	52,103,259	3
chrX	52,564,504	52,709,501	3
chrX	52,814,879	58,082,761	3
chrX	62,044,948	70,897,049	3
chrX	71,030,014	72,150,000	3
chrX	72,170,486	75,670,433	3
chrX	75,820,433	87,842,233	3
chrX	87,847,233	92,796,662	3
chrX	92,800,589	95,483,449	3
chrX	95,483,449	95,501,458	5
chrX	95,501,458	96,326,524	3
chrX	96,326,524	96,332,382	8
chrX	96,332,382	96,538,511	3
chrX	96,538,511	96,700,351	4
chrX	96,700,351	98,922,937	3
chrX	98,922,937	98,935,108	5
chrX	98,935,108	99,785,173	3
chrX	99,810,173	101,603,799	2
chrX	101,740,939	103,228,668	2
chrX	103,238,668	104,263,433	2
chrX	104,263,433	104,266,953	6
chrX	104,343,207	104,348,428	5
chrX	104,348,428	108,354,214	2
chrX	108,359,214	116,378,951	2

Table D.3.7. High-resolution copy number calls in HeLa CCL-2

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 D.3.8. Haplotype-Resolved Copy Number (HRCN) profile of HeLa CCL-2.
Proportion of the genome (UCSC hg19/GRCh37, excluding assembly gaps and segmental 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 D.3.9. Large rgions of LOH in HeLa CCL-2.

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr1	1219982	1223434	1225577	1229105	2143
chr1	9591985	9595341	9597227	9600089	1886
chr1	14432636	14436290	14437570	14441375	1280
chr1	25154228	25158676	25161514	25165093	2838
chr1	26456898	26459838	26463789	26468542	3951
chr1	34985513	34991181	34992539	34996862	1358
chr1	56827433	56831099	56834909	56838758	3810
chr1	63702441	63705334	63708226	63712091	2892
chr1	71234176	71237339	71239506	71243649	2167
chr1	73450444	73454045	73454999	73458655	954
chr1	81400468	81404472	81408850	81413863	4378
chr1	83465559	83469370	83477194	83482174	7824
chr1	84513313	84517907	84524680	84528485	6773
chr1	84708940	84711986	84715625	84719559	3639
chr1	97657769	97660911	97669287	97672164	8376
chr1	106011959	106016763	106023133	106026856	6370
chr1	107217125	107222220	107223107	107229248	887
chr1	108730241	108733240	108737139	108741157	3899
chr1	110183007	110188076	110190877	110195347	2801
chr1	113215882	113221119	113224360	113228986	3241
chr1	116224845	116229439	116232523	116237054	3084
chr1	145087506	145093102	145096826	145101547	3724
chr1	158862935	158867476	158869332	158873313	1856
chr1	179325893	179330458	179332696	179337887	2238
chr1	179450982	179455462	179457941	179461858	2479
chr1	184810138	184814613	184820856	184824172	6243
chr1	197496174	197500497	197503019	197506713	2522
chr1	206196105	206201100	206204558	206209774	3458
chr1	207538358	207541947	207544901	207549865	2954
chr1	210073183	210076156	210083759	210086702	7603
chr1	229808620	229812448	229820888	229823967	8440
chr1	232454490	232458021	232459720	232463478	1699
chr1	238479704	238483552	238485664	238489584	2112
chr1	248048061	248051468	248057471	248061344	6003
chr2	1097647	1101967	1104548	1107542	2581
chr2	1214589	1218811	1227052	1231902	8241
chr2	1341926	1346043	1348762	1351743	2719
chr2	4777856	4781284	4787289	4790798	6005
chr2	11391361	11394666	11399696	11403933	5030
chr2	16267901	16271547	16274058	16278112	2511
chr2	33088738	33091602	33095837	33098992	4235
chr2	36212682	36216149	36219159	36222182	3010
chr2	40761679	40764780	40772187	40776152	7407
chr2	76522203	76526482	76528580	76533319	2098
chr2	79072942	79076298	79080535	79084314	4237
chr2	79329560	79332427	79340085	79343563	7658
chr2	89024107	89029186	89032250	89036680	3064
chr2	106876453	106879315	106885950	106888812	6635
chr2	107895379	107899331	107902985	107907350	3654
chr2	108272240	108275202	108278766	108282045	3564

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr2	119649907	119653243	119659341	119663593	6098
chr2	126439903	126442995	126451374	126455168	8379
chr2	130432522	130437800	130444772	130449065	6972
chr2	130545696	130549556	130558302	130562716	8746
chr2	134962174	134966692	134969953	134975262	3261
chr2	159955084	159959408	159961184	159964933	1776
chr2	162329257	162332119	162334457	162337434	2338
chr2	170636358	170640750	170643778	170646698	3028
chr2	177262743	177265605	177272587	177275449	6982
chr2	202142972	202146622	202149209	202152734	2587
chr2	209939026	209943299	209948576	209951775	5277
chr2	217085890	217089914	217092137	217096117	2223
chr2	220048764	220052266	220056195	220059926	3929
chr2	223756860	223760158	223762503	223766340	2345
chr2	227160464	227165399	227171059	227173971	5660
chr2	232393348	232397193	232400141	232403838	2948
chr2	235551147	235554212	235557680	235560808	3468
chr2	238447706	238451483	238453865	238457095	2382
chr2	242741860	242744722	242748878	242751740	4156
chr3	4062722	4066883	4069516	4072897	2633
chr3	17537316	17540852	17545150	17548889	4298
chr3	22088622	22092324	22097412	22101461	5088
chr3	25969591	25973347	25978418	25982836	5071
chr3	26436264	26439477	26445541	26449165	6064
chr3	42834857	42838188	42840278	42844399	2090
chr3	58531873	58534735	58537295	58540157	2560
chr3	67490793	67493658	67496840	67499988	3182
chr3	83849460	83853088	83855293	83858701	2205
chr3	93665268	93669340	93670413	93673763	1073
chr3	98896090	98898968	98902468	98905388	3500
chr3	98940888	98943750	98949179	98952041	5429
chr3	107034093	107037882	107040259	107043431	2377
chr3	120157900	120161334	120164909	120168300	3575
chr3	128669066	128672057	128674484	128678337	2427
chr3	129072955	129075817	129082766	129085628	6949
chr3	130344301	130347557	130353005	130356623	5448
chr3	136016782	136020961	136025873	136029914	4912
chr3	146380894	146385163	146390166	146395114	5003
chr3	160720549	160723432	160724021	160728820	589
chr3	162760638	162764845	162769155	162773222	4310
chr3	163655020	163658805	163664987	163669216	6182
chr3	177290212	177294442	177297279	177300567	2837
chr3	186577626	186581002	186585067	186588403	4065
chr3	187729174	187732036	187734734	187737784	2698
chr3	192871950	192875296	192885136	192890631	9840
chr3	194395510	194398372	194400805	194403802	2433
chr4	3609034	3615017	3618599	3622303	3582
chr4	10388516	10392404	10401988	10404957	9584
chr4	16941202	16945672	16950703	16954250	5031
chr4	19075217	19079377	19085506	19089076	6129

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr4	21155964	21160900	21166624	21170601	5724
chr4	42758944	42762416	42769284	42772785	6868
chr4	59941173	59944496	59950359	59953869	5863
chr4	61325715	61331232	61333288	61337393	2056
chr4	73549356	73552218	73561379	73564878	9161
chr4	79266181	79269090	79275358	79278638	6268
chr4	80883345	80888023	80893952	80897541	5929
chr4	83698257	83701956	83704282	83708226	2326
chr4	88479734	88483755	88486261	88490292	2506
chr4	91593326	91596775	91602687	91606898	5912
chr4	94555083	94560418	94565089	94568899	4671
chr4	108124800	108127787	108131477	108135185	3690
chr4	115171374	115174662	115182912	115186293	8250
chr4	115924550	115928682	115931788	115935113	3106
chr4	119122241	119125371	119129611	119134289	4240
chr4	130062111	130065916	130068324	130071292	2408
chr4	135429353	135433008	135435646	135438845	2638
chr4	163389791	163392653	163395118	163397980	2465
chr4	165999409	166002454	166004638	166008852	2184
chr4	171174890	171177752	171180241	171183103	2489
chr4	172370862	172374373	172379378	172383665	5005
chr4	172984775	172989093	172992341	172995638	3248
chr4	174529208	174532070	174535491	174538353	3421
chr4	184652174	184655728	184657947	184661728	2219
chr4	186437444	186441598	186443309	186447300	1711
chr4	187090227	187093089	187097953	187101137	4864
chr4	188342968	188345863	188347481	188351509	1618
chr4	190053655	190057911	190063600	190066849	5689
chr4	190601603	190606920	190610201	190614584	3281
chr5	1964353	1968378	1971566	1977522	3188
chr5	12808034	12810896	12820756	12824573	9860
chr5	18452208	18456085	18459830	18464834	3745
chr5	18868972	18873112	18874750	18879547	1638
chr5	21606857	21611796	21614749	21619556	2953
chr5	24366970	24370519	24373378	24376908	2859
chr5	29706095	29709904	29711508	29715584	1604
chr5	34300578	34304708	34307834	34311977	3126
chr5	57676480	57679976	57686094	57690155	6118
chr5	59944841	59947703	59952238	59955100	4535
chr5	61804524	61807544	61810268	61813744	2724
chr5	63695220	63698297	63701125	63704390	2828
chr5	83944725	83947913	83955395	83959656	7482
chr5	103850608	103854256	103860183	103864439	5927
chr5	108591764	108595067	108601043	108604636	5976
chr5	114322312	114325286	114335069	114338724	9783
chr5	120645717	120649524	120651986	120656240	2462
chr5	132914272	132918941	132924153	132928517	5212
chr5	136850836	136855041	136856214	136859445	1173
chr5	147549442	147552304	147554462	147557580	2158
chr5	151451430	151456401	151461965	151467110	5564

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr5	174126397	174129667	174132180	174135042	2513
chr5	176383400	176387450	176390002	176393920	2552
chr5	177818326	177821621	177823817	177827476	2196
chr6	19761376	19765096	19770947	19774217	5851
chr6	29681178	29685415	29688062	29691865	2647
chr6	33934874	33937783	33942660	33946136	4877
chr6	54658958	54662425	54665884	54668746	3459
chr6	56755301	56758261	56760886	56764618	2625
chr6	57281043	57286376	57289240	57299261	2864
chr6	57417631	57423054	57428871	57434459	5817
chr6	70010354	70014091	70018028	70023018	3937
chr6	81622752	81625911	81630504	81633366	4593
chr6	85314645	85318086	85323815	85327744	5729
chr6	94536687	94539866	94544939	94548748	5073
chr6	110473032	110476699	110479698	110484106	2999
chr6	112221307	112224184	112229905	112233885	5721
chr6	119161560	119165019	119172467	119176620	7448
chr6	125704314	125708707	125711004	125714486	2297
chr6	126179223	126183221	126187093	126191334	3872
chr6	129315217	129319424	129325491	129328684	6067
chr6	133337915	133341795	133347710	133350963	5915
chr6	153026313	153029695	153033276	153036638	3581
chr6	153954736	153957939	153960993	153964090	3054
chr6	164540432	164543306	164547190	164551036	3884
chr6	165720010	165724618	165732147	165737402	7529
chr7	4412772	4416537	4419798	4423386	3261
chr7	25053948	25057329	25059883	25064862	2554
chr7	32386359	32390021	32392974	32397656	2953
chr7	49716462	49719822	49725861	49730062	6039
chr7	51591196	51594222	51598615	51601577	4393
chr7	55283365	55286622	55288797	55292643	2175
chr7	64311223	64315743	64317818	64321662	2075
chr7	73824236	73828408	73830500	73834295	2092
chr7	82702779	82705905	82709158	82712806	3253
chr7	84154057	84156919	84157308	84160170	389
chr7	93413023	93416917	93422950	93427712	6033
chr7	96472104	96475869	96481971	96486191	6102
chr7	100998223	101001128	101003833	101007745	2705
chr7	110177718	110181937	110186691	110191834	4754
chr7	113413233	113416122	113422237	113425982	6115
chr7	118827732	118831226	118834735	118838701	3509
chr7	126041620	126045845	126050534	126054644	4689
chr7	129296863	129299725	129302048	129305743	2323
chr7	148069310	148072808	148076022	148079779	3214
chr7	158497185	158500690	158504533	158508133	3843
chr8	590749	594379	599355	603532	4976
chr8	960050	963586	967640	971642	4054
chr8	11489120	11492924	11495214	11498329	2290
chr8	16197004	16201246	16207422	16211454	6176
chr8	25063428	25068680	25069734	25073623	1054

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr8	37385095	37389594	37395165	37399765	5571
chr8	40771502	40774384	40779321	40782398	4937
chr8	42187032	42190226	42194190	42197389	3964
chr8	63031034	63034929	63040195	63044278	5266
chr8	68153164	68156524	68161879	68165119	5355
chr8	72211728	72216132	72217805	72221536	1673
chr8	73784191	73787709	73793865	73797618	6156
chr8	96871796	96875167	96879800	96882884	4633
chr8	124868793	124871757	124877828	124881223	6071
chr8	126590077	126595099	126601036	126604775	5937
chr8	129460647	129465135	129471211	129475398	6076
chr8	130136975	130140897	130145096	130149468	4199
chr8	131197197	131201216	131210164	131214235	8948
chr8	135079988	135082868	135088694	135092827	5826
chr8	137155470	137160159	137163801	137166868	3642
chr8	140252931	140256694	140258784	140261842	2090
chr8	144631089	144633951	144636568	144639430	2617
chr9	11206454	11210588	11215783	11220964	5195
chr9	12553691	12556788	12559609	12563421	2821
chr9	71735066	71738808	71743435	71747247	4627
chr9	78915817	78919396	78921842	78925556	2446
chr9	84320336	84324318	84326926	84329880	2608
chr9	110030144	110033227	110035423	110039462	2196
chr9	110533733	110537352	110540349	110544682	2997
chr9	110655176	110658038	110662636	110665975	4598
chr9	130029812	130033075	130035087	130038732	2012
chr9	130178578	130181748	130186149	130189179	4401
chr9	138210732	138214044	138217637	138223194	3593
chr9	140249678	140253150	140255891	140260453	2741
chr9	140769763	140774831	140776787	140781361	1956
chr10	1738044	1740906	1743233	1746095	2327
chr10	5886388	5889250	5892561	5895423	3311
chr10	6407764	6411429	6417250	6420683	5821
chr10	26995815	26998677	27002113	27004975	3436
chr10	38789414	38795054	38795690	38798552	636
chr10	39120715	39123577	39126206	39129068	2629
chr10	42527061	42531930	42532103	42536455	173
chr10	54012016	54015508	54017817	54021852	2309
chr10	55079052	55082142	55090229	55093495	8087
chr10	94131212	94134575	94137687	94140596	3112
chr10	99031606	99034833	99036936	99041673	2103
chr10	102351438	102354381	102361681	102365007	7300
chr10	106101204	106105530	106106748	106110569	1218
chr10	111568568	111572041	111578149	111581553	6108
chr10	114109068	114113550	114116185	114120321	2635
chr10	132985112	132988406	132990664	132994838	2258
chr10	134945906	134948856	134952221	134955205	3365
chr10	135287172	135290042	135292769	135296621	2727
chr11	1651318	1654180	1656419	1659775	2239
chr11	3235138	3239843	3243981	3248044	4138

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr11	4963794	4969340	4975118	4980155	5778
chr11	5870561	5873812	5883130	5886408	9318
chr11	29003653	29007043	29012492	29015900	5449
chr11	45189321	45193409	45197529	45202241	4120
chr11	47948075	47952718	47961833	47967336	9115
chr11	48363049	48367484	48373427	48376696	5943
chr11	48597949	48600818	48603687	48607581	2869
chr11	60683731	60688496	60696165	60700201	7669
chr11	60847945	60851764	60852973	60857721	1209
chr11	61940706	61943860	61953525	61956387	9665
chr11	62386981	62390413	62391975	62395874	1562
chr11	65930191	65933592	65939273	65942184	5681
chr11	90564121	90567866	90570955	90574529	3089
chr11	90754571	90758723	90762388	90766313	3665
chr11	92866072	92869774	92875728	92879344	5954
chr11	105289901	105293446	105298911	105302354	5465
chr11	114421196	114424162	114430417	114435034	6255
chr12	14210042	14213704	14216250	14219286	2546
chr12	31948887	31951749	31954578	31957440	2829
chr12	40895986	40899172	40905755	40908617	6583
chr12	42694320	42697494	42699407	42703523	1913
chr12	43973586	43977274	43979425	43982846	2151
chr12	53372356	53375563	53377667	53382430	2104
chr12	60749838	60753056	60757766	60761357	4710
chr12	70590726	70594709	70597310	70601286	2601
chr12	80154645	80157975	80159284	80163038	1309
chr12	96230306	96233568	96236300	96239956	2732
chr12	96335769	96340325	96342567	96346494	2242
chr12	125096520	125100077	125103553	125107106	3476
chr12	130336701	130339565	130342896	130345811	3331
chr13	32529212	32532511	32538401	32541263	5890
chr13	51066023	51069000	51074302	51077570	5302
chr13	66890009	66893486	66898191	66902976	4705
chr13	67332822	67335994	67341384	67345605	5390
chr13	70121550	70127032	70127746	70131942	714
chr13	72804623	72807485	72812415	72815277	4930
chr13	93005230	93009159	93018571	93022150	9412
chr13	113258980	113263099	113267897	113271835	4798
chr13	114846089	114848975	114852049	114855115	3074
chr14	23090931	23094822	23098199	23101877	3377
chr14	28865638	28868500	28871065	28873927	2565
chr14	35300560	35304095	35306749	35310016	2654
chr14	40605658	40609724	40617207	40621353	7483
chr14	45145804	45150241	45150983	45153845	742
chr14	65356215	65360233	65362928	65366678	2695
chr14	85293392	85296483	85300974	85304137	4491
chr14	107082421	107089780	107096156	107101361	6376
chr15	22333139	22336731	22343873	22346735	7142
chr15	55215033	55218012	55224187	55227360	6175
chr15	65811756	65815996	65818685	65822402	2689

Deletions						
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)	
chr15	68421997	68425988	68428866	68432392	2878	
chr15	71018514	71021790	71027370	71030916	5580	
chr15	72382675	72385927	72388119	72391177	2192	
chr15	72568386	72571753	72575386	72579157	3633	
chr15	82392972	82396611	82401629	82404932	5018	
chr15	86052677	86056986	86058863	86062283	1877	
chr15	89481083	89483945	89486536	89489398	2591	
chr15	98529175	98532455	98536201	98540010	3746	
chr15	101091519	101096432	101098048	101102625	1616	
chr16	2577003	2581082	2583485	2587100	2403	
chr16	5423016	5426814	5428971	5432575	2157	
chr16	16930590	16934353	16940338	16944506	5985	
chr16	18829325	18833363	18838259	18842053	4896	
chr16	25336572	25340082	25342763	25345812	2681	
chr17	133109	136632	137505	141889	873	
chr17	1638040	1642119	1644324	1647840	2205	
chr17	15784780	15789905	15793225	15796747	3320	
chr17	21209111	21212881	21221154	21224956	8273	
chr17	25532745	25536493	25540200	25543062	3707	
chr17	41433726	41437789	41439183	41443554	1394	
chr17	61987062	61991907	61992175	61996921	268	
chr17	74358659	74362597	74364647	74368977	2050	
chr17	74783545	74786407	74790556	74793418	4149	
chr17	77988865	77991991	78000543	78004820	8552	
chr18	18511313	18514175	18514606	18520184	431	
chr18	23744557	23747743	23751607	23755400	3864	
chr18	28766762	28771283	28775833	28778985	4550	
chr18	47691169	47694933	47696990	47702476	2057	
chr18	49192416	49196558	49198912	49202473	2354	
chr18	51949005	51952873	51956960	51960826	4087	
chr18	63330779	63334307	63340253	63343151	5946	
chr18	63632474	63635437	63640109	63644928	4672	
chr18	63763561	63766790	63768932	63772464	2142	
chr18	65384504	65389148	65395773	65399696	6625	
chr18	76126587	76129723	76132332	76135376	2609	
chr18	77306394	77309838	77312090	77315980	2252	
chr19	6854816	6858402	6861194	6864665	2792	
chr19	10610090	10612952	10619988	10622850	7036	
chr19	12604002	12606896	12611359	12614251	4463	
chr19	16100072	16103130	16106143	16109481	3013	
chr19	29946128	29950164	29955441	29959454	5277	
chr19	30384161	30388736	30393054	30396813	4318	
chr19	36954780	36957661	36960301	36963428	2640	
chr19	51327437	51330388	51333808	51337493	3420	
chr19	53685399	53689152	53691864	53696139	2712	
chr19	54552136	54554998	54561022	54563884	6024	
chr20	2800125	2803124	2806433	2810270	3309	
chr20	7092919	7096580	7102260	7105990	5680	
chr20	13813627	13817220	13822767	13825693	5547	
chr20	32812906	32815842	32819269	32822254	3427	

Deletions					
Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr20	42267303	42271498	42273921	42277745	2423
chr20	54430683	54434533	54440590	54444187	6057
chr20	55100447	55103457	55105825	55108797	2368
chr21	11086192	11089780	11092602	11096079	2822
chr21	22001696	22005523	22010528	22013668	5005
chr21	26859173	26862280	26865644	26869506	3364
chr21	44967071	44970317	44972908	44977498	2591
chr22	18053950	18057628	18060280	18064827	2652
chr22	24191796	24195520	24197910	24200982	2390
chr22	36568300	36571774	36575439	36579177	3665
chr22	37140005	37143031	37147941	37151550	4910
chrX	1752882	1755750	1760166	1763577	4416
chrX	2191898	2195084	2200086	2203381	5002
chrX	10524701	10528605	10530743	10534521	2138
chrX	11720569	11725339	11731046	11734365	5707
chrX	26359541	26363107	26366052	26369692	2945
chrX	32983817	32986679	32988932	32992432	2253
chrX	38580371	38584330	38587444	38591335	3114
chrX	43941575	43944781	43947196	43950879	2415
chrX	70133984	70137078	70139934	70143743	2856
chrX	81092002	81096623	81101754	81106339	5131
chrX	85601335	85604868	85609650	85613695	4782
chrX	92792544	92796248	92801304	92805182	5056
chrX	118870544	118873929	118876324	118880223	2395
chrX	140473801	140476676	140479553	140483125	2877
chrX	144418334	144422010	144424420	144428180	2410
chrX	150704050	150706972	150712230	150715382	5258

Interchromosomal

BP1 Chr	BP1 Start	BP1 End	BP2 Chr	BP2 Start	BP2 End	Supp.
chr2	3930448	3935237	chr12	124494444	124,500,974	100
chr2	11948918	11954246	chr21	18143482	18,149,316	104
chr3	73158440	73162742	chr17	41379324	41384937	80
chr3	73158440	73162742	chr17	41397704	41,403,894	132
chr3	75985574	75999286	chr20	26198677	26,213,907	271
chr3	97545352	97548214	chr20	51820873	51826137	54
chr3	111271573	111275045	chr8	128529232	128,538,779	138
chr5	45603558	45620710	chr22	39709007	39715736	64
chr7	81789161	81794473	chr10	60900343	60904922	66
chr7	100936524	100940377	chr11	114417947	114425186	45
chr7	111050826	111056166	chr12	108199590	108203183	55
chr8	28431852	28436443	chr17	79758758	79762966	51
chr9	12624450	12627475	chr22	29065133	29067995	11
chr9	96381359	96386424	chr13	19645775	19650874	29
chr9	101473886	101479887	chr22	32575696	32579748	67
chr9	109772276	109777707	chr20	29873930	29878940	70
chr9	121298728	121305690	chr11	11980608	12002075	178
chr12	56987510	56991801	chr15	39995972	39999958	25
chr13	55877158	55882393	chr19	12896441	12904258	72
chr13	63618669	63624153	chr17	21666588	21670266	115

Inversions

Chr	BP1 Start	BP1 End	BP2 Start	BP2 End	Size (bp)
chr2	117,788,943	117,791,805	117,795,872	117,800,947	4,067
chr4	88,847,496	88,850,544	88,856,950	88,862,945	6,406
chr4	188,866,434	188,869,296	188,873,075	188,878,354	3,779
chr5	6,344,556	6,351,102	6,352,996	6,358,295	1,894
chr7	112,769,507	112,772,567	112,774,843	112,778,107	2,276
chr10	37,391,949	37,395,530	37,400,024	37,403,895	4,494
chr11	307,347	310,209	316,100	321,191	5,891
chr22	39,608,077	39,612,118	39,614,886	39,618,298	2,768

Table D.3.10. Structural Rearrangements in heLa CCL-2

Chr	Start	End	Type	GenBank ID	Gene Name	ENSEMBL ID	Description	In CGC?	Hom/Het	Repeat	Exonic?	HGDP
chr1	1223434	1225577	DEL,60	NM_006625	FUSIP1	ENSG00000188529|ENSG00000215699	FUS interacting protein (serine/arginine-rich) 1	no	Hemi	Simple	No	Yes
chr1	106016763	106023133	DEL,76	NM_001005221	OR4F29	ENSG00000177799|ENSG00000183127|ENSG00000185097	Olfactory receptor, family 4, subfamily F, member 29	no	Hemi	None	No	Yes
chr1	108733240	108737139	DEL,53	NM_013386	SLC25A24	ENSG00000085491	Solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 24	no	Hemi	None	Yes	Yes
chr1	116229439	116232523	DEL,117	NM_138959	VANGL1	ENSG00000173218	Vang-like 1 (van gogh, Drosophila)	no	Hemi	None	No	Yes
chr1	145093102	145096826	DEL,405	NM_004892	SEC22B	N/A	SEC22 vesicle trafficking protein homolog B (S. cerevisiae)	no	Hemi	None	Yes	Yes
chr1	145093102	145096826	DEL,405	NM_178230	PPIAL4	ENSG00000198161|ENSG00000198360|ENSG00000198936	Peptidylprolyl isomerase A (cyclophilin A)-like 4	no				
chr1	152555480	152590169	AMP_RD	NM_178433	LCE3B	ENSG00000187238	Late cornified envelope 3B	no	Amp	None	Yes	No
chr1	152555480	152590169	AMP_RD	NM_178434	LCE3C	ENSG00000187225	Late cornified envelope 3C	no				
chr1	155182045	155190755	AMP_RD	NM_002455	MTX1	ENSG00000164418	Metaxin 1	no	Amp	SegDup	Yes	Yes
chr1	179455462	179457941	DEL,75	NM_144696	C1orf125	ENSG00000162779	Chromosome 1 open reading frame 125	no	Hemi	None	No	No
chr1	184814613	184820856	DEL,93	NM_052966	FAM129A	ENSG00000135842	Family with sequence similarity 129, member A	no	Hemi	None	No	Yes
chr2	1101967	1104548	DEL,79	NM_018968	SNTG2	ENSG00000172554	Syntrophin, gamma 2	no	Hemi	SegDup	No	Yes
chr2	11948918	11954246	chr21:18143482-18149316,104	NM_145693	LPIN1	ENSG00000134324	Lipin 1	no	Inter	None	No	N/a
chr2	183090568	183091068	DEL_RD	NM_001003683	PDE1A	ENSG00000115252	Phosphodiesterase 1A, calmodulin-dependent	no	Hom	None	No	No
chr2	202146622	202149209	DEL,140	NM_001080124	CASP8	ENSG00000064012	Caspase 8, apoptosis-related cysteine peptidase	no	Hemi	Yes	No	Yes
chr3	17540852	17545150	DEL,108	NM_014744	TBC1D5	ENSG00000131374	TBC1 domain family, member 5	no	Hemi	None	No	No
chr3	60555218	60658452	DEL,53	NM_002012	FHIT	ENSG00000189283	Fragile histidine triad gene	YES	Hom	None	No	No
chr3	97545352	97548214	chr20:51820873-51826137,54	NM_153605	DKFZp667G2110	ENSG00000080200	Hypothetical protein DKFZp667G2110	no	Inter	None	No	N/A
chr3	120161334	120164909	DEL,76	NM_007085	FSTL1	ENSG00000163430	Follistatin-like 1	no	Hom	None	No	No
chr3	136020961	136025873	DEL,87	NM_000532	PCCB	N/A	Propionyl Coenzyme A carboxylase, beta polypeptide	no	Hom	None	No	Yes
chr4	115174662	115182912	DEL,101	NM_001005217	FRG2	ENSG00000205097	FSHD region gene 2	no	Hemi	None	No	Yes
chr4	115174662	115182912	DEL,101	NM_033178	DUX4	ENSG00000178067	Double homeobox, 4	YES				
chr4	115928682	115931788	DEL,79	NM_022569	NDST4	ENSG00000138653	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 4	no	Hemi	Yes	No	Yes
chr4	119125371	119129611	DEL,66	NM_004784	NDST3	ENSG00000164100	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 3	no	Hemi	None	No	Yes
chr4	166002454	166004638	DEL,54	NM_001100389	TMEM192	ENSG00000170088	Transmembrane protein 192	no	Hom	None	No	Yes
chr4	186441598	186443309	DEL,95	NM_014476	PDLIM3	ENSG00000154553	PDZ and LIM domain 3	no	Hom	None	No	Yes
chr5	37670470	37692616	AMP_RD	NM_018034	WDR70	ENSG00000082068	WD repeat domain 70	no	Amp	Yes	No	N/A
chr5	45603558	45620710	chr22:39709007-39715736,64	NM_021072	HCN1	ENSG00000164588	Hyperpolarization activated cyclic nucleotide-gated potassium channel 1	no	Inter	Yes	No	N/A
chr5	147553050	147553908	DEL_RD	NM_001001325	SPINK5L2	ENSG00000196800	Kazal type serine protease inhibitor 5-like 2	no	Hom	None	Yes	Yes
chr5	176387450	176390002	DEL,133	NM_016290	UIMC1	ENSG00000087206	Ubiquitin interaction motif containing 1	no	Hemi	Yes	No	Yes
chr6	57286376	57289240	DEL,118	NM_000947	PRIM2	ENSG00000146143	Primase, DNA, polypeptide 2 (58kDa)	no	Hemi	None	No	Yes
chr6	70014091	70018028	DEL,178	NM_001704	BAI3	ENSG00000135298	Brain-specific angiogenesis inhibitor 3	no	Hom	None	No	No
chr6	126183221	126187093	DEL,97	NM_181782	NCOA7	ENSG00000111912	Nuclear receptor coactivator 7	no	Hom	None	No	Yes
chr7	81789161	81794473	chr10:60900343-60904922,66	NM_000722	CACNA2D1	ENSG00000153956	Calcium channel, voltage-dependent, alpha 2/delta subunit 1	no	Inter	None	No	N/A
chr7	111050826	111056166	chr12:108199590-108203183,55	NM_032549	IMMP2L	ENSG00000184903	IMP2 inner mitochondrial membrane peptidase-like (S. cerevisiae)	no	Inter	None	No	N/A
chr7	143309234	143402897	AMP_RD	NM_173678	FAM139A	ENSG00000159860|ENSG00000170379	Family with sequence similarity 139, member A	no	Amp	SegDup	No	Yes
chr7	148072808	148076022	DEL,72	NM_014141	CNTNAP2	ENSG00000174469	Contactin associated protein-like 2	no	Hemi	None	No	Yes
chr8	25066854	25070450	DEL_RD	NM_024940	DOCK5	ENSG00000147459	Dedicator of cytokinesis 5	no	Hom	None	No	Yes
chr8	68156524	68161879	DEL,62	NM_006421	ARFGEF1	ENSG00000066777	ADP-ribosylation factor guanine nucleotide-exchange factor 1(brefeldin A-inhibited)	no	Hemi	None	No	Yes
chr8	131201216	131210164	DEL,141	NM_018482	DDEF1	ENSG00000153317	Development and differentiation enhancing factor 1	no	Hemi	None	No	No
chr9	96381359	96386424	chr13:19645775-19650874,29	NM_005392	PHF2	ENSG00000197724	PHD finger protein 2	no	Inter	Near SegDup	No	N/A
chr9	109772276	109777707	chr20:29873930-29878940,70	NM_021224	ZNF462	ENSG00000148143	Zinc finger protein 462	no	Inter	None	Yes	N/A
chr9	140253150	140255891	DEL,253	NM_017820	FLJ20433	ENSG00000187609	Hypothetical protein FLJ20433	no	Hom	None	No	Yes
chr10	54015508	54017817	DEL,71	NM_001098512	PRKG1	ENSG00000185532	Protein kinase, cGMP-dependent, type I	no	Hemi	None	No	Yes
chr10	99034833	99036936	DEL,83	NM_032900	ARHGAP19	N/A	Rho GTPase activating protein 19	no	Hom	Yes	No	Yes
chr11	11980608	12002075	chr9:121298728-121305690,178	NM_001018057	DKK3	ENSG00000050165	Dickkopf homolog 3 (Xenopus laevis)	no	Inter	None	No	N/A
chr11	11980608	12002075	chr9:121298728-121305690,178	NM_017944	USP47	ENSG00000170242	Ubiquitin specific peptidase 47	no				
chr11	48367484	48373427	DEL,52	NM_001005513	OR4C45	ENSG00000197161	Olfactory receptor, family 4, subfamily C, member 45	no	Hemi	Yes	No	Yes
chr11	65933592	65939273	DEL,209	NM_018026	PACS1	ENSG00000175115	Phosphofurin acidic cluster sorting protein 1	no	Hom	None	No	No
chr11	114417947	114425186	chr7:100936524-100940377,45	NM_152315	FAM55D	N/A	Family with sequence similarity 55, member D	no	Inter	None	No	N/A

Chr	Start	End	Type	GenBank ID	Gene Name	ENSEMBL ID	Description	In CGC?	Hom/Het	Repeat	Exonic?	HGDP
chr12	56987510	56991801	chr15:39995972-39999958,25	NM_002898	RBMS2	ENSG00000076067	RNA binding motif, single stranded interacting protein 2	no	Inter	None	Yes	N/A
chr12	56987510	56991801	chr15:39995972-39999958,25	NM_013449	BAZ2A	ENSG00000076108	Bromodomain adjacent to zinc finger domain, 2A	no				
chr12	96340325	96342567	DEL,70	NM_152435	AMDHD1	ENSG00000139344	Amidohydrolase domain containing 1	no	Hemi	None	No	Yes
chr13	66893486	66898191	DEL,51	NM_020403	PCDH9	ENSG00000005073	Protocadherin 9	no	Hom	None	No	No
chr13	93009159	93018571	DEL,66	NM_004466	GPC5	ENSG00000179399	Glypican 5	no	Hom	None	No	No
chr13	114848975	114852049	DEL,52	NM_007368	RASA3	ENSG00000185989	RAS p21 protein activator 3	no	Hemi	None	No	Yes
chr15	39995972	39999958	chr12:56987510-56991801,25	NM_152597	FSIP1	ENSG00000150667	Fibrous sheath interacting protein 1	no	Inter	None	No	N/A
chr15	68425988	68428866	DEL,224	NM_016166	PIAS1	ENSG00000033800|ENSG00000070756	Protein inhibitor of activated STAT, 1	no	Hom	Yes	No	Yes
chr15	86056986	86058863	DEL,51	NM_006738	AKAP13	ENSG00000170776	A kinase (PRKA) anchor protein 13	no	Hemi	None	No	Yes
chr16	78371592	78385046	DEL,138	NM_016373	WWOX	ENSG00000186153	WW domain containing oxidoreductase	no	Hemi	None	No	Yes
chr17	79758758	79762966	chr8:28431852-28436443,51	NM_000160	GCGR	ENSG00000215644	Glucagon receptor	no	Inter	None	No	N/A
chr18	23747743	23751607	DEL,114	NM_001025096	PSMA8	N/A	Proteasome (prosome, macropain) subunit, alpha type, 8	no	hemi	SegDup	No	Yes
chr18	47694933	47696990	DEL,79	NM_001080467	ACAA2	ENSG00000167315	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)	no	Hemi	None	No	Yes
chr18	64219986	64220948	DEL_RD	NM_021153	CDH19	ENSG00000071991	Cadherin 19, type 2	no	Hom	None	No	Yes
chr19	1194824	1219594	DEL,88	NM_000455	STK11	N/A	Serine/threonine kinase 11	YES				
chr19	1194824	1219594	DEL,88	NM_001080770	KIR2DL4	ENSG00000189013	Killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4	no	Hom	None	Yes	No
chr19	12896441	12904258	chr13:55877158-55882393,72	NM_002229	JUNB	ENSG00000171223	Jun B proto-oncogene	no	Inter	None	Yes	N/A
chr19	52132546	52149113	DEL,59	NM_001098612	SIGLEC5	ENSG00000105501	Sialic acid binding Ig-like lectin 5	no	Hemi	SegDup	Yes	Yes
chr19	53689152	53691864	DEL,71	NM_024733	ZNF665	ENSG00000197497	Zinc finger protein 665	no	Hom	Yes	No	Yes
chr19	54558395	54560049	DEL_RD	NM_198481	VSTM1	ENSG00000189068	V-set and transmembrane domain containing 1	no	Hom	None	No	Yes
chr20	1560991	1593998	DEL,84	NM_001083910	SIRPB1	ENSG00000101307	Signal-regulatory protein beta 1	no	Hom	SegDup	Yes	Yes
chr20	42271498	42273921	DEL,81	NM_016004	IFT52	ENSG00000101052	Intraflagellar transport 52 homolog (Chlamydomonas)	no	Hom	None	No	Yes
chr20	51820873	51826137	chr3:97545352-97548214,54	NM_173485	TSHZ2	ENSG00000182463	Teashirt zinc finger homeobox 2	no	Inter	None	No	N/A
chr21	44970317	44972908	DEL,125	NM_007031	HSF2BP	ENSG00000160207|ENSG00000180251	Heat shock transcription factor 2 binding protein	no	Hom	None	No	Yes
chr22	18057628	18060280	DEL,177	NM_031481	SLC25A18	ENSG00000182902	Solute carrier family 25 (mitochondrial carrier), member 18	no	Hom	None	No	Yes
chr22	39709007	39715736	chr5:45603558-45620710,64	NM_000967	RPL3	ENSG00000185507	Ribosomal protein L3	no	Inter	Ribosomal Protein	Yes	N/A
chrX	118873929	118876324	DEL,52	NM_002186	IL9R	ENSG00000124334	Interleukin 9 receptor	no				
chrX	118873929	118876324	DEL,52	NM_005638	SYBL1	ENSG00000124333	Synaptobrevin-like 1	no	Hom	None	No	Yes
chrX	118873929	118876324	DEL,52	NM_005840	SPRY3	ENSG00000168939	Sprouty homolog 3 (Drosophila)	no				

Table D.3.11. Genes affected by rearrangements 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**	**5871**	**9051**	**47410**	**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)	**1147**	**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**	**10230**	**11799**	**14376**	**30783**	**11705**	**12942**
Phased unanimously among clones, both alleles observed	545114	563384	17552	17858	21966	29179	5541	6023

Table D.3.12. 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).

Chr	Start	End	Copy Number	HapA CN	HapB CN
chr1	695426	92544807	3	2	1
chr1	92544808	93250555	3	3	0
chr1	93250556	108304451	3	2	1
chr1	108304452	118983114	4	2	2
chr1	118983115	121351595	5	3	2
chr1	145354416	147424817	4	2	2
chr1	148511358	148952924	4	3	1
chr1	148947929	214739993	4	2	2
chr1	214739994	236162977	3	2	1
chr1	236162978	249225340	4	2	2
chr2	13454	40339749	3	2	1
chr2	40339750	41208042	3	3	0
chr2	41208043	41845726	2	2	0
chr2	41845727	41992745	3	3	0
chr2	41992746	57119918	3	2	1
chr2	57119919	57848366	3	3	0
chr2	57848367	63879763	3	2	1
chr2	63879764	64368793	3	3	0
chr2	64368794	90265886	3	2	1
chr2	91753384	92200960	3	2	1
chr2	95326171	106690344	3	2	1
chr2	106690345	106740428	3	3	0
chr2	106740429	222388784	2	2	0
chr2	222388785	222715816	1	1	0
chr2	222715817	243100282	2	2	0
chr3	60000	9675192	3	2	1
chr3	9690192	9992900	4	2	2
chr3	9992901	60027144	3	2	1
chr3	60027145	60260647	2	1	1
chr3	60260648	61126589	1	1	0
chr3	61126590	65556903	2	1	1
chr3	65556904	66016933	1	1	0
chr3	66016934	80281399	2	1	1
chr3	80281400	81385274	2	2	0
chr3	81385275	90091219	2	1	1
chr3	90091220	90311949	3	2	1
chr3	93527460	94463381	3	2	1
chr3	94463382	95016325	3	3	0
chr3	95016326	99198292	2	2	0
chr3	99198293	145492035	3	3	0
chr3	145492036	197919651	4	4	0
chr4	27175	46967020	2	1	1
chr4	46967021	47386717	2	2	0
chr4	47386718	49565614	2	1	1
chr4	52687650	128066863	2	1	1
chr4	128066864	128442232	1	1	0
chr4	128442233	158267825	2	1	1
chr4	158267826	158517367	2	2	0
chr4	158517368	161280735	1	1	0
chr4	161280736	172475206	2	1	1
chr4	172475207	173703673	2	1	1
chr4	173703674	190915842	2	1	1
chr5	17963	1834795	6	6	0
chr5	1834796	4316302	5	5	0
chr5	4316303	34155914	6	6	0
chr5	34155915	38591973	5	5	0
chr5	38591974	40572128	6	6	0
chr5	40572129	45273458	5	5	0
chr5	45273459	46123423	4	4	0
chr5	49558423	58355617	3	2	1
chr5	58355618	58425354	3	3	0
chr5	58425355	58853543	1	1	0
chr5	58853544	58907872	3	3	0
chr5	58907873	68880445	3	2	1
chr6	202273	58720004	3	3	0
chr6	61966206	64435780	3	3	0
chr6	64435781	69136438	2	2	0
chr6	69136439	116684701	3	3	0
chr6	116684702	170921789	2	2	0
chr7	29843	15684942	3	2	1
chr7	15684943	16895503	3	3	0
chr7	16895504	57938791	3	2	1
chr7	62451118	68981361	3	2	1
chr7	68981362	69230394	3	3	0
chr7	69230395	71484217	3	2	1
chr7	71484218	72034104	3	3	0
chr7	72034105	80414808	3	2	1
chr7	80414809	81134166	3	3	0
chr7	81134167	92054951	3	2	1
chr7	92054952	92916009	3	3	0
chr7	92916010	102119384	3	2	1
chr7	102379717	103018223	3	3	0
chr7	103018224	123832241	3	2	1
chr7	123832242	126478678	3	3	0
chr7	126478679	159128494	3	2	1
chr8	100779	43403452	3	2	1
chr8	47454041	116975237	3	2	1
chr8	117120123	119370546	4	3	1
chr8	119370547	120426072	5	4	1
chr8	120426073	144654644	4	3	1
chr8	144654645	146239185	3	2	1
chr9	10843	39189384	3	2	1
chr9	40475834	40837321	3	2	1
chr9	41917223	42249271	3	2	1
chr9	43422269	44908286	3	2	1
chr9	66242215	66556462	3	2	1
chr9	68137998	69866907	3	2	1
chr9	70942723	121299044	3	2	1
chr9	121299045	141112247	5	4	1
chr10	111192	38310982	2	1	1
chr10	38340982	39076503	3	2	1
chr10	42602109	135477933	3	2	1
chr11	184961	1837136	2	2	0
chr11	1837137	2165916	1	1	0
chr11	2165916	2876557	2	2	0
chr11	2876558	10993915	4	2	2
chr11	10993916	12026919	6	4	2
chr11	12026920	13186527	4	2	2
chr11	13186528	20436737	3	2	1
chr11	20436738	21995915	4	2	2
chr11	21995916	22372308	3	2	1
chr11	22372309	22437070	3	3	0
chr11	22437071	23242441	2	2	0
chr11	23242442	24136643	3	3	0
chr11	24136644	24503054	2	2	0
chr11	24503055	24563856	2	1	1
chr11	24563857	25242040	3	2	1
chr11	25336309	25586289	4	2	2
chr11	25586290	27618048	6	4	2
chr11	27681979	28945431	4	2	2
chr11	28945432	29623756	6	4	2
chr11	29623757	36307039	4	2	2
chr11	36307040	45136796	3	2	1
chr11	45151796	46304858	5	3	2
chr11	46304859	46921853	4	2	2
chr11	46921854	49788080	6	4	2
chr11	49788081	50305370	3	2	1
chr11	51376934	51552637	3	2	1
chr11	55027625	59655847	3	2	1
chr11	59655848	85459235	4	2	2
chr11	85459236	88245446	3	2	1
chr11	88799962	89356880	4	2	2
chr11	89729893	89892012	4	3	1
chr11	90071609	91735457	4	2	2
chr11	91735458	92786960	3	2	1
chr11	92786961	93761246	6	4	2
chr11	93761247	102153898	4	2	2
chr11	102153899	102239619	4	4	0
chr11	102239620	134949041	2	2	0
chr12	147379	34435433	3	2	1
chr12	38439886	56390108	4	2	2
chr12	56390109	133841615	3	2	1
chr13	19167980	55866991	2	2	0
chr13	55866992	115108157	3	3	0
chr14	19055798	107293124	3	2	1
chr15	20101189	40790452	3	2	1
chr15	40790453	41413744	5	3	2
chr15	41443744	48354693	4	2	2
chr15	48469520	49335651	4	4	0
chr15	49335652	96361922	4	2	2
chr15	96361923	96612286	3	2	1
chr15	96612287	102517772	4	2	2
chr16	60426	133917	4	3	1
chr16	133918	32227049	3	2	1
chr16	32432133	33669226	2	1	1
chr16	33702945	35153974	3	2	1
chr16	46457240	66670979	3	2	1
chr16	66670980	67625255	3	3	0
chr16	67625256	90174337	3	2	1
chr17	1	22228332	3	2	1
chr17	25268076	46082480	3	2	1
chr17	46082481	46550758	3	3	0
chr17	46550759	81163132	3	2	1
chr18	63620	15383596	3	2	1
chr18	18520339	78018950	3	2	1
chr19	219874	5214799	3	3	0
chr19	5214800	6617274	2	2	0
chr19	6632274	8336454	3	3	0
chr19	8366563	9949072	2	2	0
chr19	9949073	12893034	3	3	0
chr19	12893035	13260609	4	3	1
chr19	13260610	14000595	3	2	1
chr19	14000596	18207258	4	3	1
chr19	18207259	21519480	3	2	1
chr19	21539480	23930875	2	1	1
chr19	23930876	24366062	3	2	1
chr19	28263494	38625599	3	2	1
chr19	38625600	39370289	2	1	1
chr19	39370290	41314803	3	2	1
chr19	41314804	48599012	2	1	1
chr19	48599013	50294132	3	2	1
chr19	50294133	51537384	4	3	1
chr19	51537385	52006722	3	2	1
chr19	52006723	55478045	2	1	1
chr19	55478046	56098991	4	3	1
chr19	56098992	56892562	3	2	1
chr19	56892563	59103179	2	1	1
chr20	60001	26258753	2	1	1
chr20	29419647	62923146	4	3	1
chr21	9827582	11181667	2	1	1
chr21	14369463	14491992	2	1	1
chr21	14594419	48105087	3	2	1
chr22	16385650	51231477	2	2	0
chrX	2699520	4104600	2	2	0
chrX	4104601	57961123	3	3	0
chrX	62044948	99785173	3	3	0
chrX	99810173	154801667	2	2	0

Table D.3.13. Haplotype-resolved copy number calls for HeLa CCL-2

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 D.3.14. 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 D.3.15. 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.

D.4 Supplementary figures for Chapter 6

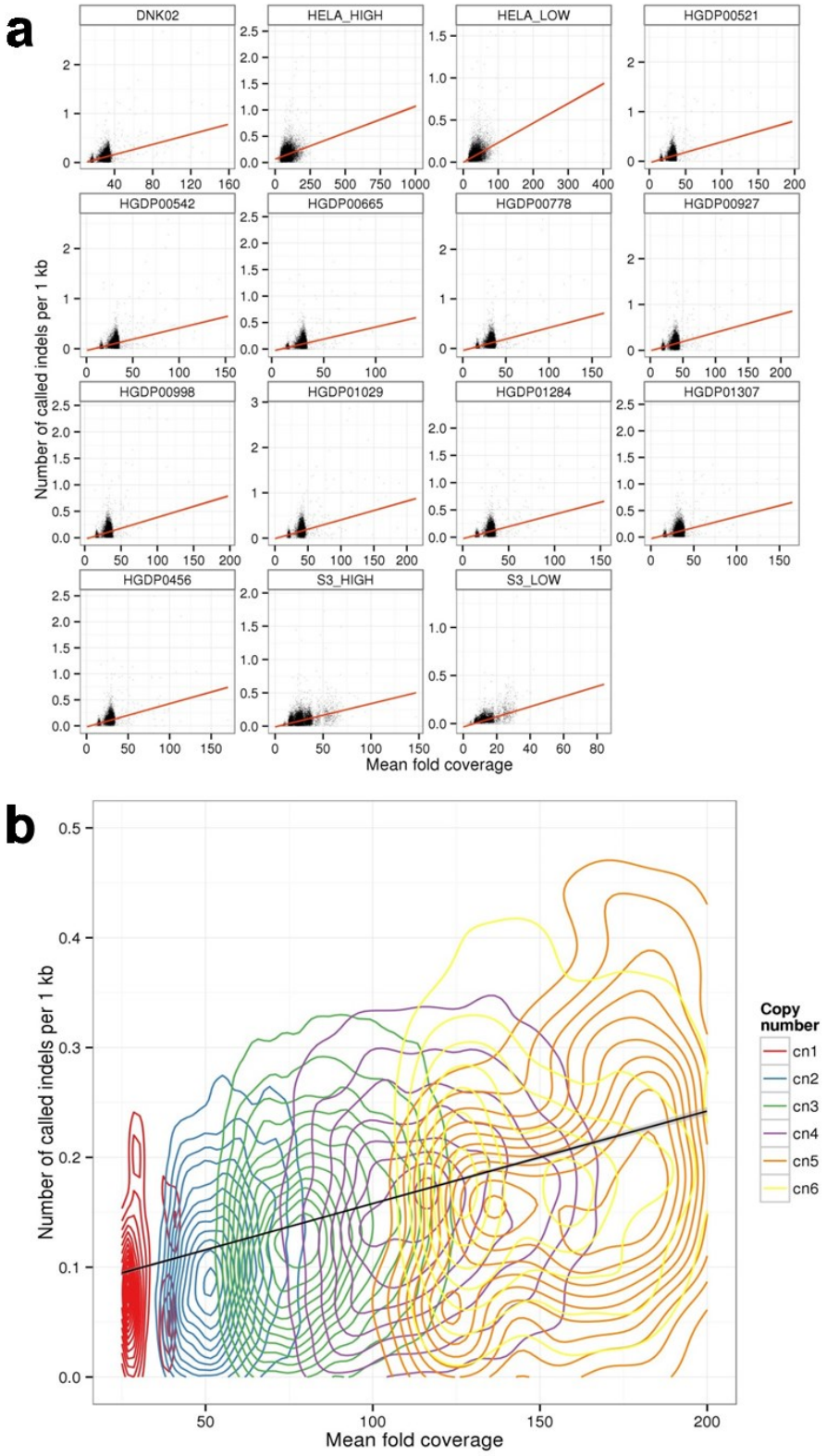


Figure D.4.1. Indel calling by coverage.
a. Counts of indels called is plotted versus read depth at each indel for HeLa and 11 HGDP controls. Shotgun reads from HeLa CCL-2 as well as HeLa S3 were randomly downsampled in order to study the effects of lower coverage upon indel counts in each genome. Mean 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.

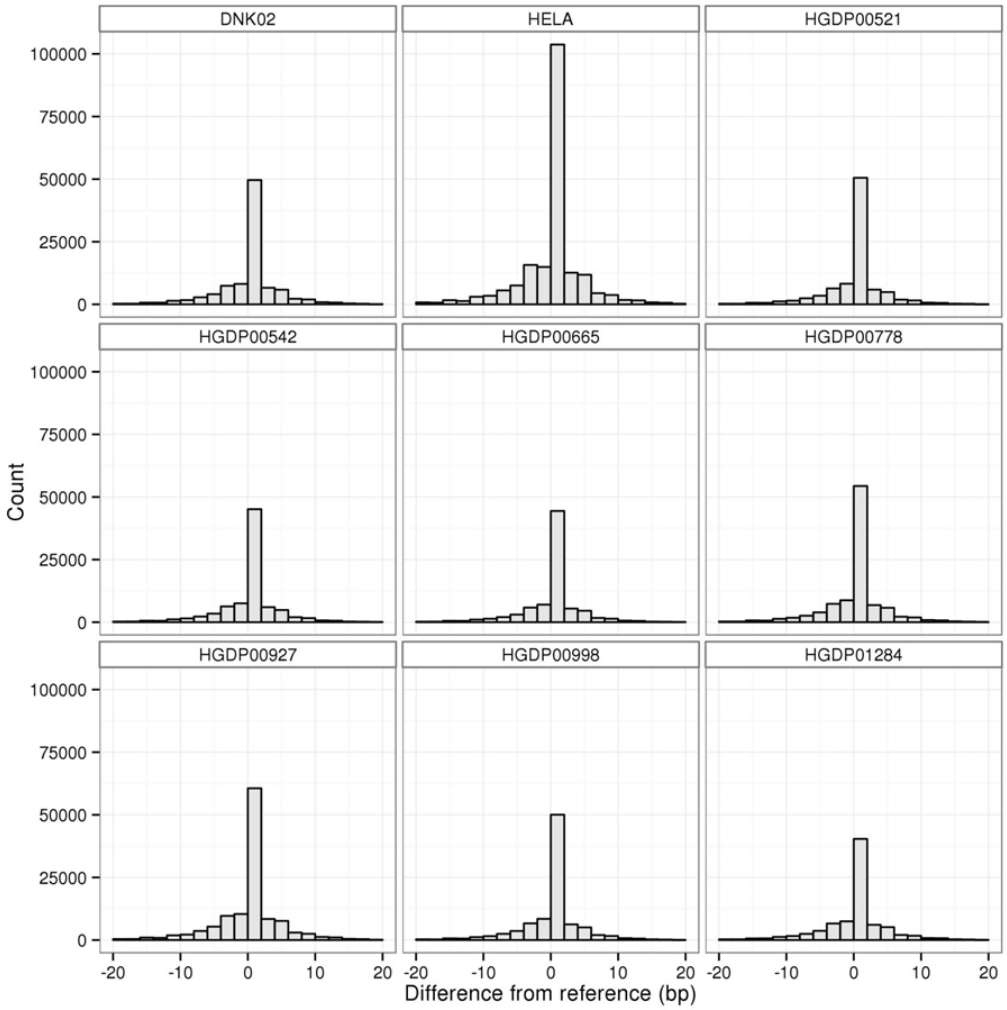


Figure D.4.2. STR profiling with lobSTR.
Short Tandem Repeats (STRs) were identified using lobSTR (Gymrek, M, et. al. (2012)) for HeLa as well as eight of the diversity panel control individuals (Rohland, N; Reich, D (2012)). Repeats with a coverage of at least 10 are represented above as a histogram 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.

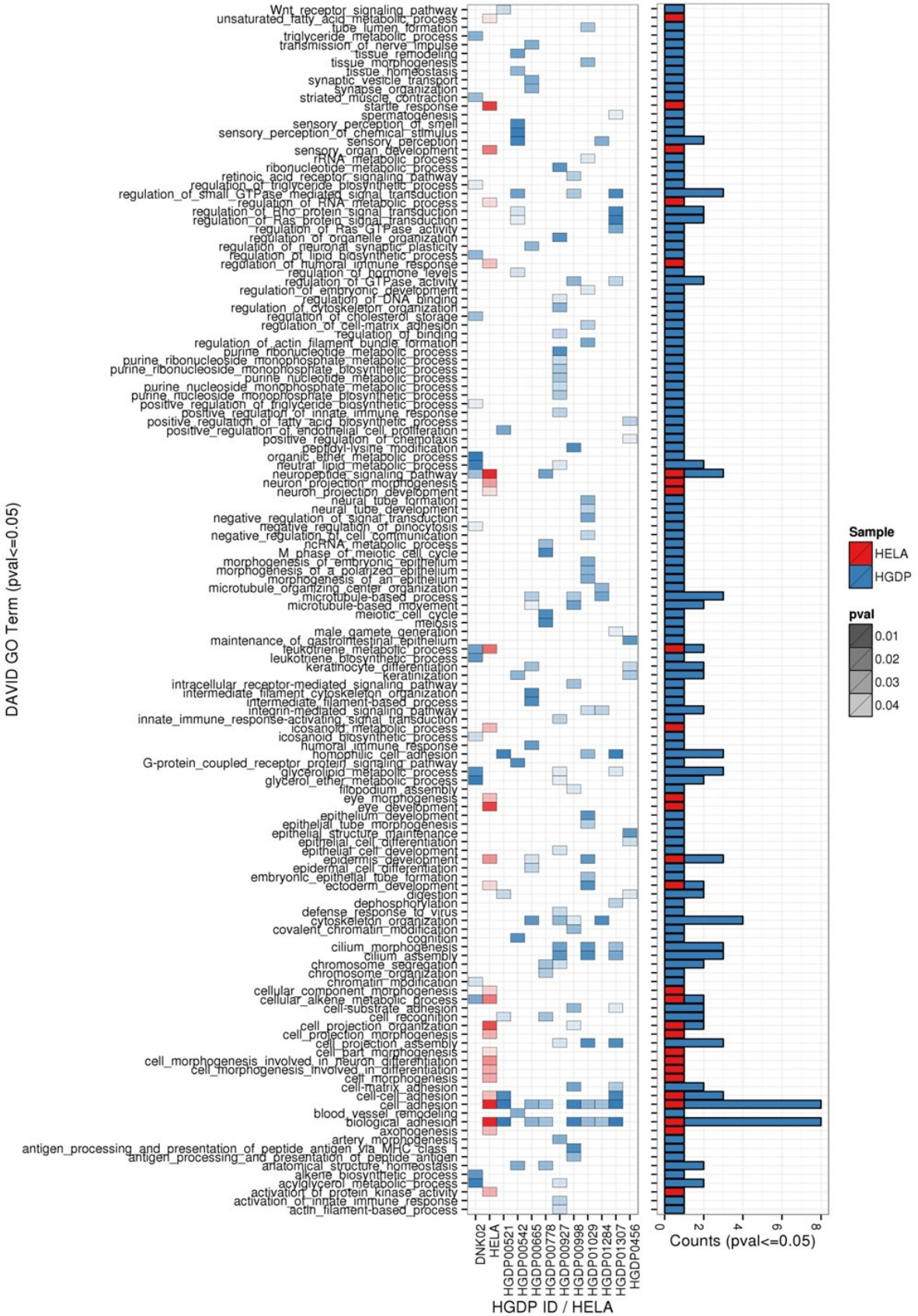

DAVID GO Term (pval<=0.05)
Wnt_receptor_signaling_pathway
unsaturated_fatty_acid_metabolic_process
tube_lumen_formation
triglyceride_metabolic_process
transmission_of_nerve_impulse
tissue_remodeling
tissue_morphogenesis
tissue_homeostasis
synaptic_vesicle_transport
synapse_organization
striated_muscle_contraction
startle_response
spermatogenesis
sensory_perception_of_smell
sensory_perception_of_chemical_stimulus
sensory_perception
sensory_organ_development
rRNA_metabolic_process
ribonucleotide_metabolic_process
retinoic_acid_receptor_signaling_pathway
regulation_of_triglyceride_biosynthetic_process
regulation_of_small_GTPase_mediated_signal_transduction
regulation_of_RNA_metabolic_process
regulation_of_Rho_protein_signal_transduction
regulation_of_Ras_protein_signal_transduction
regulation_of_Ras_GTPase_activity
regulation_of_organelle_organization
regulation_of_neuronal_synaptic_plasticity
regulation_of_lipid_biosynthetic_process
regulation_of_humoral_immune_response
regulation_of_hormone_levels
regulation_of_GTPase_activity
regulation_of_embryonic_development
regulation_of_DNA_binding
regulation_of_cytoskeleton_organization
regulation_of_cholesterol_storage
regulation_of_cell-matrix_adhesion
regulation_of_binding
regulation_of_actin_filament_bundle_formation
purine_ribonucleotide_metabolic_process
purine_ribonucleoside_monophosphate_metabolic_process
purine_ribonucleoside_monophosphate_biosynthetic_process
purine_nucleotide_metabolic_process
purine_nucleoside_monophosphate_metabolic_process
purine_nucleoside_monophosphate_biosynthetic_process
positive_regulation_of_triglyceride_biosynthetic_process
positive_regulation_of_innate_immune_response
positive_regulation_of_fatty_acid_biosynthetic_process
positive_regulation_of_endothelial_cell_proliferation
positive_regulation_of_chemotaxis
peptidyl-lysine_modification
organic_ether_metabolic_process
neutral_lipid_metabolic_process
neuropeptide_signaling_pathway
neuron_projection_morphogenesis
neuron_projection_development
neural_tube_formation
neural_tube_development
negative_regulation_of_signal_transduction
negative_regulation_of_pinocytosis
negative_regulation_of_cell_communication
ncRNA_metabolic_process
M_phase_of_meiotic_cell_cycle
morphogenesis_of_embryonic_epithelium
morphogenesis_of_a_polarized_epithelium
morphogenesis_of_an_epithelium
microtubule_organizing_center_organization
microtubule-based_process
microtubule-based_movement
meiotic_cell_cycle
meiosis
male_gamete_generation
maintenance_of_gastrointestinal_epithelium
leukotriene_metabolic_process
leukotriene_biosynthetic_process
keratinocyte_differentiation
keratinization
intracellular_receptor-mediated_signaling_pathway
intermediate_filament_cytoskeleton_organization
intermediate_filament-based_process
integrin-mediated_signaling_pathway
innate_immune_response-activating_signal_transduction
icosanoid_metabolic_process
icosanoid_biosynthetic_process
humoral_immune_response
homophilic_cell_adhesion
G-protein_coupled_receptor_protein_signaling_pathway
glycerolipid_metabolic_process
glycerol_ether_metabolic_process
filopodium_assembly
eye_morphogenesis
eye_development
epithelium_development
epithelial_tube_morphogenesis
epithelial_structure_maintenance
epithelial_cell_differentiation
epithelial_cell_development
epidermis_development
epidermal_cell_differentiation
embryonic_epithelial_tube_formation
ectoderm_development
digestion
dephosphorylation
defense_response_to_virus
cytoskeleton_organization
covalent_chromatin_modification
cognition
cilium_morphogenesis
cilium_assembly
chromosome_segregation
chromosome_organization
chromatin_modification
cellular_component_morphogenesis
cellular_alkene_metabolic_process
cell-substrate_adhesion
cell_recognition
cell_projection_organization
cell_projection_morphogenesis
cell_projection_assembly
cell_part_morphogenesis
cell_morphogenesis_involved_in_neuron_differentiation
cell_morphogenesis_involved_in_differentiation
cell_morphogenesis
cell-matrix_adhesion
cell-cell_adhesion
cell_adhesion
blood_vessel_remodeling
biological_adhesion
axonogenesis
artery_morphogenesis
antigen_processing_and_presentation_of_peptide_antigen_via_MHC_class_I
antigen_processing_and_presentation_of_peptide_antigen
anatomical_structure_homeostasis
alkene_biosynthetic_process
acylglycerol_metabolic_process
activation_of_protein_kinase_activity
activation_of_innate_immune_response
actin_filament-based_process
DNK02
HELA
HGDP00521
HGDP00542
HGDP00665
HGDP00778
HGDP00927
HGDP00998
HGDP01029
HGDP01284
HGDP01307
HGDP0456
HGDP ID / HELA
0
2
4
6
8
Counts (pval<=0.05)
Sample
HELA
HGDP
pval
0.01
0.02
0.03
0.04

Figure D.4.3. 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 D.4.4. 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 kbp), with predicted copy number state overlaid (black dots).

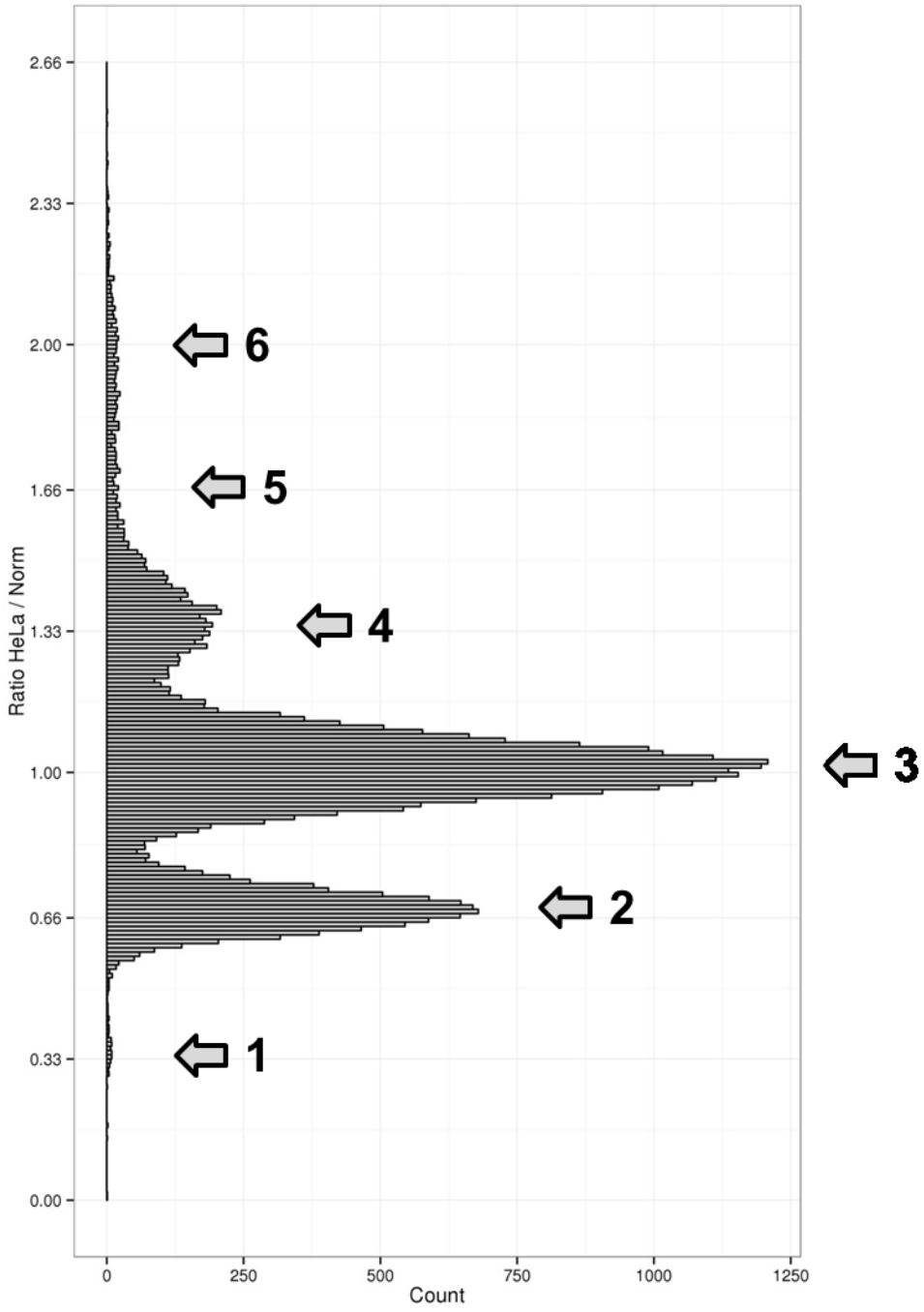
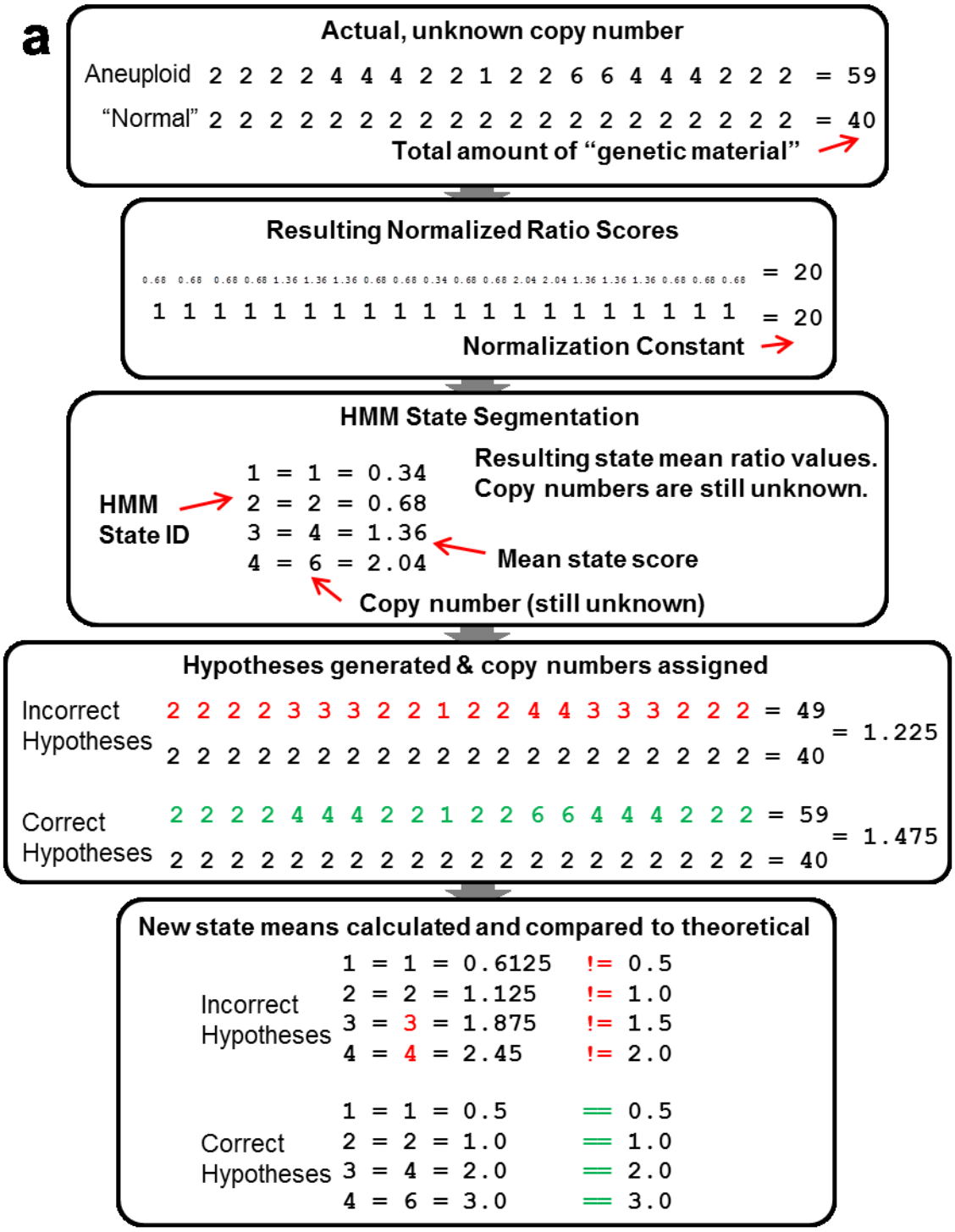
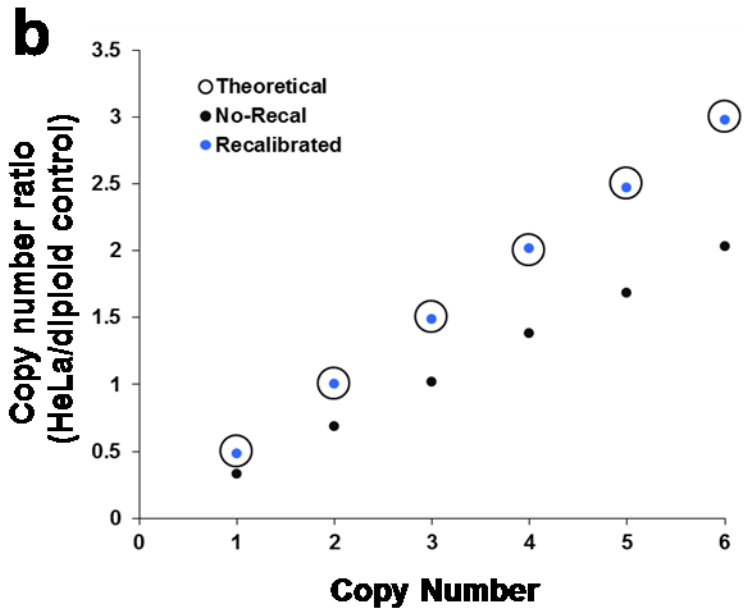
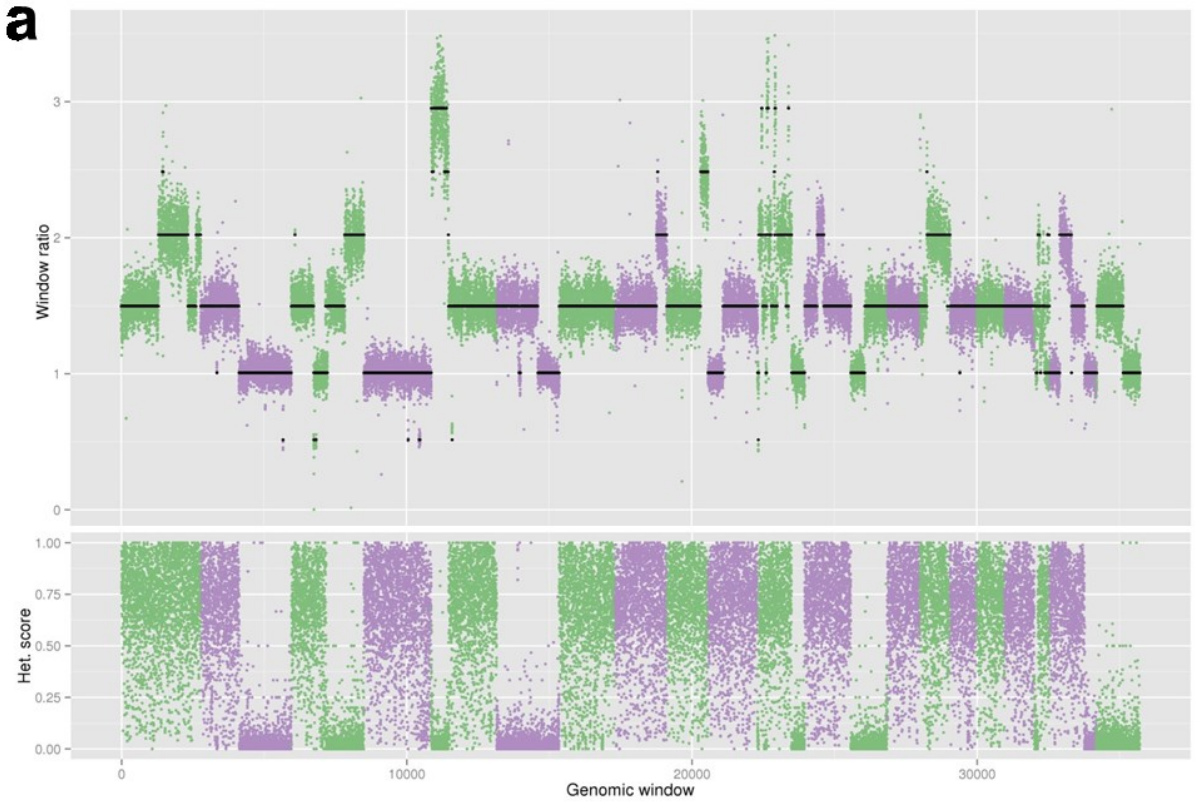


Figure D.4.5. HeLa over GC-matched control ratio histogram.
SUNK window (500 unique 30mer) resolution ratio scores plotted as a histogram. Distinct peaks are observed at approximately 0.33, 0.66, 1.0, 1.33, consistent with an approximately triploid numerator sample (HeLa) over a diploid denominator sample (GC-matched control). Inferred copy numbers are indicated by arrows.


a
Actual, unknown copy number
Aneuploid 2 2 2 2 4 4 4 2 2 1 2 2 6 6 4 4 4 2 2 2 = 59
"Normal" 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 = 40
Total amount of "genetic material"
Resulting Normalized Ratio Scores
0.68 0.68 0.68 0.68 1.36 1.36 1.36 0.68 0.68 0.34 0.68 0.68 2.04 2.04 1.36 1.36 1.36 0.68 0.68 0.68 = 20
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 = 20
Normalization Constant
HMM State Segmentation
HMM State ID
1 = 1 = 0.34
2 = 2 = 0.68
3 = 4 = 1.36
4 = 6 = 2.04
Resulting state mean ratio values.
Copy numbers are still unknown.
Mean state score
Copy number (still unknown)
Hypotheses generated & copy numbers assigned
Incorrect Hypotheses
2 2 2 2 3 3 3 2 2 1 2 2 4 4 3 3 3 2 2 2 = 49
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 = 40
= 1.225
Correct Hypotheses
2 2 2 2 4 4 4 2 2 1 2 2 6 6 4 4 4 2 2 2 = 59
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 = 40
= 1.475
New state means calculated and compared to theoretical
Incorrect Hypotheses
1 = 1 = 0.6125 != 0.5
2 = 2 = 1.125 != 1.0
3 = 3 = 1.875 != 1.5
4 = 4 = 2.45 != 2.0
Correct Hypotheses
1 = 1 = 0.5 == 0.5
2 = 2 = 1.0 == 1.0
3 = 4 = 2.0 == 2.0
4 = 6 = 3.0 == 3.0



Figue D.4.6. 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.


a
Window ratio
Het. score
Genomic window
0
10000
20000
30000

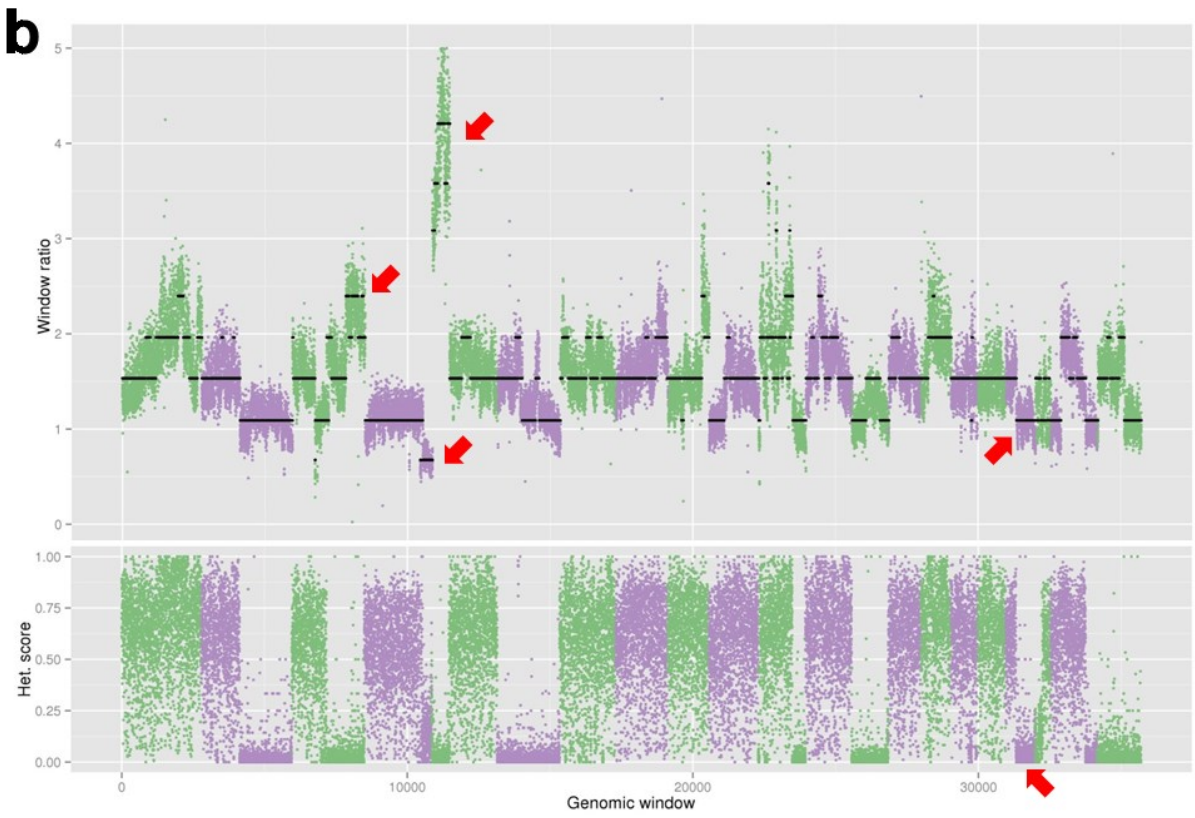

b
Window ratio
Het. score
Genomic window
0
10000
20000
30000

Figure D.4.7. 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.

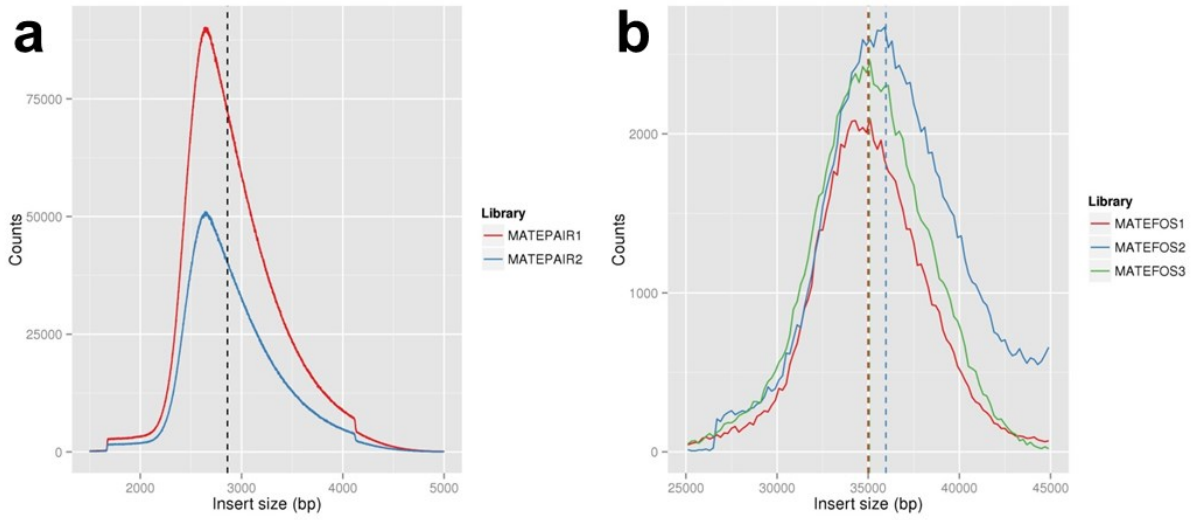


Figure D.4.8. Mate pair insert size distributions.
a, Insert size distributions of concordant pairs for the two "3 kbp" mate-pair libraries constructed using *in vitro* circularization (Talkowski *et. al.* (2011)). b, Insert size distributions of concordant pairs for the three "40 kbp" mate-pair libraries constructed using fosmid cloning.

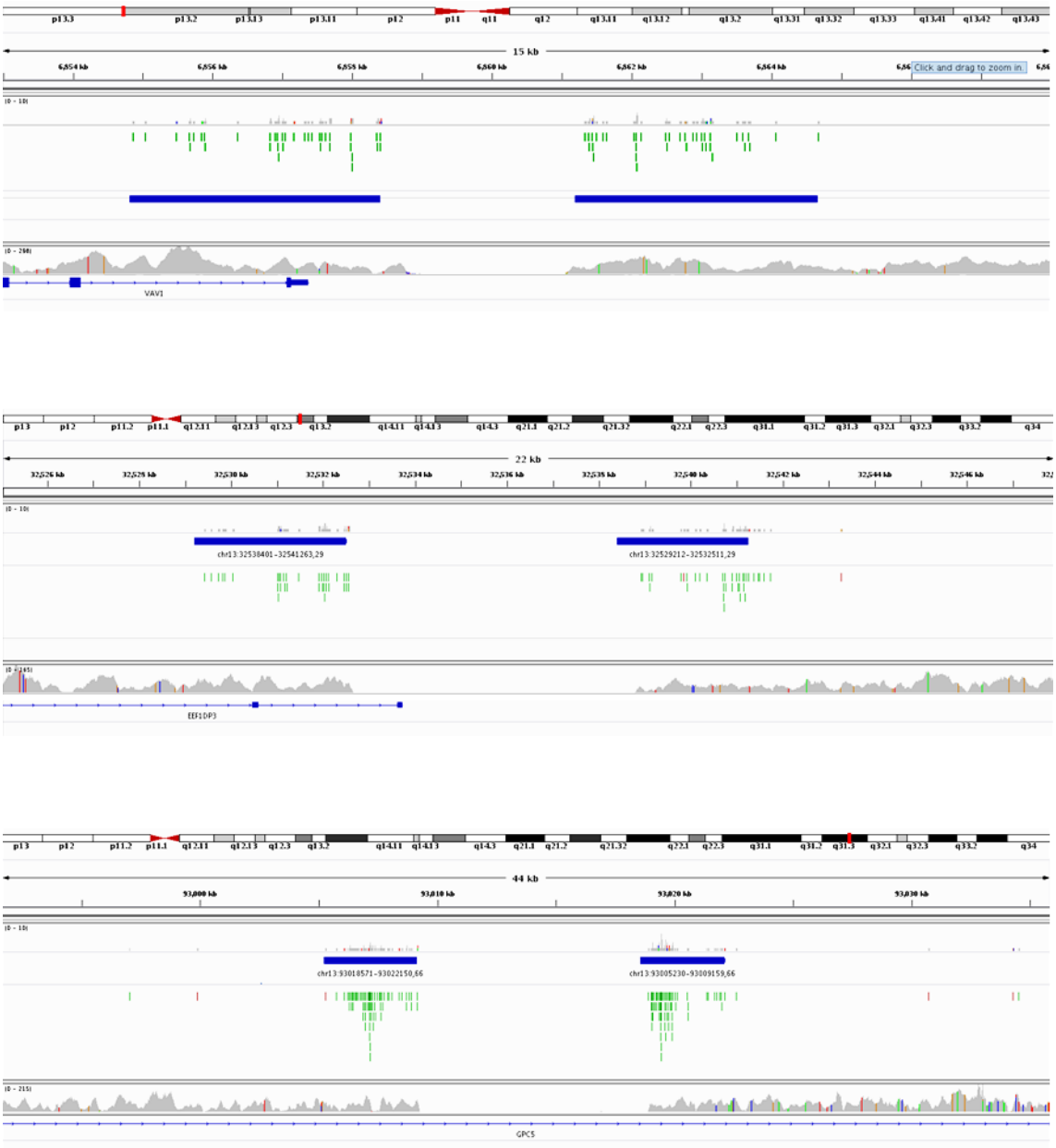


Figure D.4.9. Examples of deletions in HeLa CCL-2.
Three examples of deletions called using a sliding window approach shown in the IGV genome browser. Blue bars denote regions of coverage from supporting 3 kbp mate-paired reads (green ticks). Shotgun sequence coverage (gray bars) are plotted beneath each event.

Figure D.4.10. 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 kbp mate-pair reads. The upper example is one of the rearrangements within marker chromosome M14.

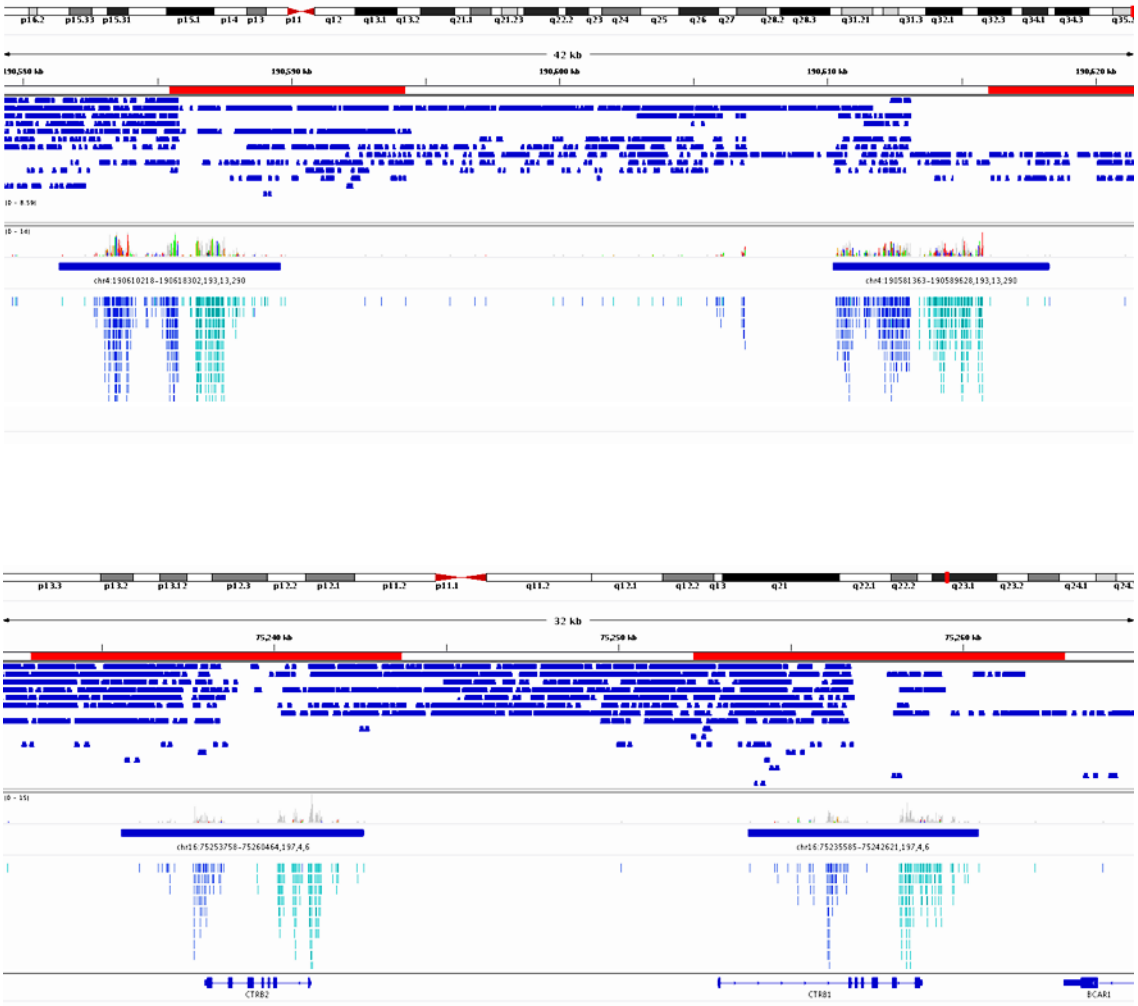


Figure D.4.11. Called inversion examples in HeLa CCL-2.
Two examples of inversions detected by a sliding window approach from discordantly-mapping 3 kbp mate-pair reads. Both inversions are supported by fosmid sequence coverage profiles (blue tracks shown below chromosome ideograms), with overlapping clones showing discontinuous patterns of coverage near each inversion breakpoint.

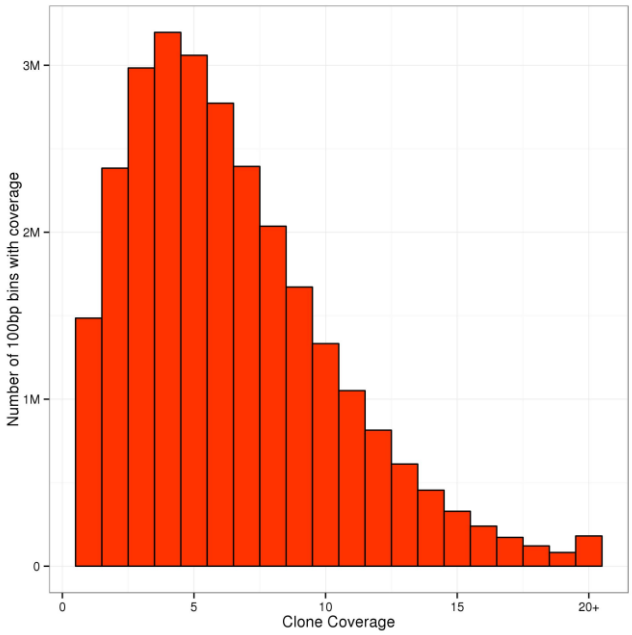


Figure D.4.12. Histogram of clone coverage.
Histogram of the physical coverage by fosmid clone inserts. Overall, 3.5% of the genome is not covered (coverage=0, excluding chromosome Y and assembly gaps).

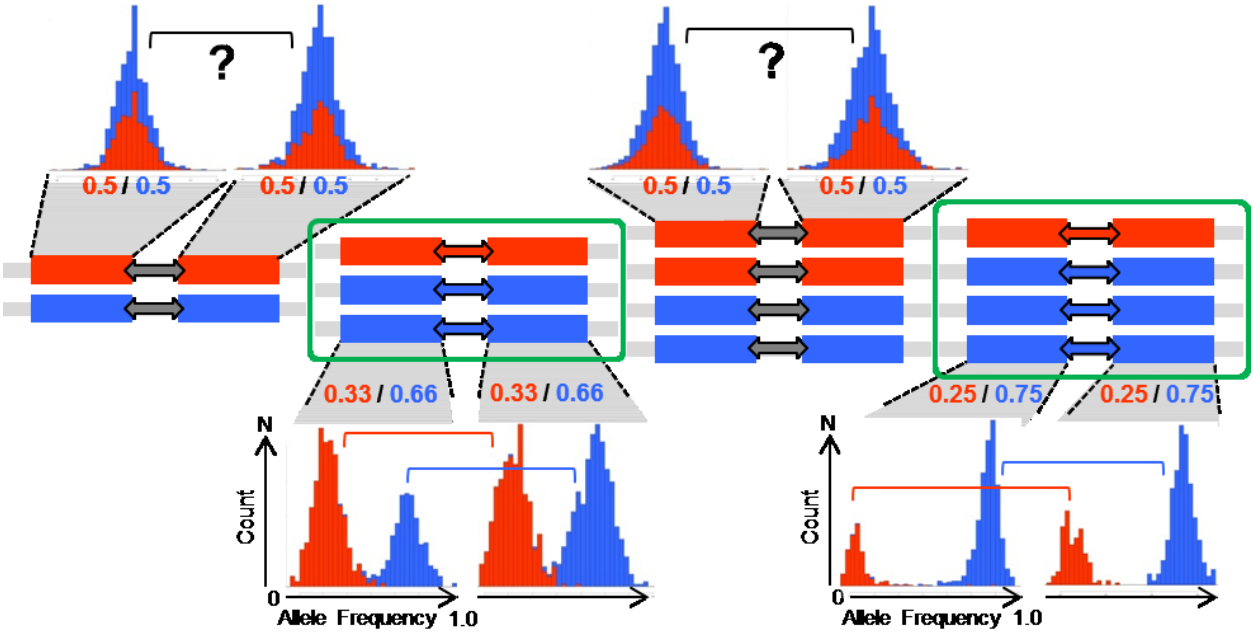


Figure D.4.13. Schematic of haplotype scaffolding approach using allele imbalance.
Consecutive haplotype blocks in regions containing imbalanced copy numbers (green boxes) between haplotypes can be merged using an HMM to form a haplotype scaffold based on the allele frequencies of phased variants within 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).

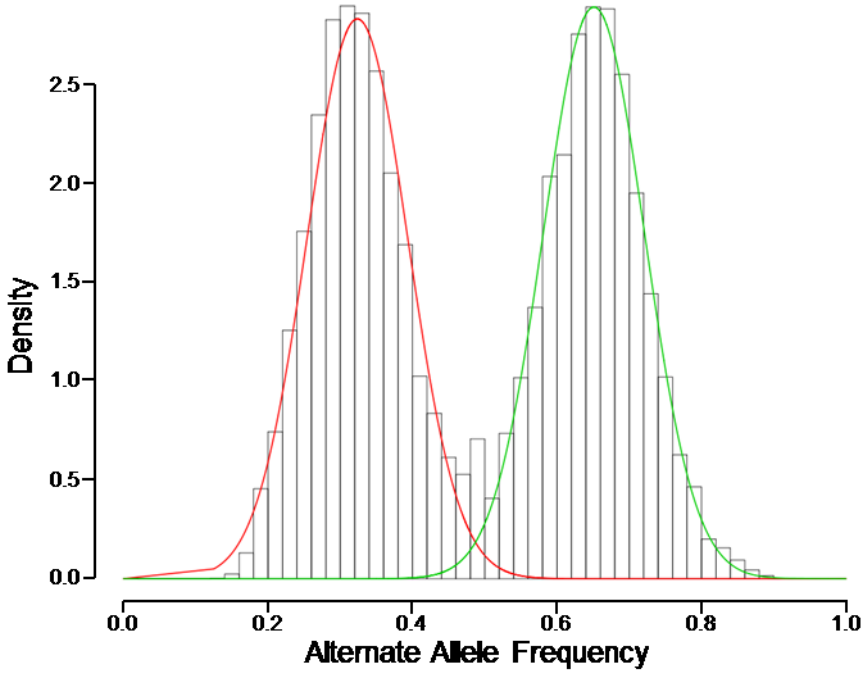


Figure D.4.14. Gaussian mixture model of AAFs in non-LOH copy number 3 regions.
A histogram of alternate allele frequencies among shotgun reads is shown for all heterozygous variants present in regions of copy number 3 in which one haplotype is at copy number 2 and the other at copy number 1. A two-component 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.

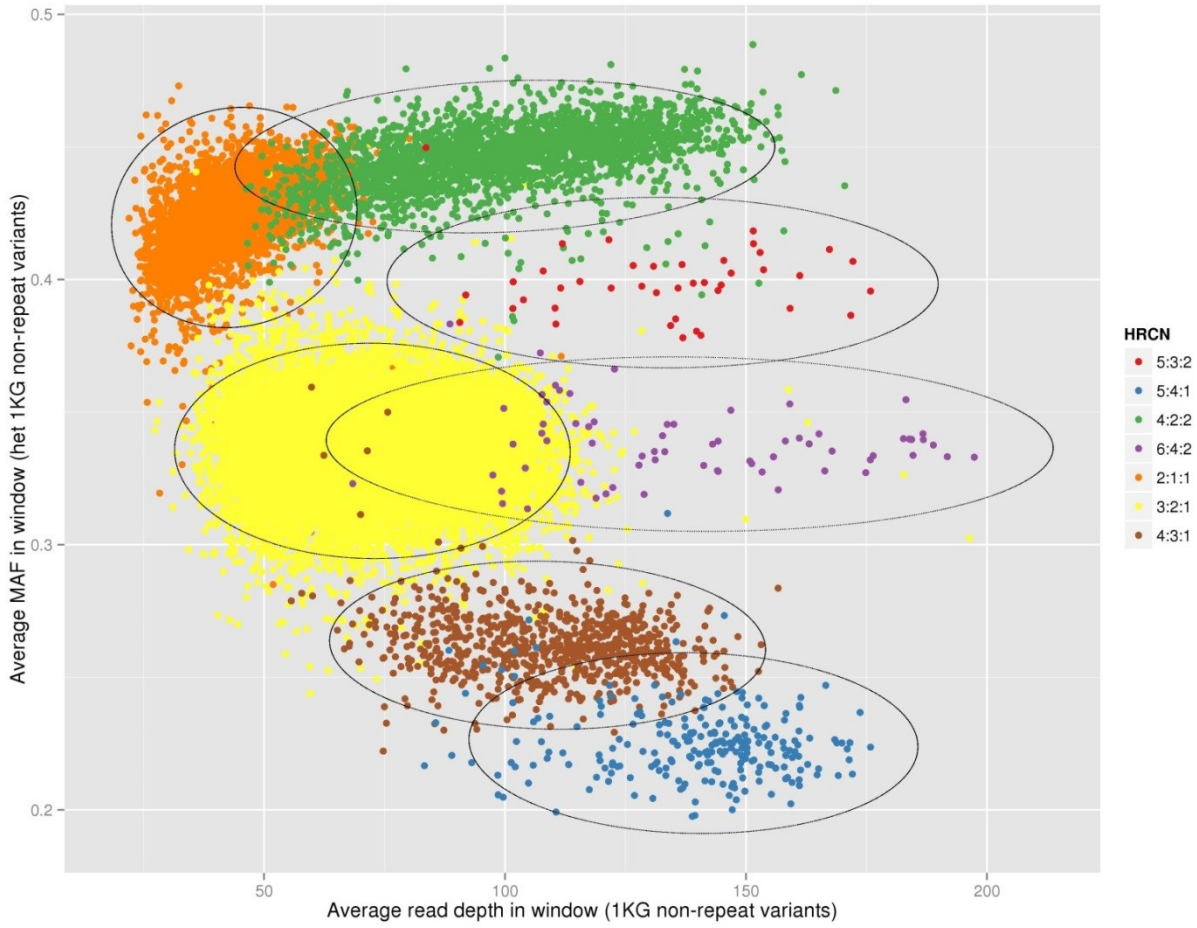


Figure D.4.15. HeLa allele balance by read depth for HRCN regions.
For each low resolution SUNK window (~77 kbp), the average minor allele frequency of all heterozygous variants was plotted against those sites' average read depth. Each point was shaded by the window's predicted HRCN (total copy number : haplotype A copy number : haplotype B copy number). Overlaid ellipses represent 95% confidence intervals for each HRCN grouping.

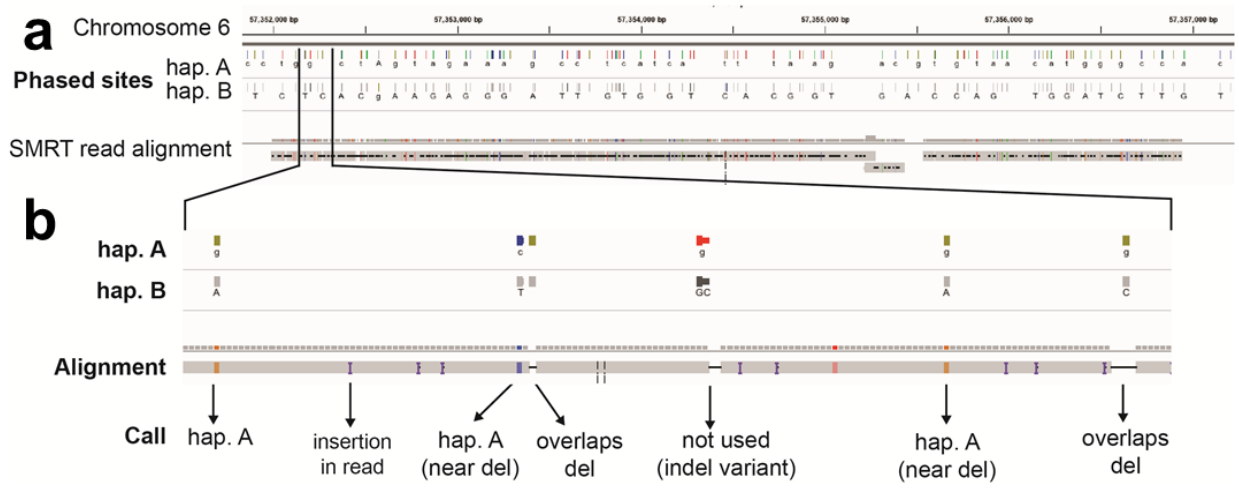


Figure D.4.16. Long-range haplotype validation.
Haplotype phase validation by single molecule long-read sequencing. An example alignment from one read spanning 4.96 kbp is shown. a. Upper track: phased variants in HeLa CCL-2 are shown for each inherited haplotype (A and B), with gray ticks indicating 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 A. b. Detail showing aligned positions matching haplotype A or rejected due to overlapping or nearby indel errors.

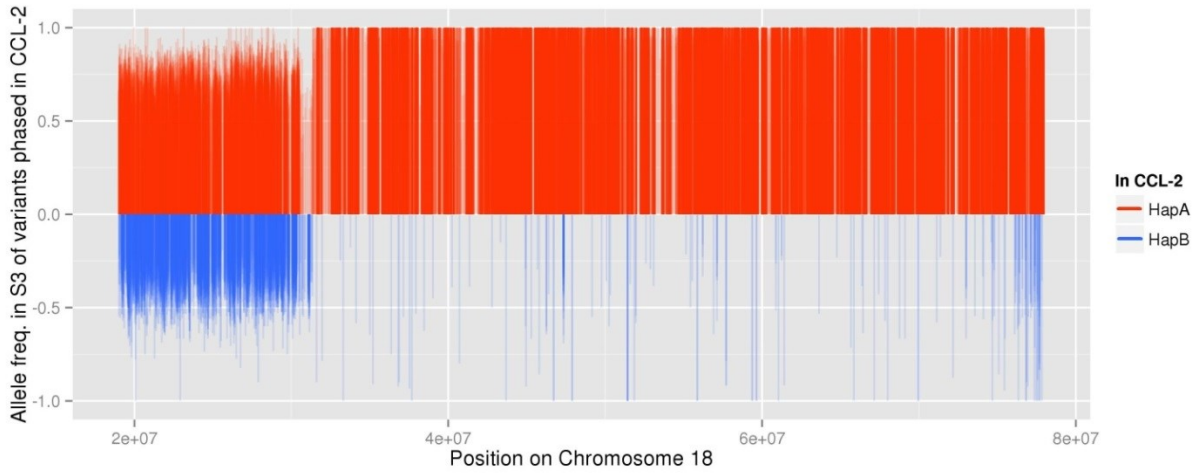


Figure D.4.17. Allelic state across LOH event specific to HeLa S3.
Allele frequencies among HeLa S3 shotgun reads are shown for all heterozygous and phased variants from HeLa CCL-2, across 78.1 Mbp of chromosome 18. Allele frequencies in S3 are plotted on the y-axis, with points' direction and color 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.

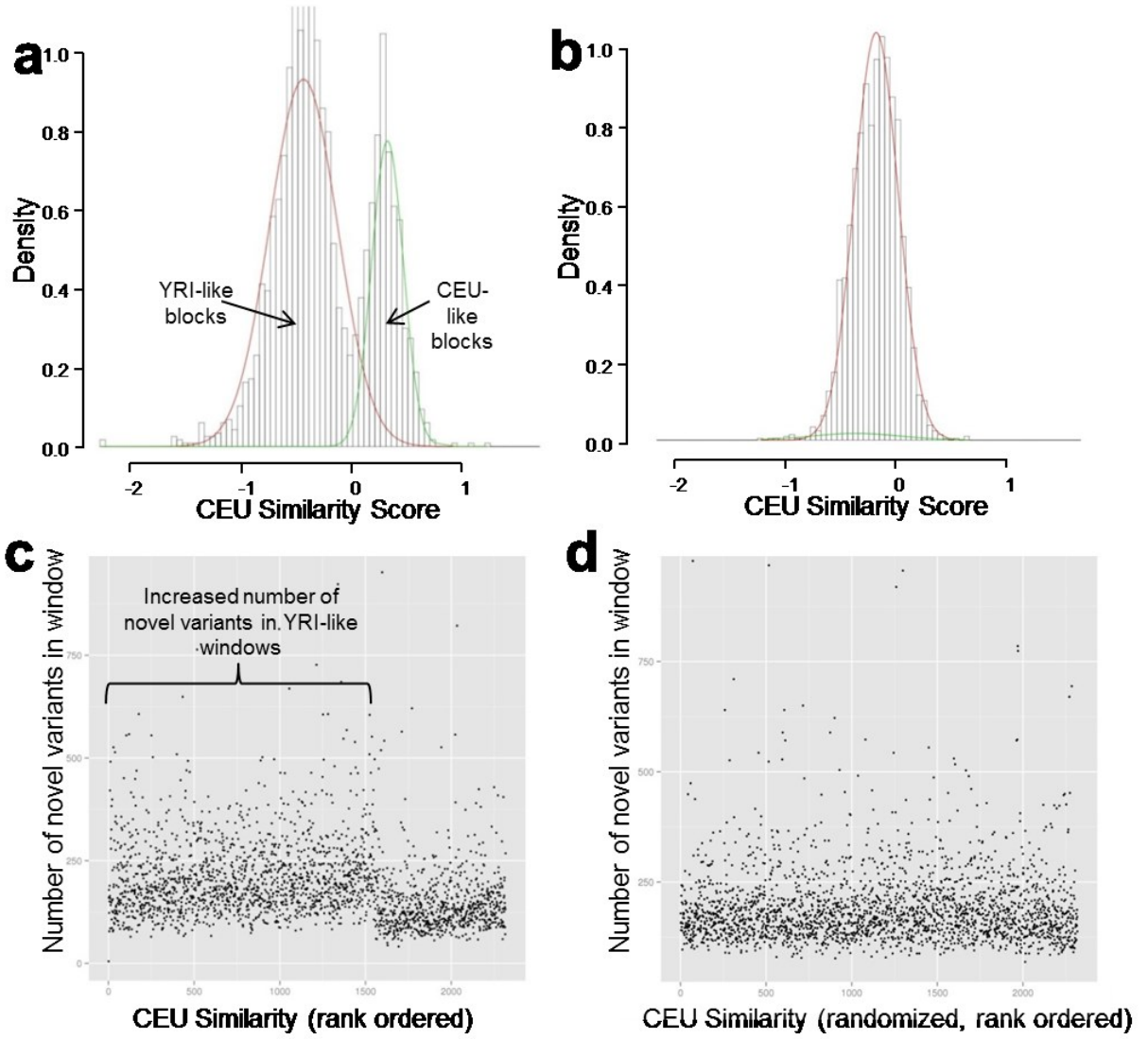
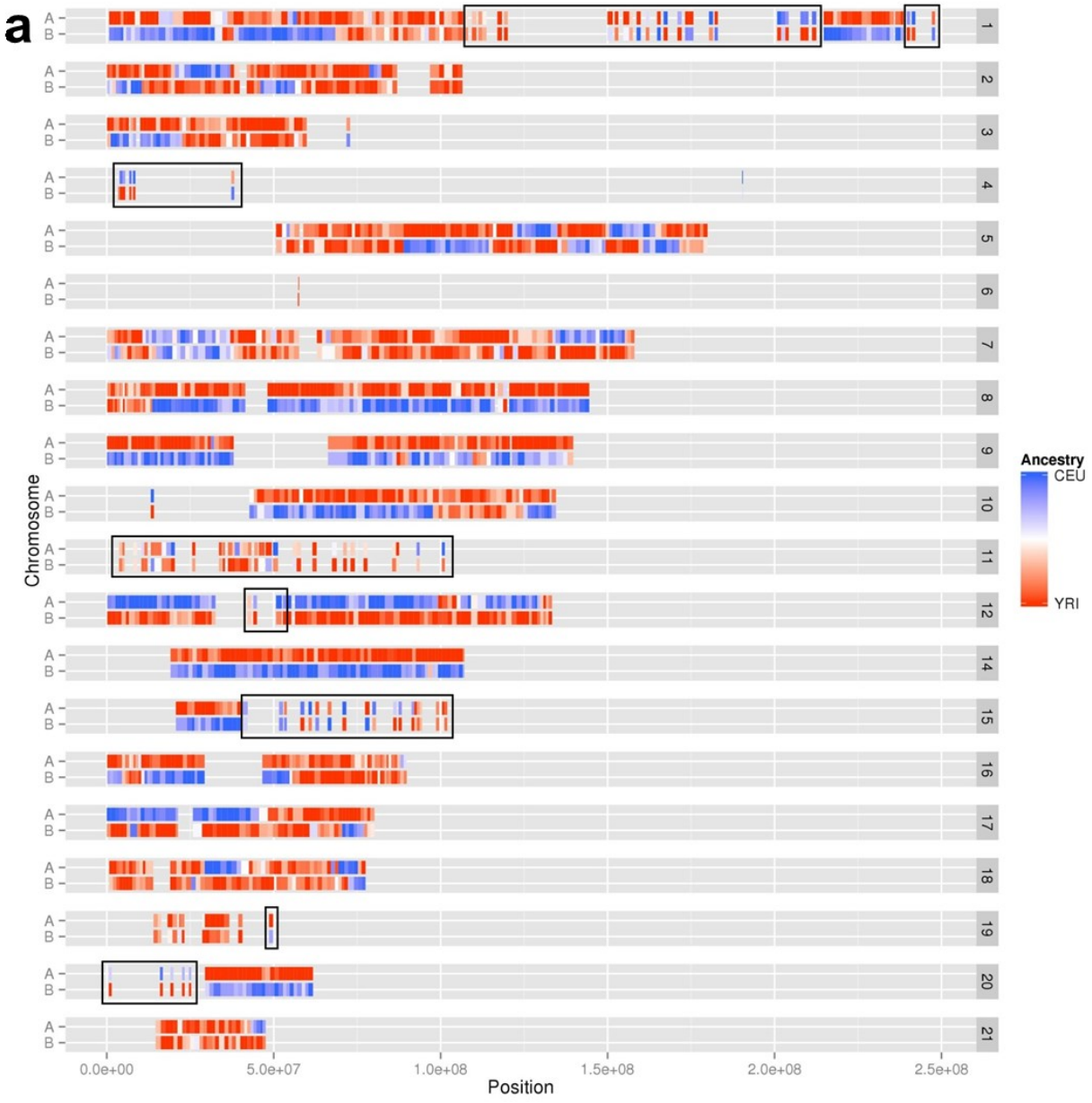


Figure D.4.18. Population-based haplotype analysis.
a. Histogram of windowed scores based based upon phased sites' popluation allele frequencies in CEU vs YRI individuals (from the 1000 Genomes Project). Red and green lines indicate density from a two-component mixture model fit. b. Randomization test. 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.


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

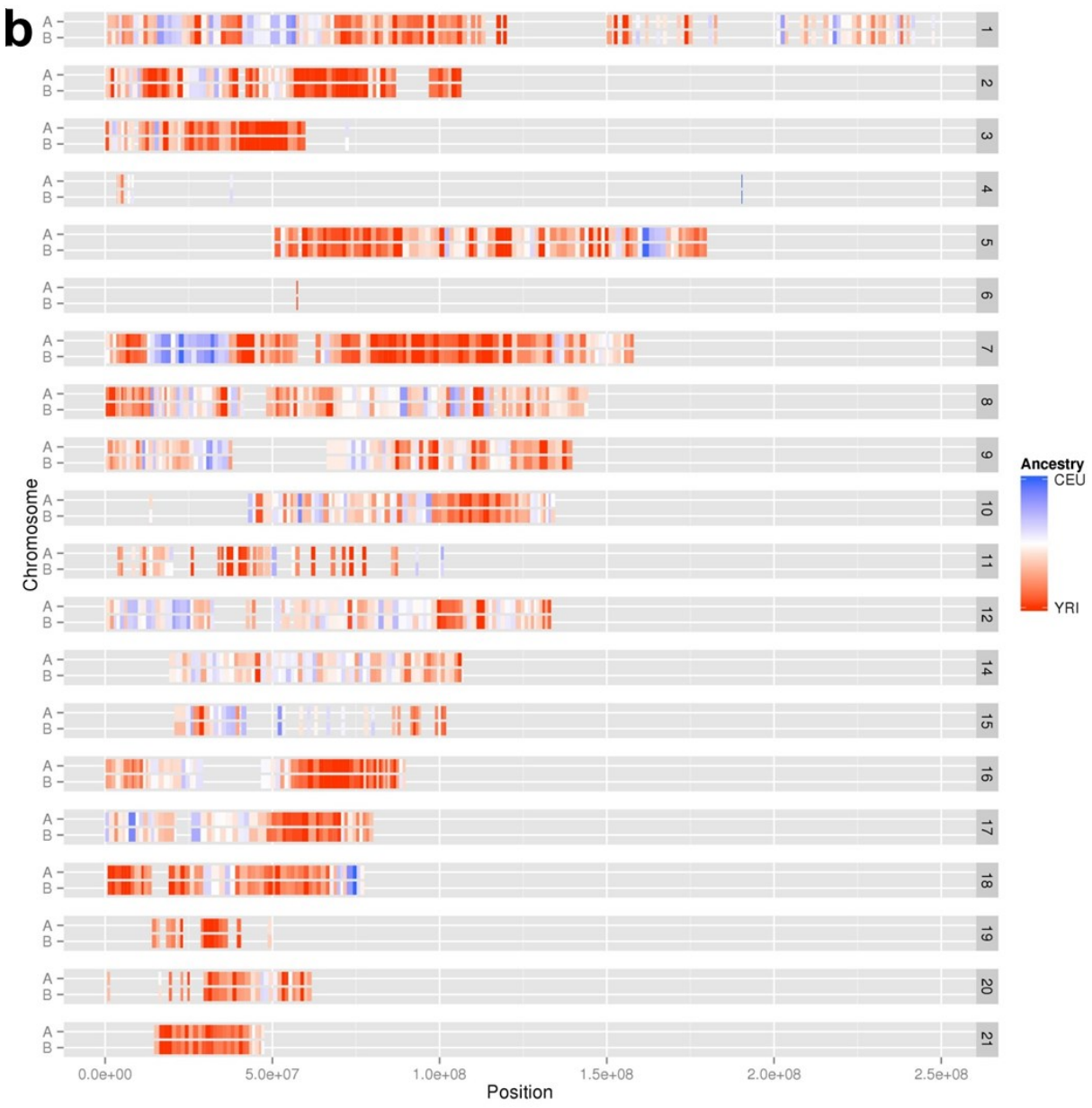


Figure D.4.19. Haplotype-based local inference of genetic ancestry.
a. Predicted genetic ancestry is shown for haplotype windows, using scores of allele frequency to CEU or YRI populations and colored by the ancestry similarity (Blue = CEU, Red = YRI). Windows in LOH regions, in haplotype scaffolds with 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.

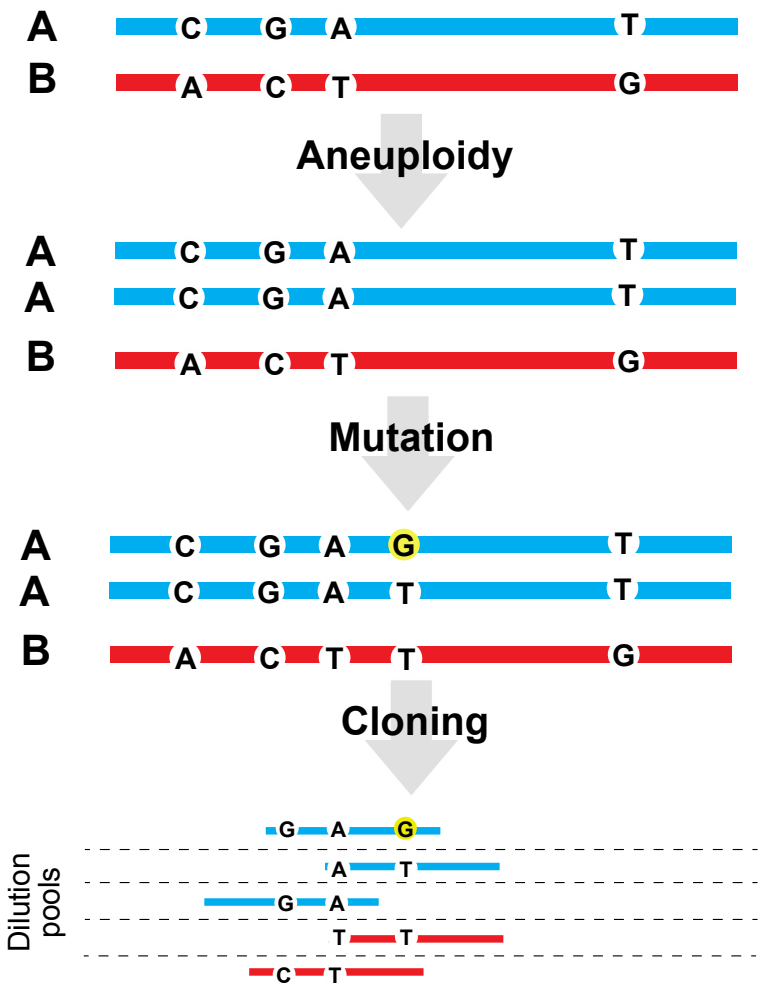


Figure D.4.20. Post-aneuploidy mutation analysis.
Schematic of validation process for somatic, post-aneuploidy mutations by large insert clone pool sequencing. Mutations arising after duplication of a germline haplotype (blue) are confirmed by the presence of both the mutant allele (yellow, "G") as well as the reference allele (T) in separate clones derived from the duplicated haplotype.

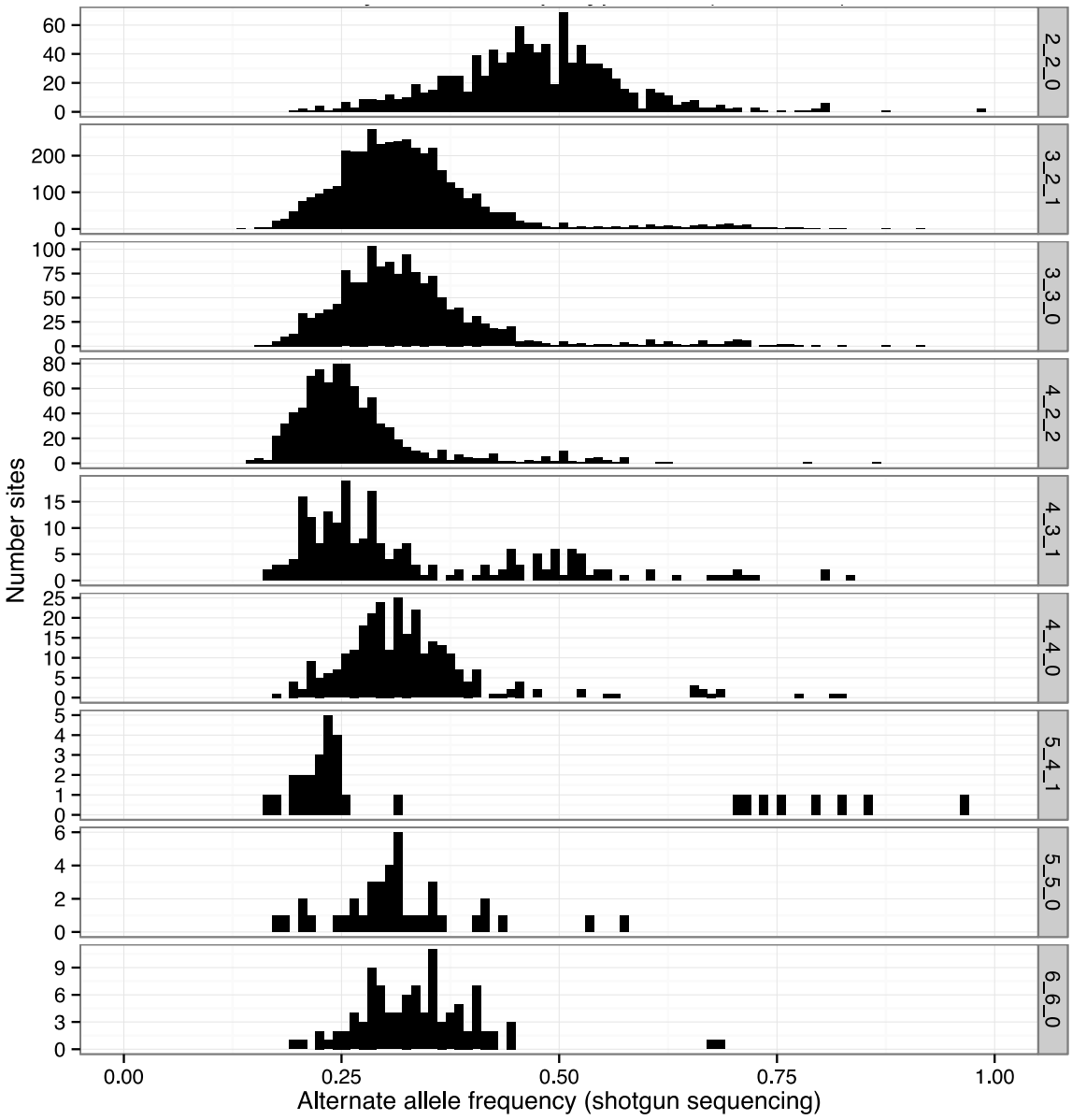


Figure D.4.21. Somatic mutation allele frequencies.
Histograms of allele frequency within shotgun data of clone-validated somatic mutations, split by haplotype-resolved copy number (HRCN) state. Regions with HRCN of 5:3:2 and 6:4:2 were omitted because there were few sites (each <= 10).

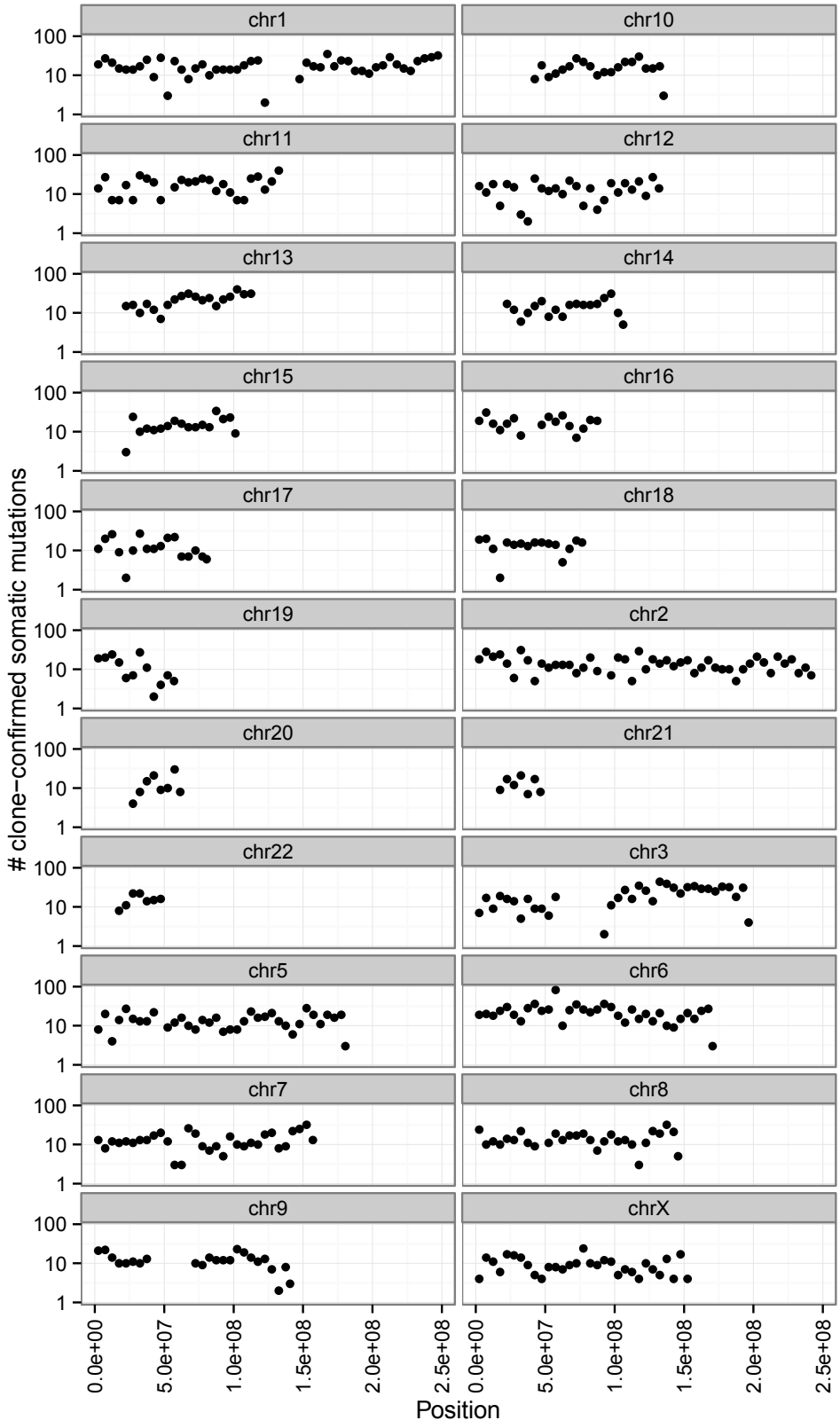


Figure D.4.22. Somatic mutation counts.
Count of somatic mutations per 5 Mbp window along each chromosome.

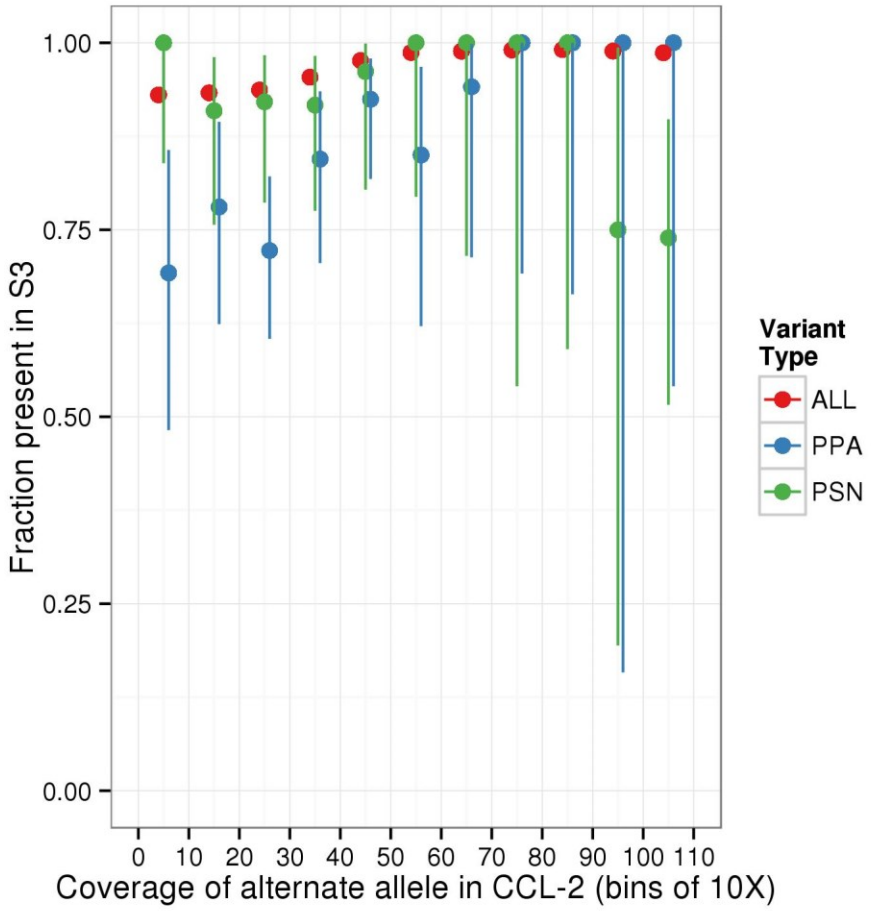


Figure D.4.23. Private alleles shared between HeLa CCL-2 and S3.
The fraction of private SNVs (not found in the 1000 Genomes Project) from HeLa CCL-2 that are also observed in S3 is shown, binned by the number of reads supporting the alternate allele in CCL-2. The fraction of shared alleles is shown 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.

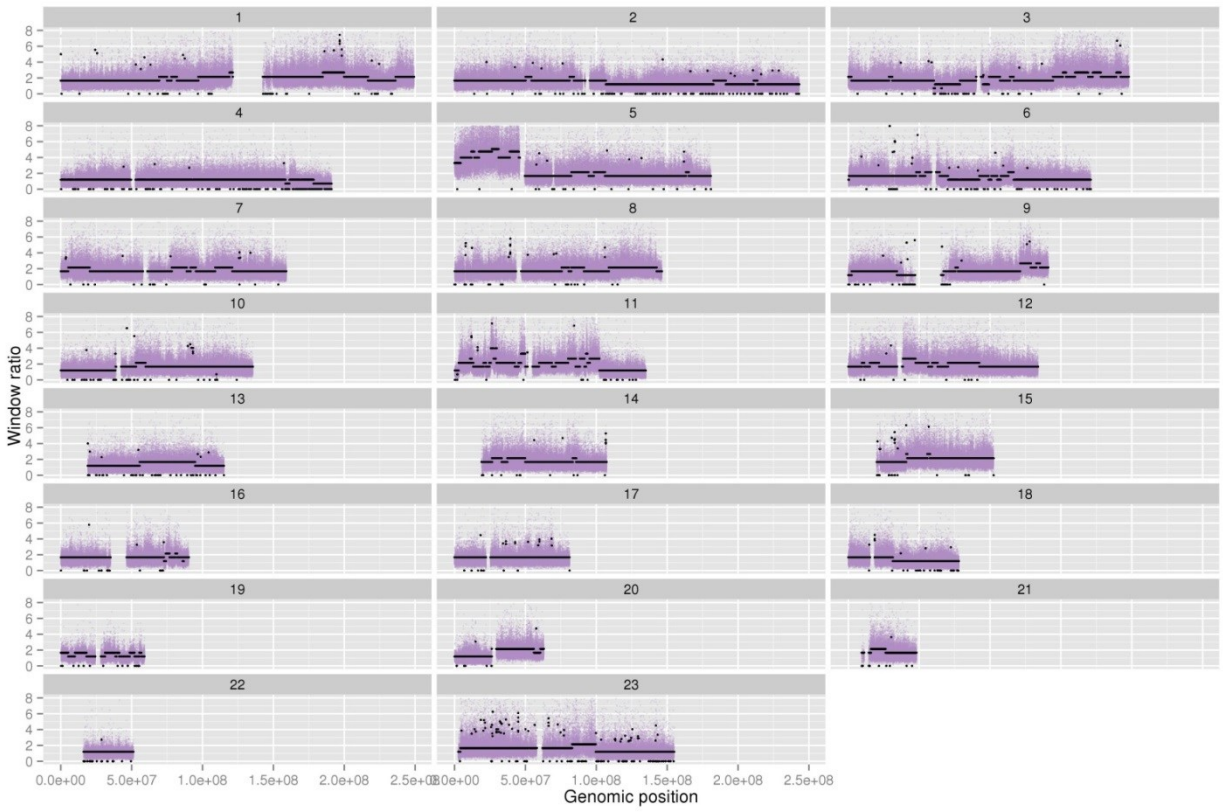


Figure D.4.24. HeLa S3 high resolution copy number calls.
Copy number ratios versus control genomes are plotted within high-resolution SUNK windows (green dots, each window size ~1.5 kbp), with predicted copy number state overlaid (black dots).

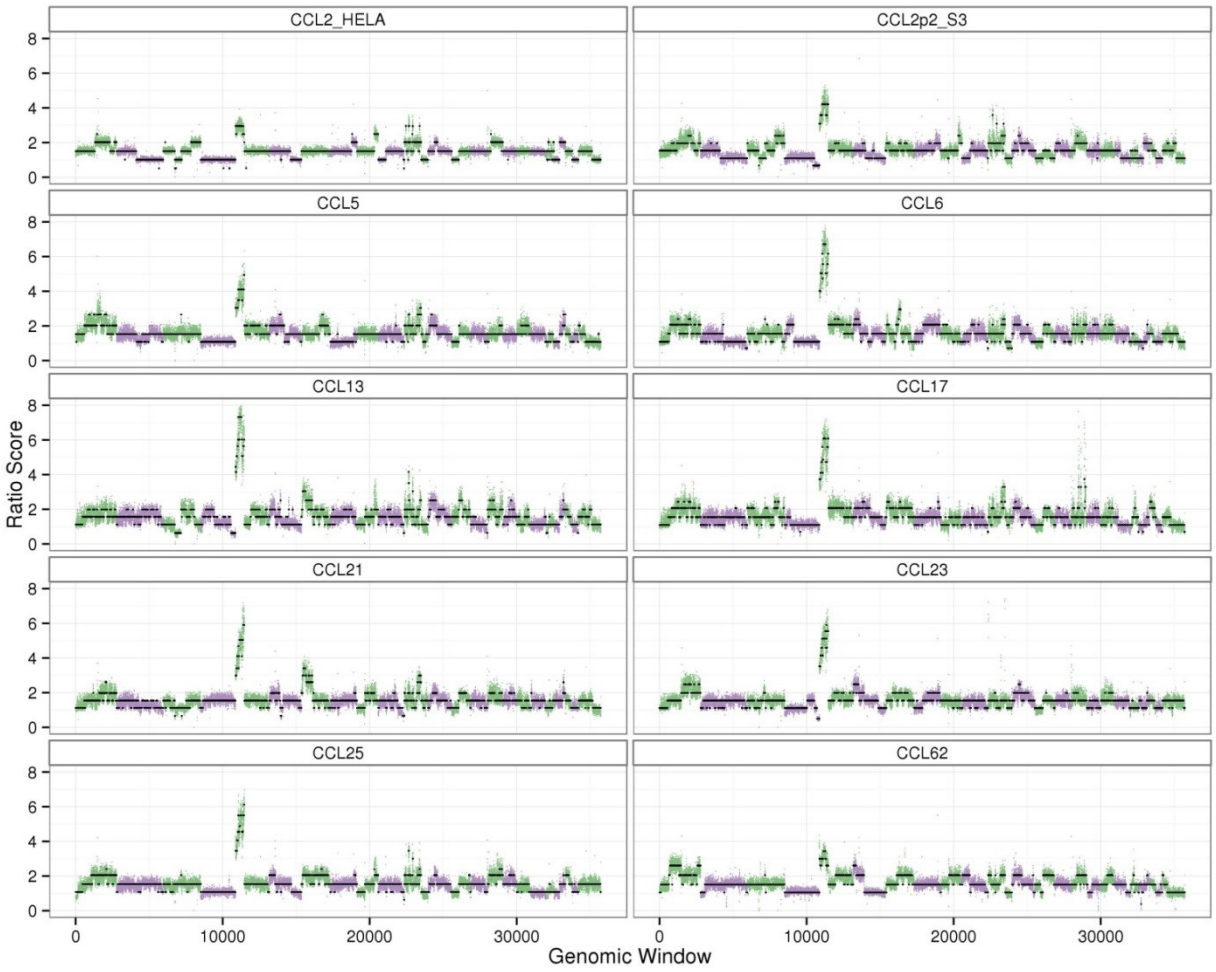


Figure D.4.25. Copy number profile for 10 HeLa strains.
Copy number ratios versus control genomes are plotted within high-resolution SUNK windows for HeLa CCL-2, HeLa S3, and eight additional HeLa strains (green and purple dots, alternating by chromosome, window contains 500 unique 30mers), with predicted copy number state overlaid (black dots).

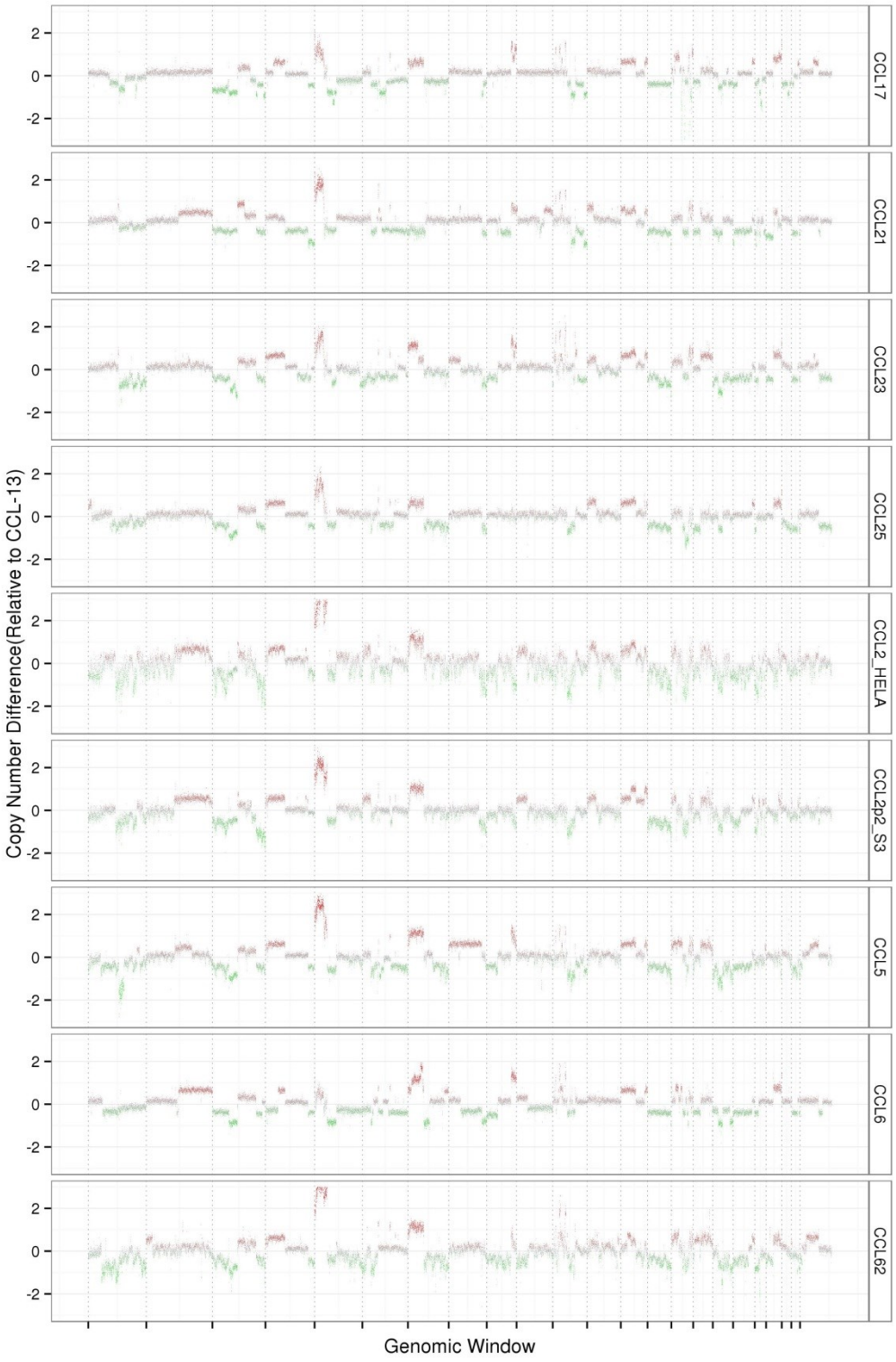


Figure D.4.26. Comparison of read depth profiles in HeLa strains.
Copy number differences across low-resolution SUNK windows relative to HeLa CCL-13 were plotted for HeLa CCL-2, S3, and 7 additional strains. Note: Increased values indicate increased copy number in CCL-13 compared to alternate strain.

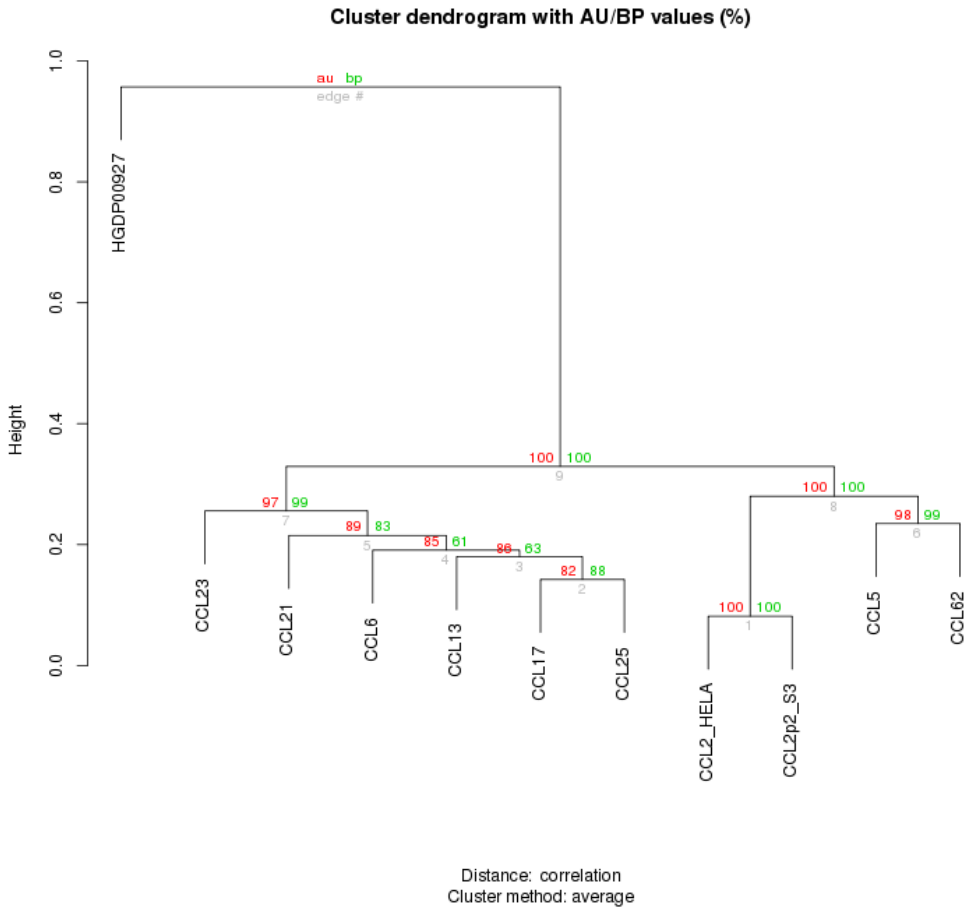


Figure D.4.27. Clustergram of 10 HeLa strains based on copy number profile similiarity.
Copy number scores were averaged within large windows (~1 Mbp) for 10 HeLa strains as well as an outgroup control genome (HGDP00927). Scores were clustered in (R package 'pvclust') with 1000 bootstrap iterations. “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.

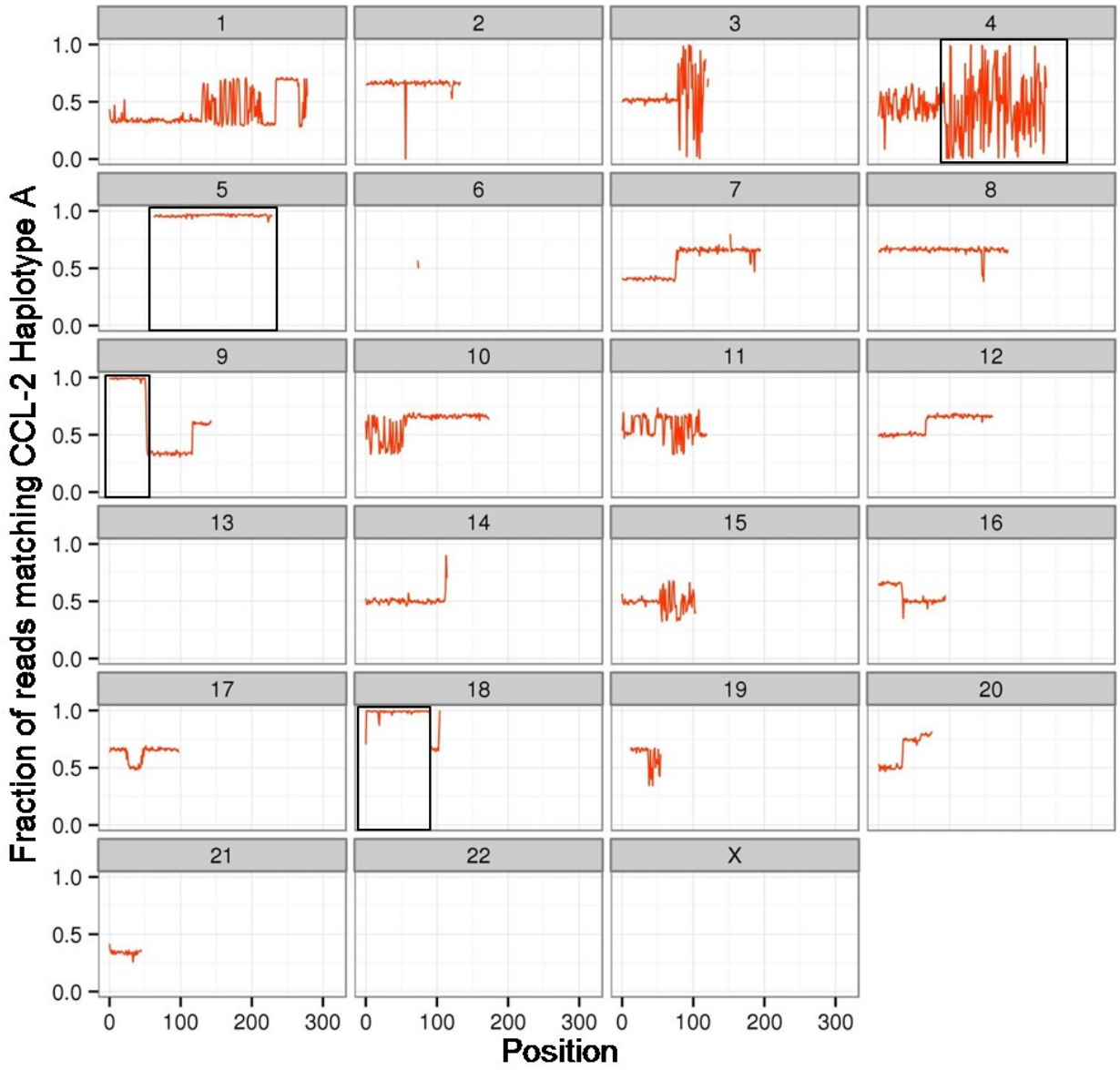


Figure D.4.28. Regions of LOH in HeLa CCL-13 by comparison to CCL-2 haplotypes.
Shown in windows (mean ~800 kbp) across each chromosome are the fraction of reads matching the allele phased to haplotype A in HeLa CCL-2. LOH in CCL-13 (but not CCL-2) manifests as long stretches where shotgun reads from 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.

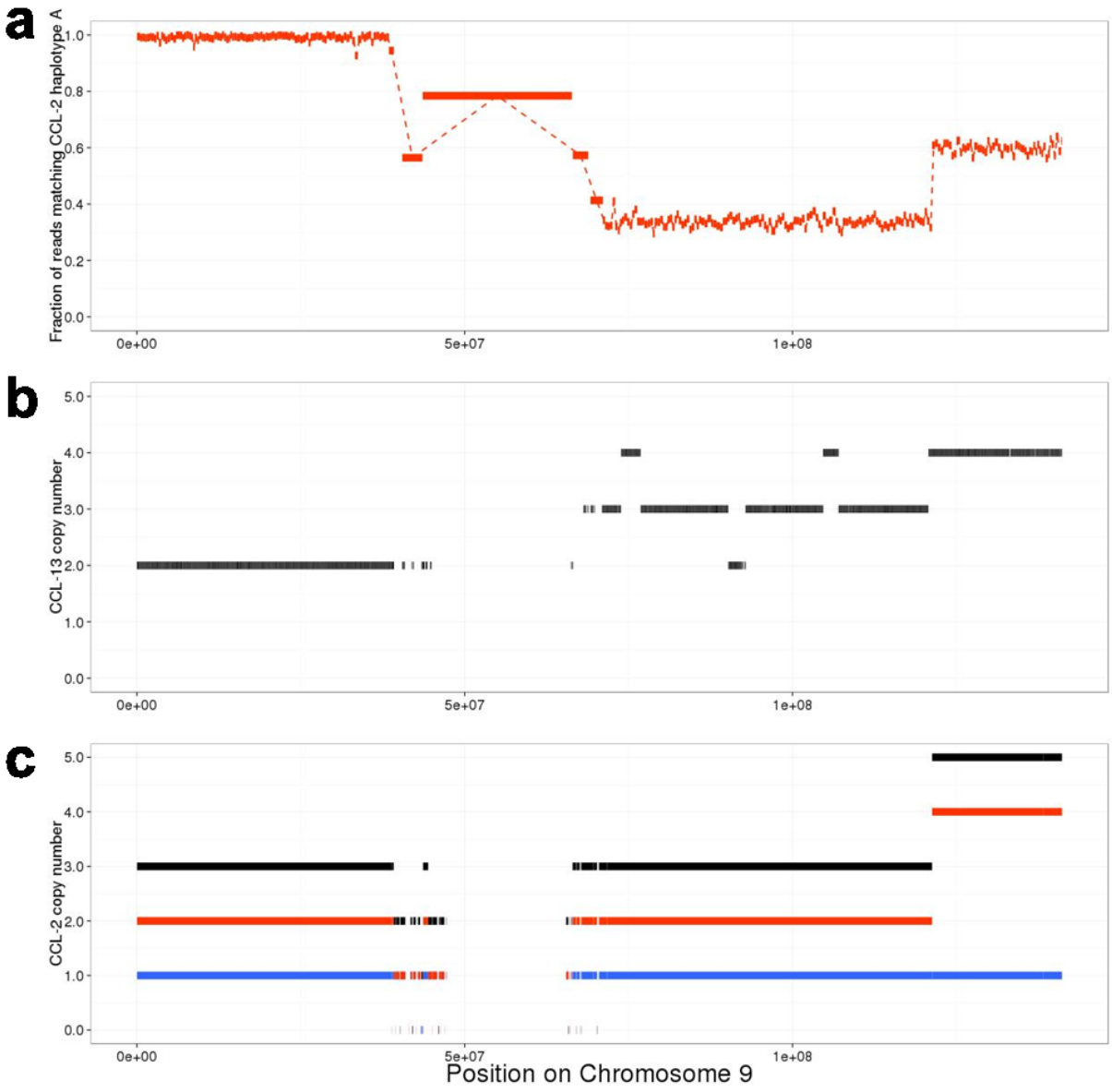
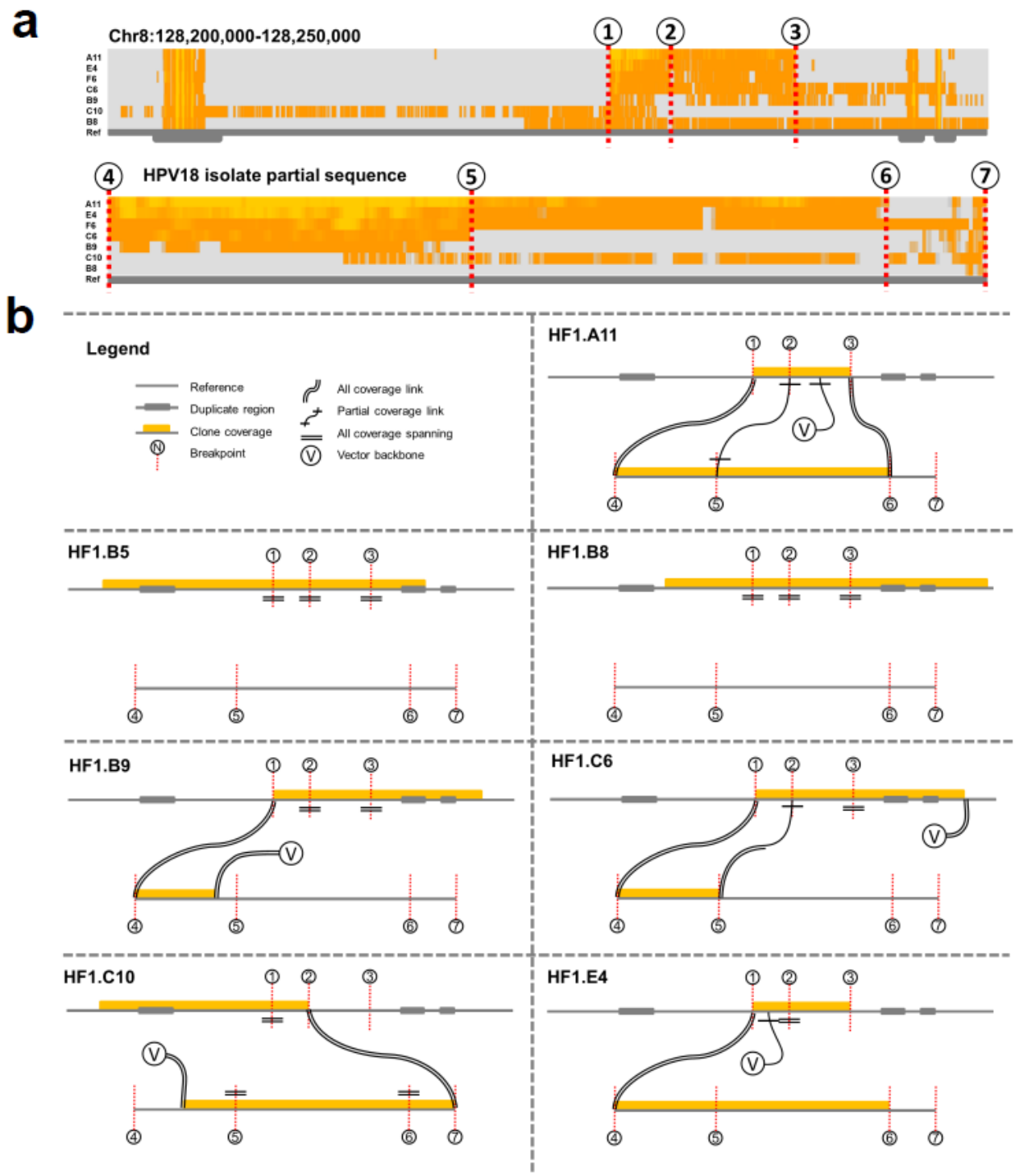
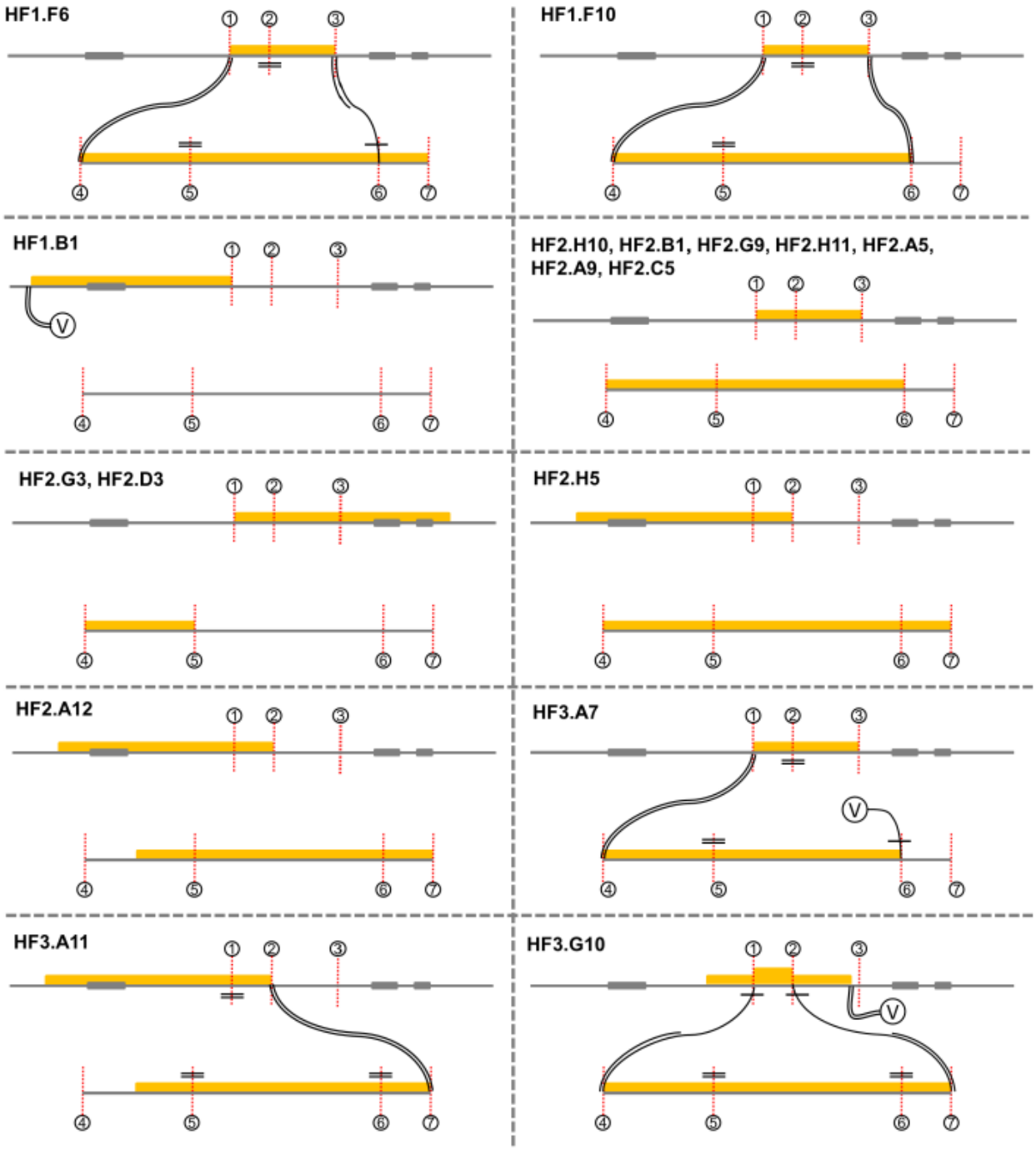


Figure D.4.29. Copy-number loss and LOH on chromosome 9 in HeLa CCL-13.
a. LOH on chromosome 9 in HeLa-CCL13, as detected by a shift towards CCL-2 haplotype A alleles was accompanied by reduction of copy number to 2 in CCL-13 in the affected region shown in b. relative to copy in HeLa CCL-2 shown in c. (Black = total copy number, Red = haplotype A copy number, Blue = haplotype B copy number).


a
Chr8:128,200,000-128,250,000
A11
E4
F6
C6
B9
C10
B8
Ref
HPV18 isolate partial sequence
b
Legend
Reference
Duplicate region
Clone coverage
Breakpoint
All coverage link
Partial coverage link
All coverage spanning
Vector backbone
HF1.A11
HF1.B5
HF1.B8
HF1.B9
HF1.C6
HF1.C10
HF1.E4


HF1.F6
HF1.F10
HF1.B1
HF2.H10, HF2.B1, HF2.G9, HF2.H11, HF2.A5, HF2.A9, HF2.C5
HF2.G3, HF2.D3
HF2.H5
HF2.A12
HF3.A7
HF3.A11
HF3.G10

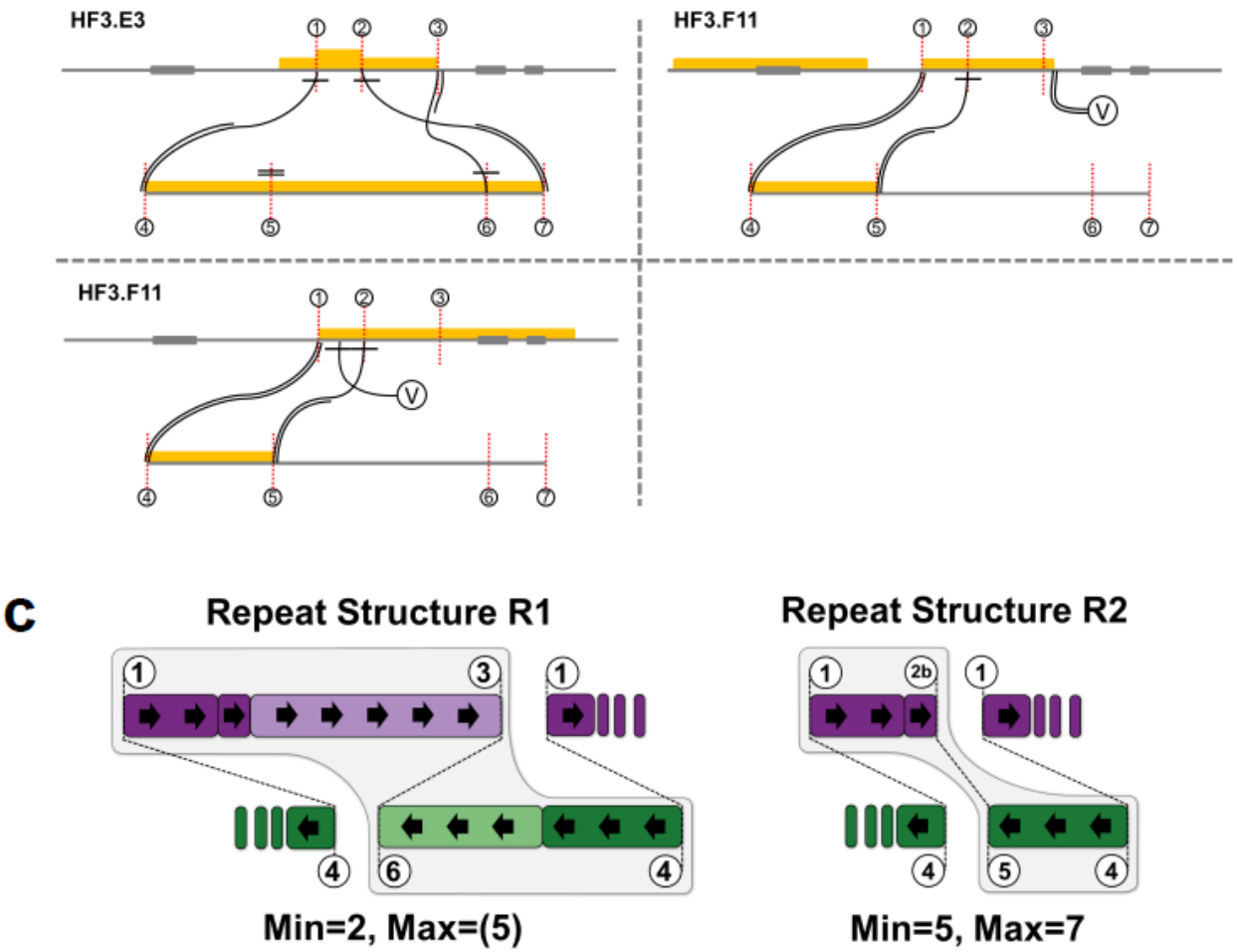
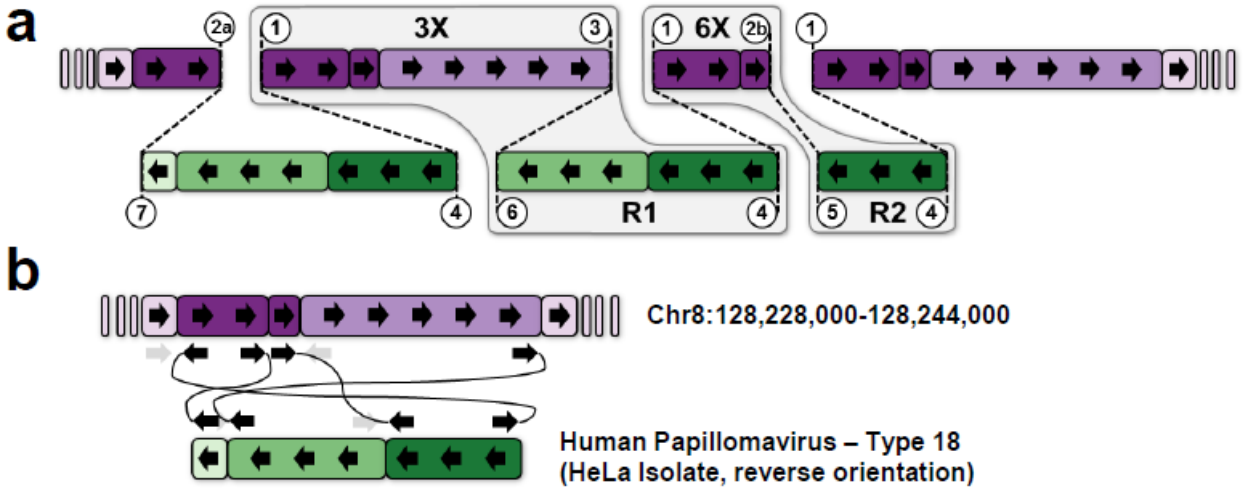


Figure D.4.30. Structure of the HPV-18 integration locus.
a. Heat maps showing coverage from fosmid clone pools across the insertion site flanking region (upper) or the HPV-18 partial sequence (lower). Circled numbers correspond to breakpoints in which discordantly mapping read pairs were found that 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 GGA 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 AGA 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 AGA 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 GGA 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 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 GGA TCA CAC CAC TGC ACT CCA GCC TGG GTG ACA GAG TGA GAC TCC CTC 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 D.4.31. 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.

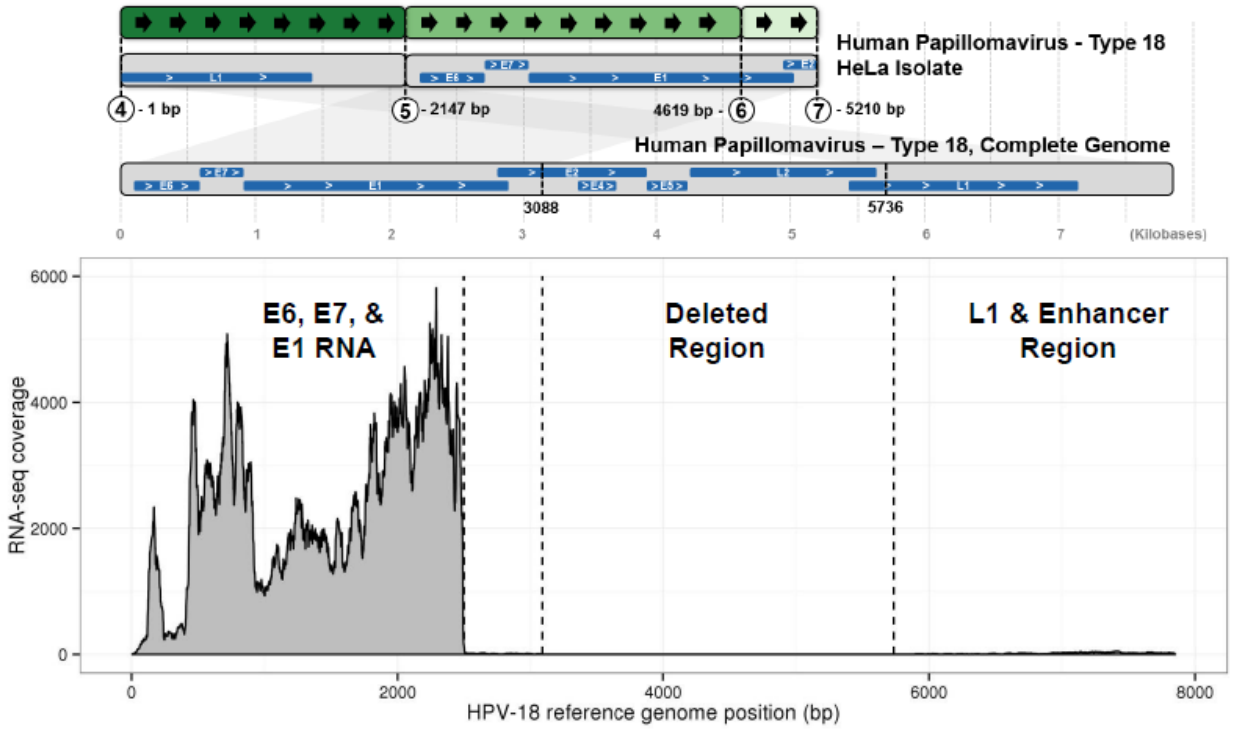


Figure D.4.32. HPV-18 RNA-Seq coverage.
Area chart (bottom panel) represents RNA-Seq level of coverage that reaches nearly 6,000 fold. Above the chart is the diagram of the HPV-18 portion of the integration locus on chromosome 8q24.21 from Figure 6.2 for reference.

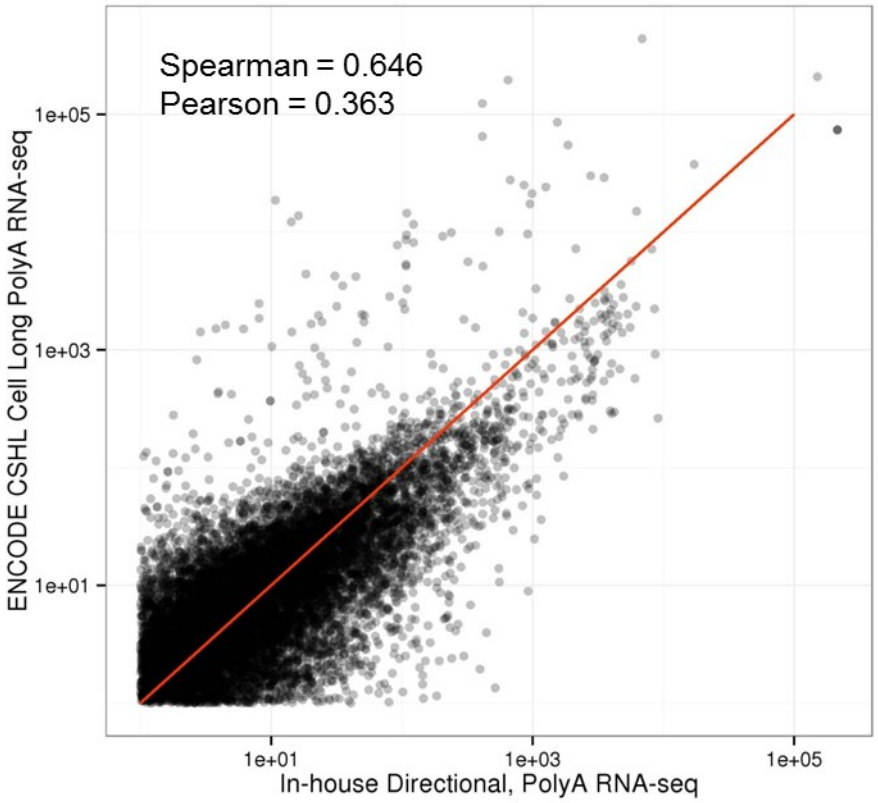


Figure D.4.33. Correlation between RNA-Seq data sets.
HeLa S3 transcript abundances (reads per kilobase per million reads, RPKM) from ENCODE RNA-Seq (Cold Spring Harbor – Cell long PolyA) were plotted against those our own RNA–Seq data. Each point represents one RefGene-annotated transcript (for transcripts with >= 1 RPKM). Red line is y=x.

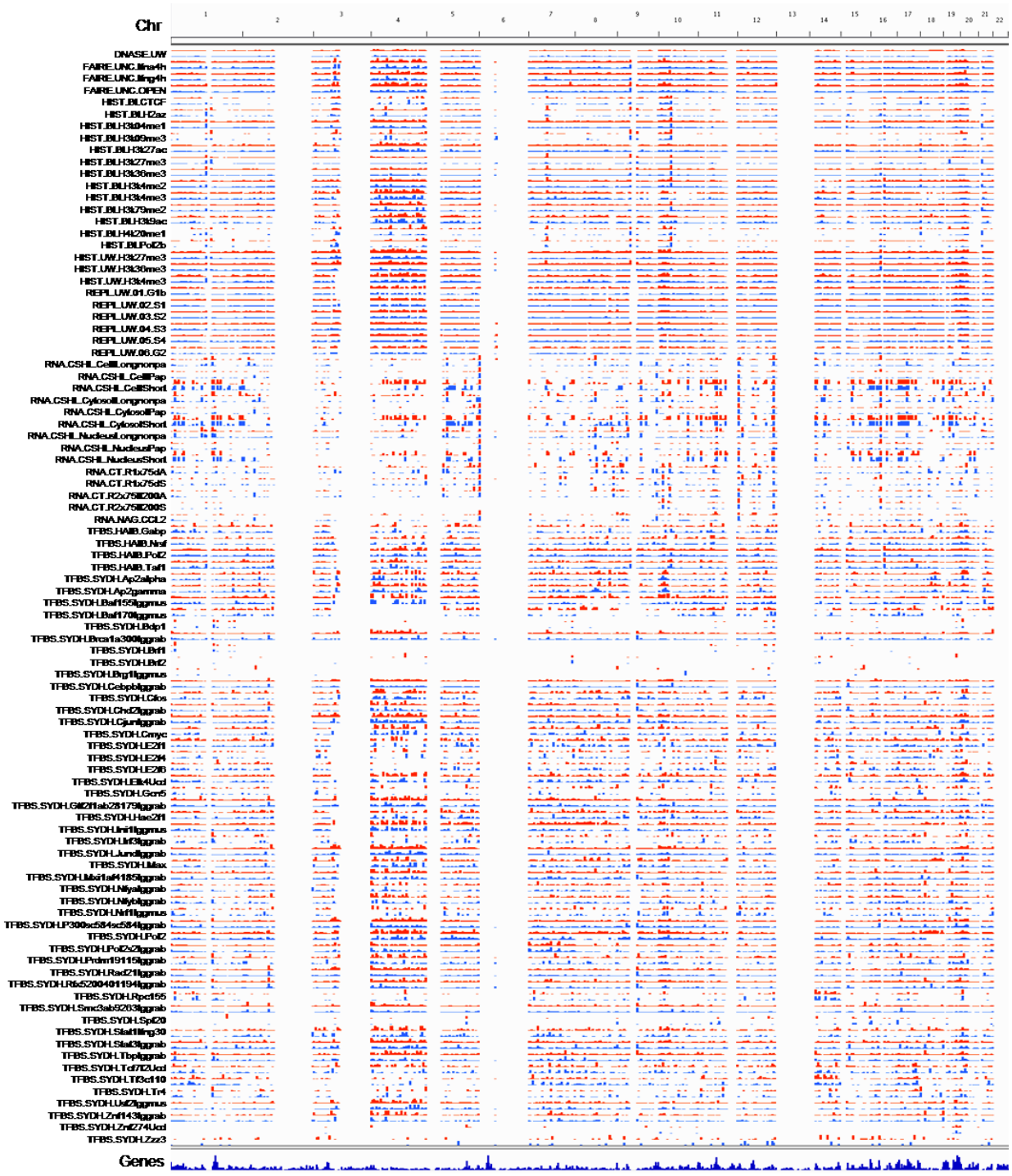


Figure D.4.34. Phased ENCODE data sets.
Per-haplotype enrichment scores for (A: red, B: blue) ENCODE HeLa S3 dataset. Height of each point indicates the degree of bias towards towards each haplotype after normalizing for its underlying copy number.

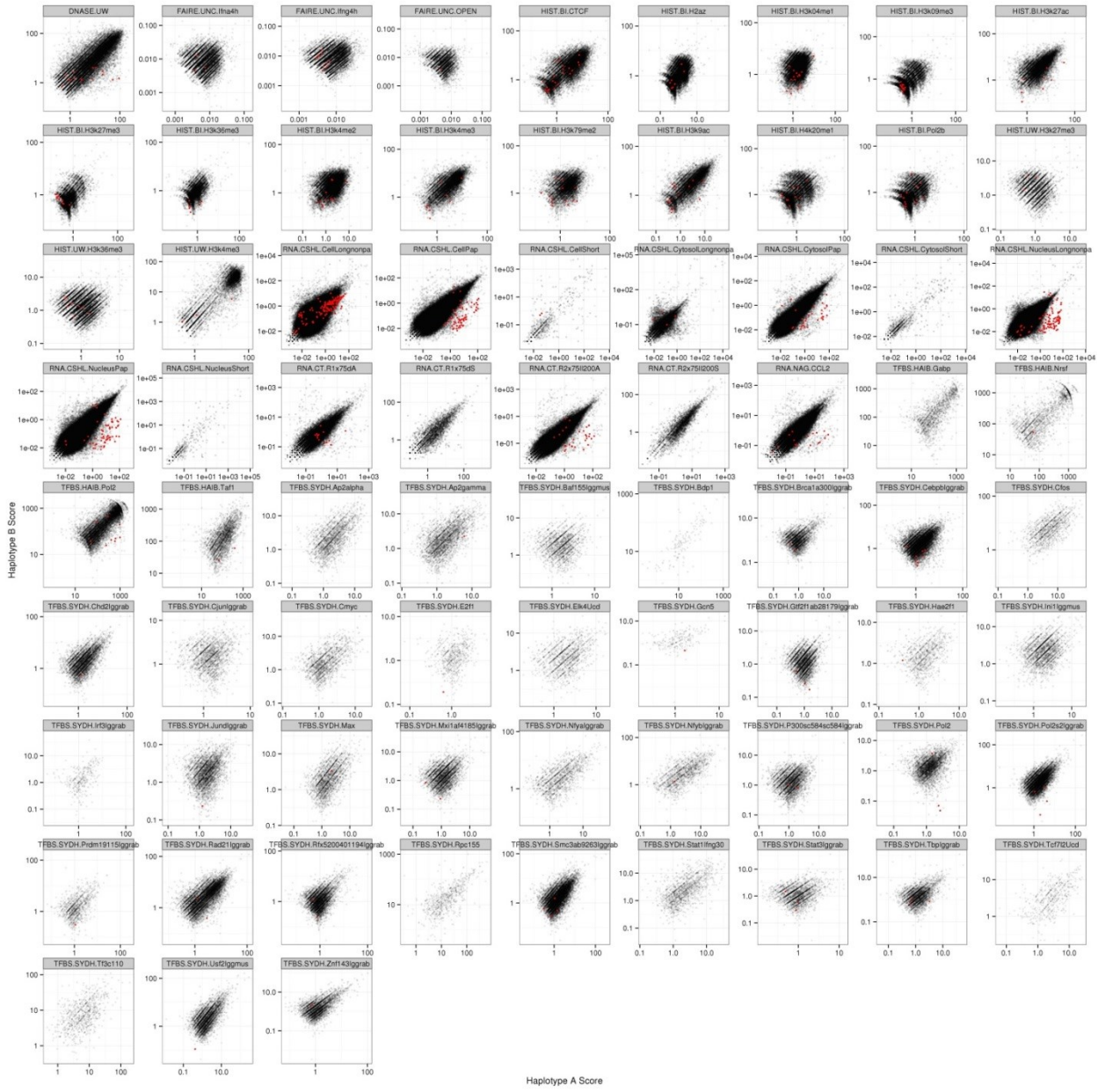


Figure D.4.35. Correlations between haplotype-specific signals for ENCODE HeLa datasets.
Copy number normalized haplotype B-specific signals are plotted against haplotype A-specific signals for ENCODE HeLa S3 datasets. Each point represents the mean haplotype-specific scores for called peaks (ChIP-seq and DNase-seq) or annotated transcripts (RNA-Seq). Peaks residing near the HPV integration site on chromosome 8q21.24 are represented by red points.

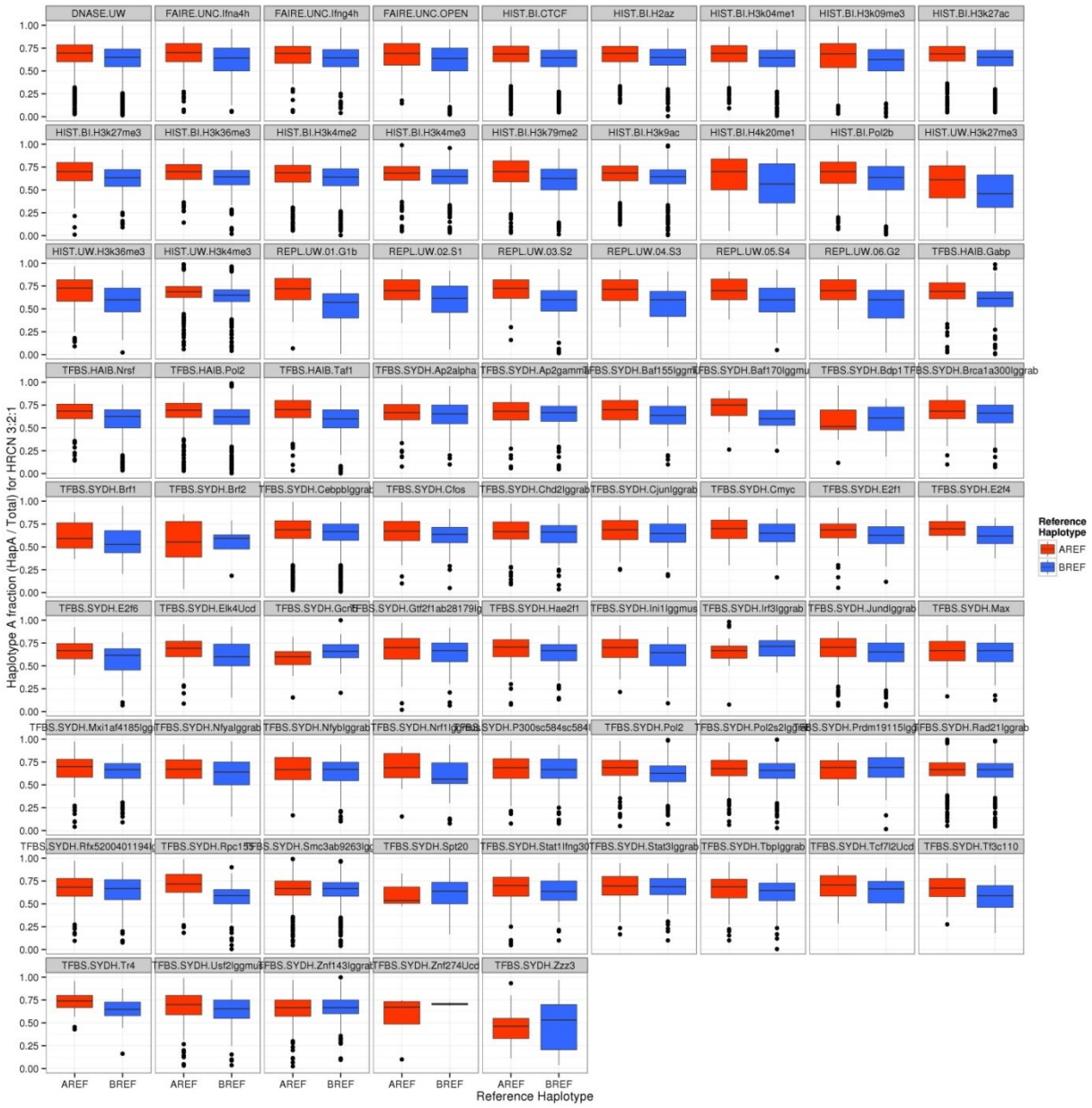


Figure D.4.36. Reference bias in ENCODE features.

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.

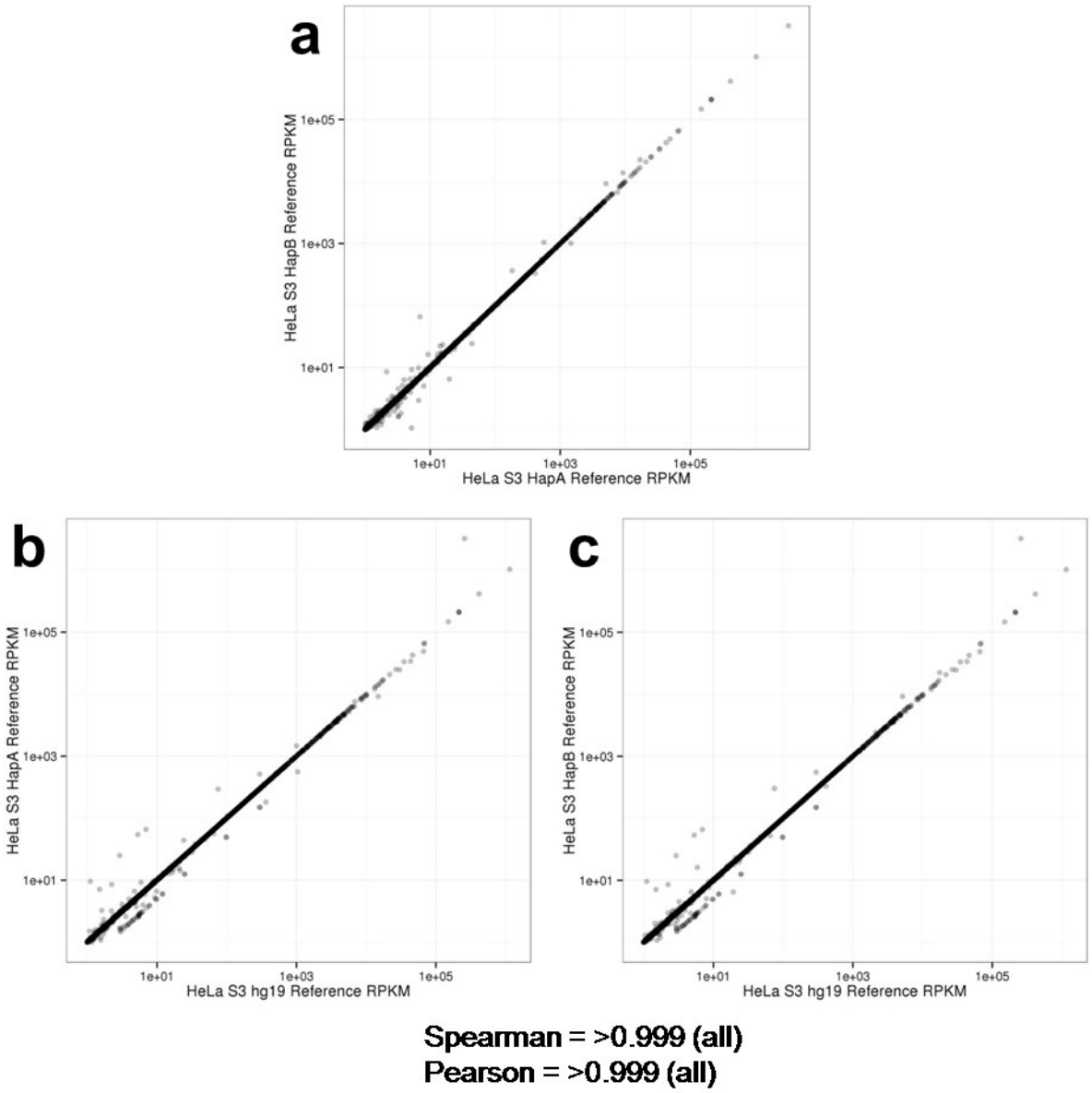
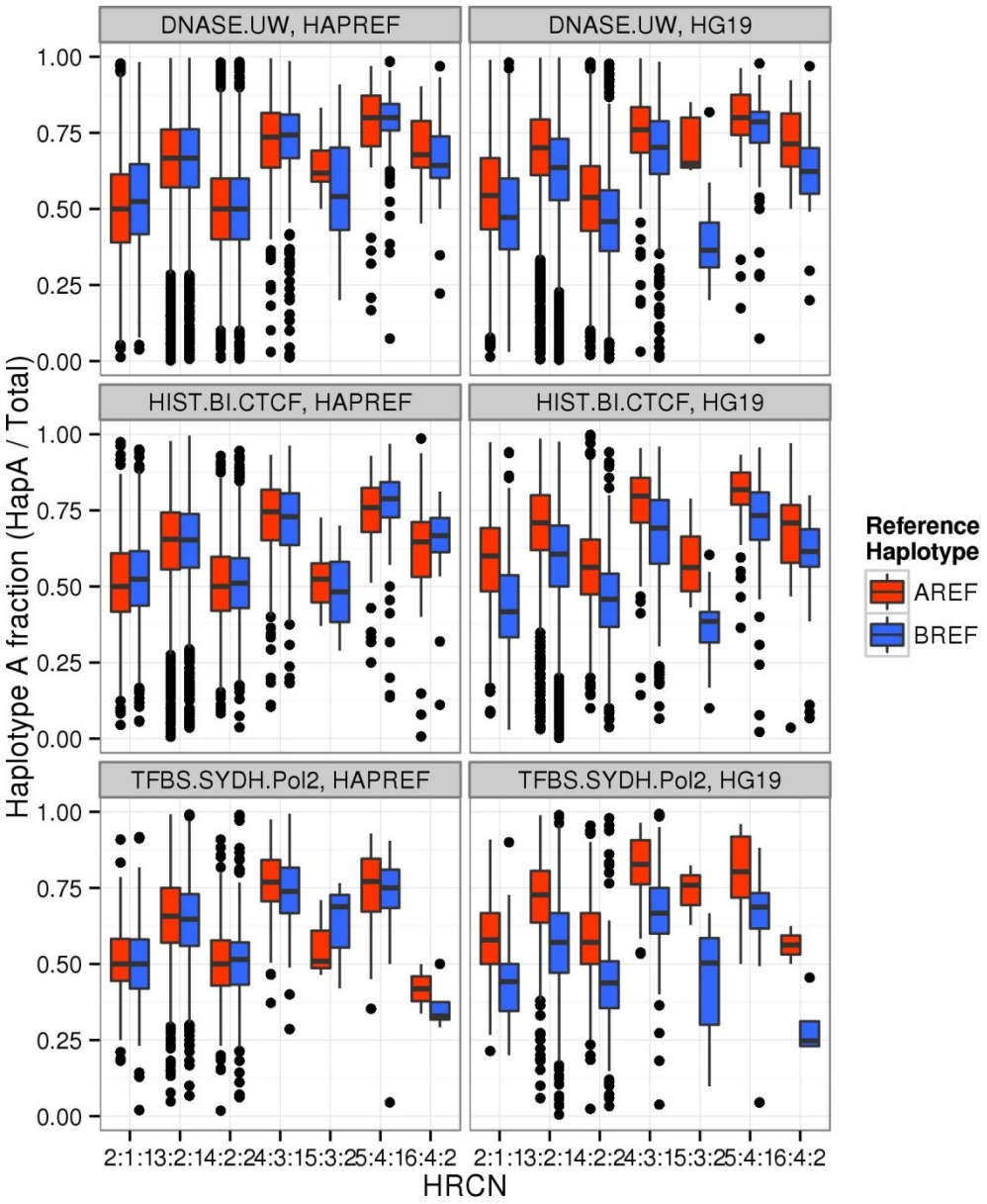


Figure D.4.37. Minimal impact of reference bias upon transcript quantitation.
HeLa S3 RNA-Seq reads (this study) were aligned using TopHat (Trapnell et. al. (2009)) to the reference genome ("hg19" (GRCh37)), as well as to HeLa haplotype-specific reference genomes ("HeLa Haplotype A" and "HeLa Haplotype B"). Transcript abundances were estimated against RefGene annotations using Cufflinks (Roberts et. al. (2011)) 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 (GRCh37) and HeLa Haplotype A reference (y-axis). c. Comparison between hg19 (GRCh37) and HeLa Haplotype B reference (y-axis).


DNASE.UW, HAPREF
DNASE.UW, HG19
HIST.BI.CTCF, HAPREF
HIST.BI.CTCF, HG19
TFBS.SYDH.Pol2, HAPREF
TFBS.SYDH.Pol2, HG19
Haplotype A fraction (HapA / Total)
1.00
0.75
0.50
0.25
0.00
2:1:13:2:14:2:24:3:15:3:25:4:16:4:2
HRCN
Reference Haplotype
AREF
BREF

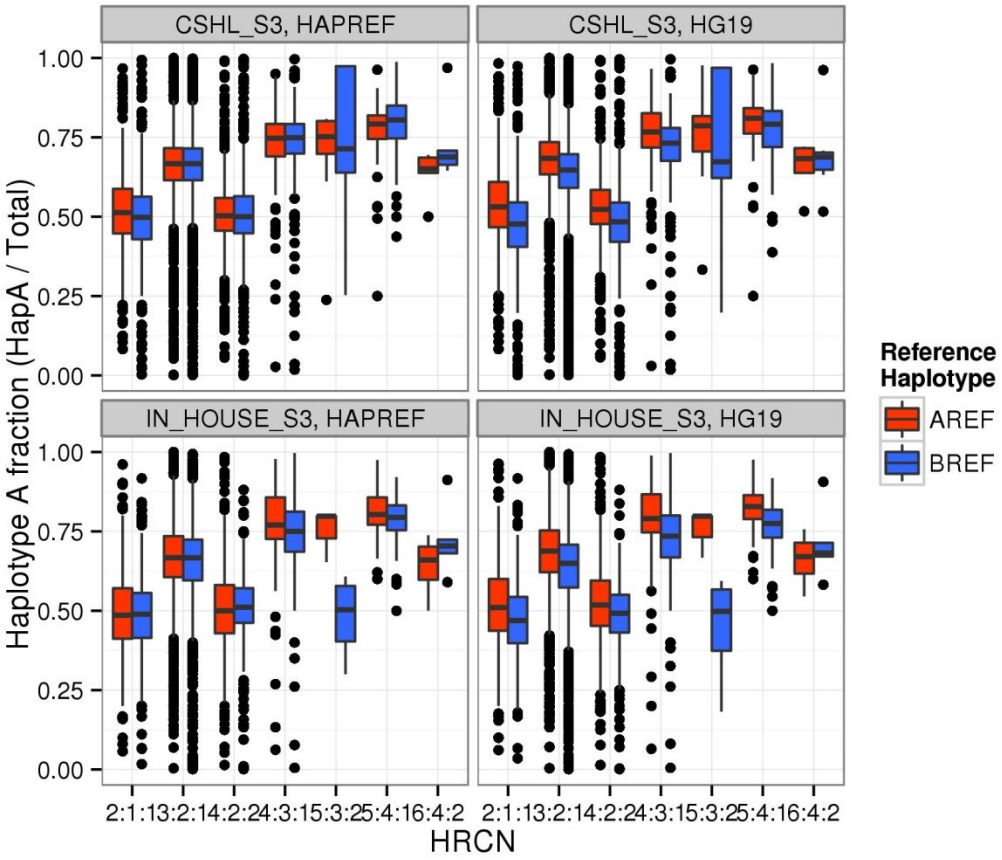


Figure D.4.38. Reference bias removal.
Reference haplotype imbalance for different HRCN classifications for in HeLa when aligning to a HeLa haplotype-resolved reference (left) or hg19 (GRCh37, right). a. Reference bias in ChIP-seq peaks. b. Reference bias in RNA-Seq. Red bars represent the haplotype 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.

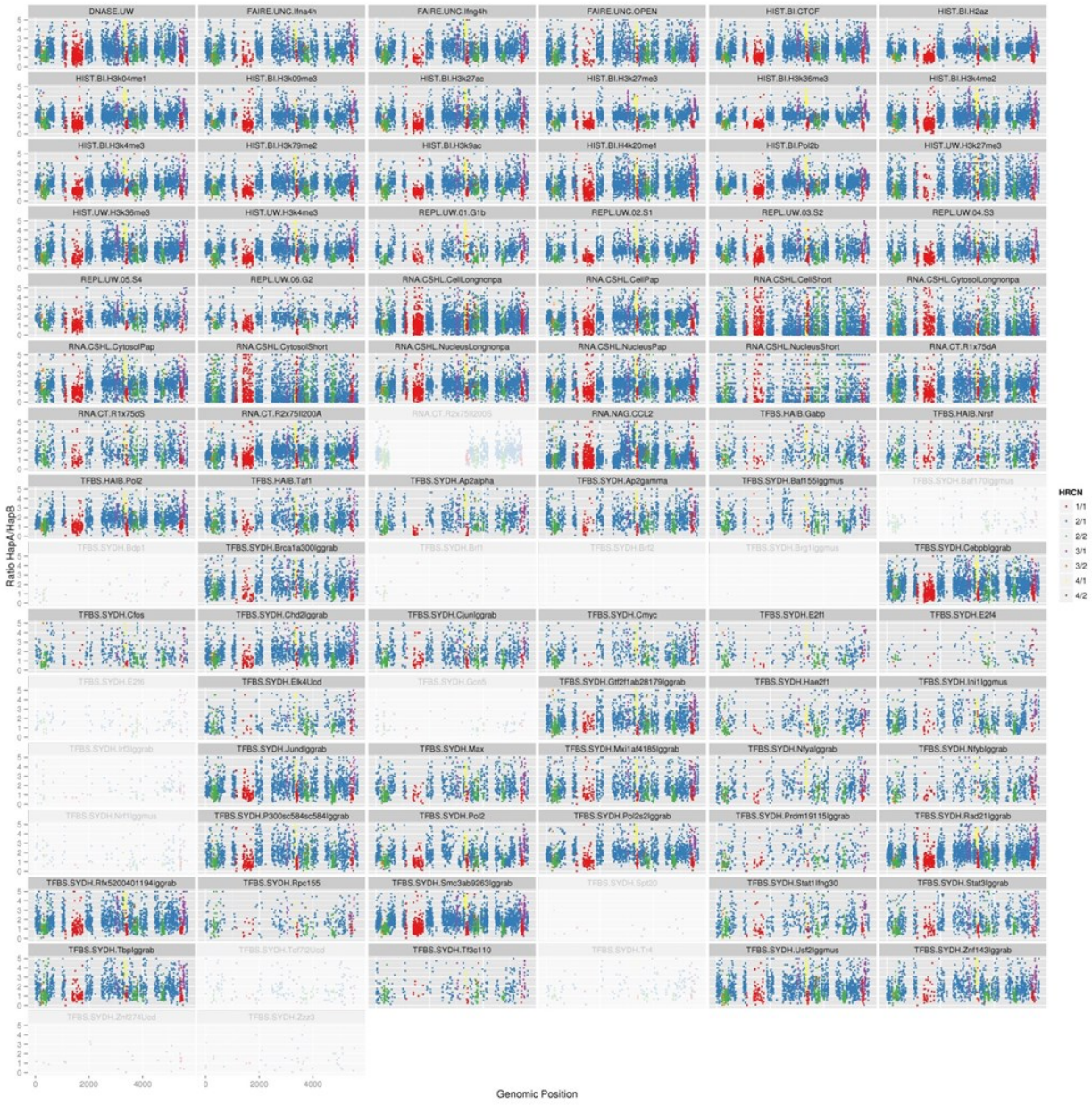


Figure D.4.39. Haplotype contributions of phased ENCODE data (windows).
Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows. Each window is color coded by the haplotype A to haplotype B ratio. Dimmed panels indicate data sets with very insufficient numbers of peaks for windowed analysis.

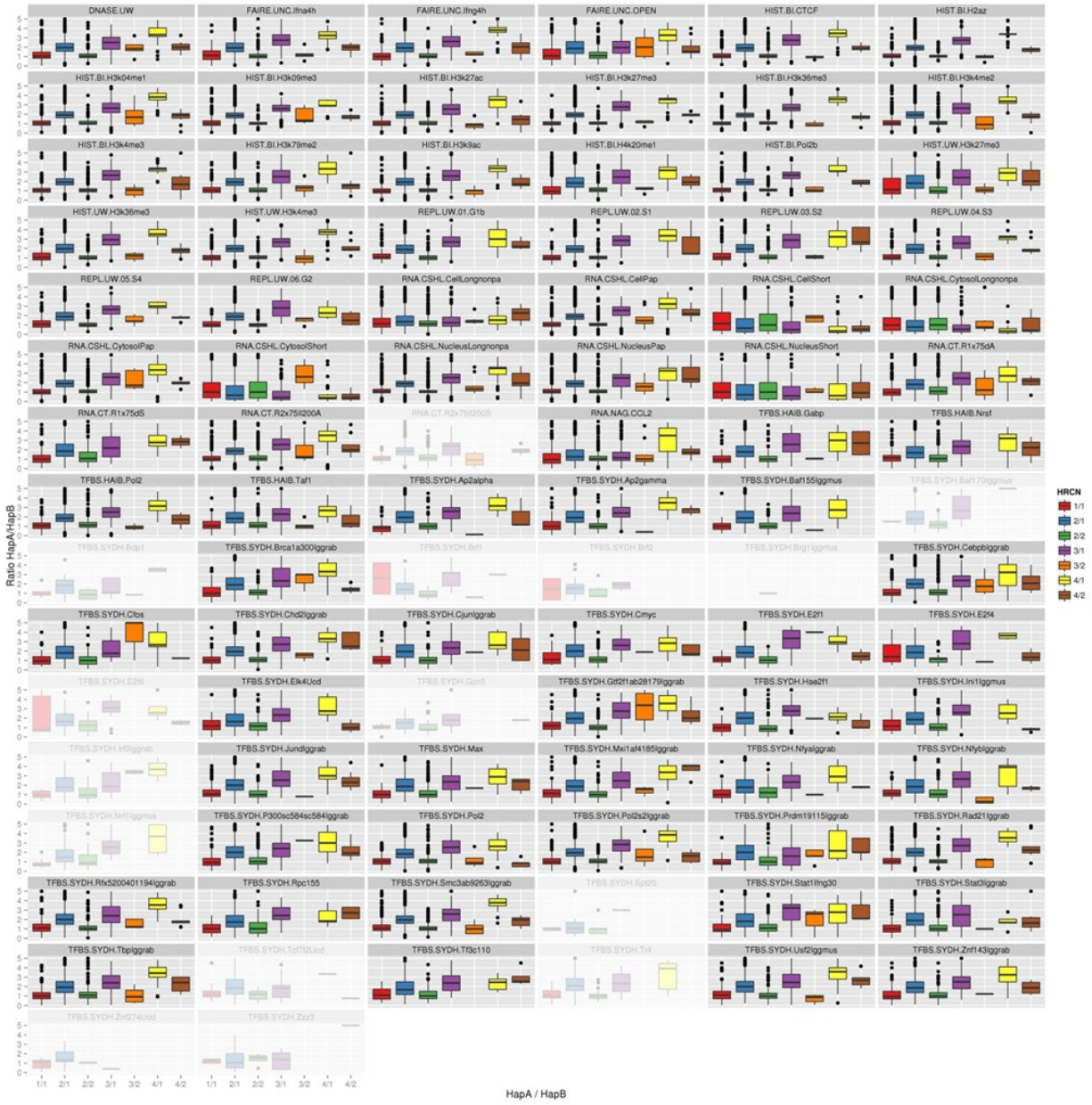


Figure D.4.40. Haplotype contributions of phased ENCODE data (box plots).
Haplotype ratios for a variety of ENCODE data tracks for haplotype A over haplotype B in 1.5 Mb sliding windows shown as box-and-whisker plots. Shaded out panels indicate data sets with very low peak counts and thus can not be reliably analyzed.

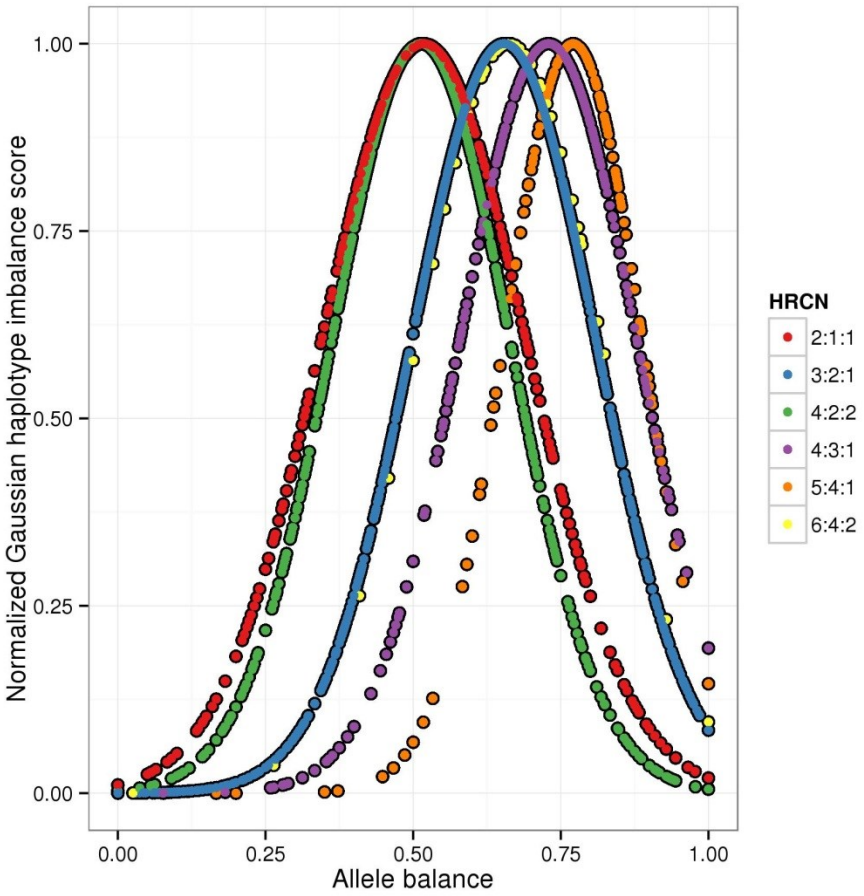
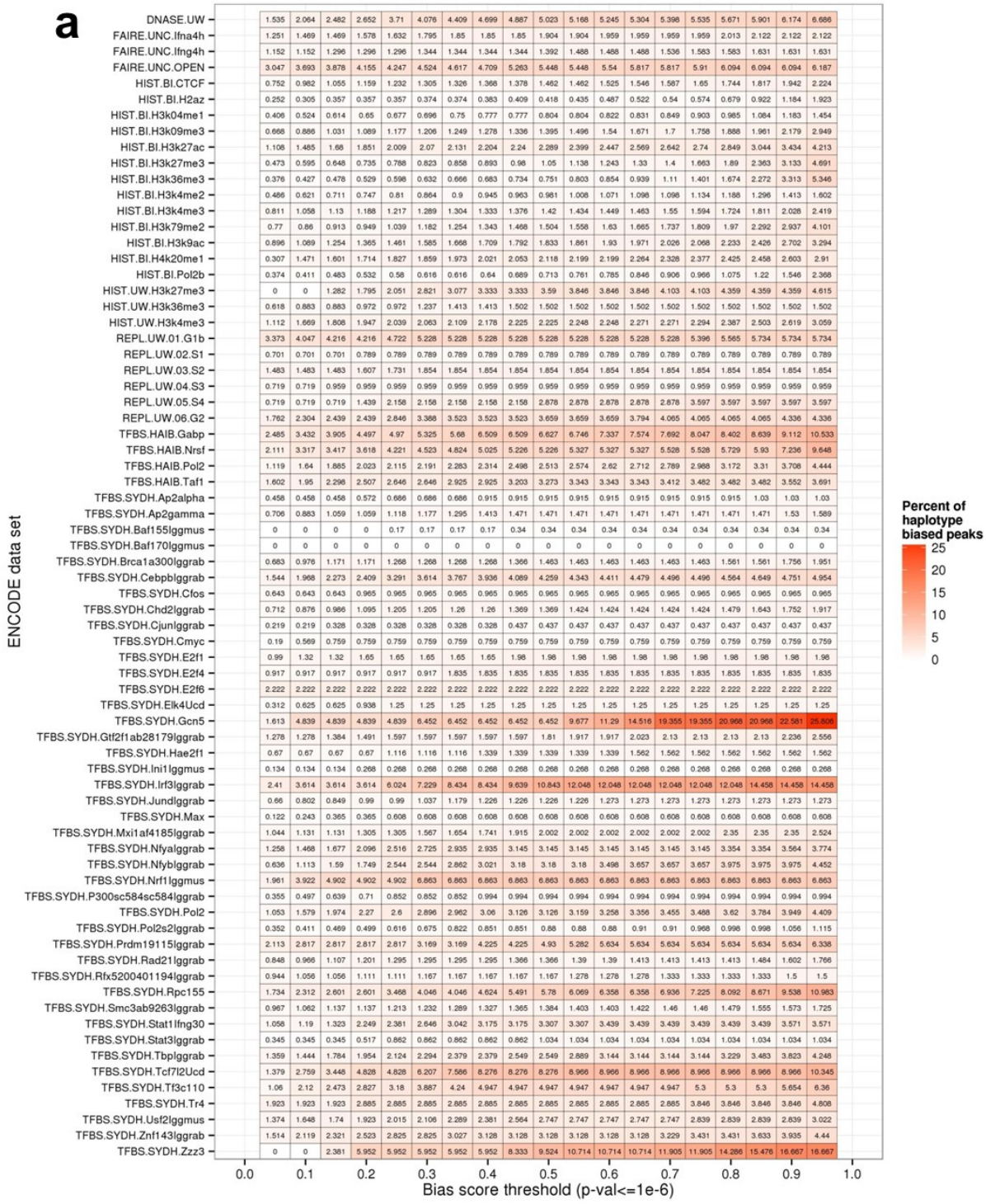
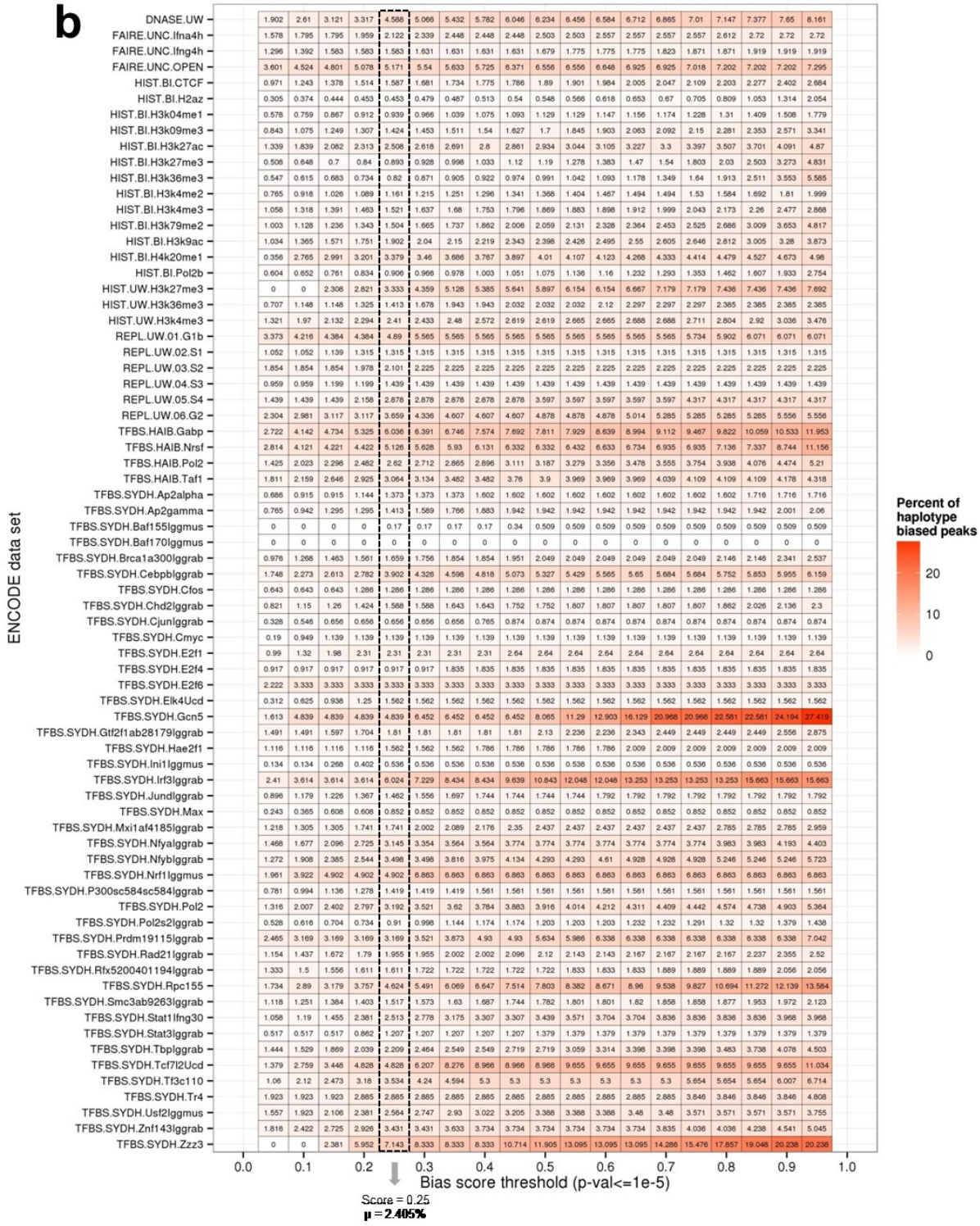


Figure D.4.41. Normalized haplotype imbalance scores by copy number.
Normalized haplotype imbalance scores were calculated and split by the underlying HRCN (total CN : hapA CN : hapB CN). The majority of the HeLa genome has a higher haplotype A copy number (as per naming conventions) and therefore expected 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).


a
ENCODE data set
DNASE.UW
FAIRE.UNC.Ifna4h
FAIRE.UNC.Ifng4h
FAIRE.UNC.OPEN
HIST.BI.CTCF
HIST.BI.H2az
HIST.BI.H3k04me1
HIST.BI.H3k09me3
HIST.BI.H3k27ac
HIST.BI.H3k27me3
HIST.BI.H3k36me3
HIST.BI.H3k4me2
HIST.BI.H3k4me3
HIST.BI.H3k79me2
HIST.BI.H3k9ac
HIST.BI.H4k20me1
HIST.BI.Pol2b
HIST.UW.H3k27me3
HIST.UW.H3k36me3
HIST.UW.H3k4me3
REPL.UW.01.G1b
REPL.UW.02.S1
REPL.UW.03.S2
REPL.UW.04.S3
REPL.UW.05.S4
REPL.UW.06.G2
TFBS.HAIB.Gabp
TFBS.HAIB.Nrsf
TFBS.HAIB.Pol2
TFBS.HAIB.Taf1
TFBS.SYDH.Ap2alpha
TFBS.SYDH.Ap2gamma
TFBS.SYDH.Baf155Iggmus
TFBS.SYDH.Baf170Iggmus
TFBS.SYDH.Brca1a300Iggrab
TFBS.SYDH.CebpbIggrab
TFBS.SYDH.Cfos
TFBS.SYDH.Chd2Iggrab
TFBS.SYDH.CjunIggrab
TFBS.SYDH.Cmyc
TFBS.SYDH.E2f1
TFBS.SYDH.E2f4
TFBS.SYDH.E2f6
TFBS.SYDH.Elk4Ucd
TFBS.SYDH.Gcn5
TFBS.SYDH.Gtf2f1ab28179Iggrab
TFBS.SYDH.Hae2f1
TFBS.SYDH.Ini1Iggmus
TFBS.SYDH.Irf3Iggrab
TFBS.SYDH.JundIggrab
TFBS.SYDH.Max
TFBS.SYDH.Mxi1af4185Iggrab
TFBS.SYDH.NfyaIggrab
TFBS.SYDH.NfybIggrab
TFBS.SYDH.Nrf1Iggmus
TFBS.SYDH.P300sc584sc584Iggrab
TFBS.SYDH.Pol2
TFBS.SYDH.Pol2s2Iggrab
TFBS.SYDH.Prdm19115Iggrab
TFBS.SYDH.Rad21Iggrab
TFBS.SYDH.Rfx5200401194Iggrab
TFBS.SYDH.Rpc155
TFBS.SYDH.Smc3ab9263Iggrab
TFBS.SYDH.Stat1Ifng30
TFBS.SYDH.Stat3Iggrab
TFBS.SYDH.TbpIggrab
TFBS.SYDH.Tcf7l2Ucd
TFBS.SYDH.Tf3c110
TFBS.SYDH.Tr4
TFBS.SYDH.Usf2Iggmus
TFBS.SYDH.Znf143Iggrab
TFBS.SYDH.Zzz3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Bias score threshold (p-val<=1e-6)
Percent of haplotype biased peaks
25
20
15
10
5
0


b
ENCODE data set
DNASE.UW
FAIRE.UNC.Ifna4h
FAIRE.UNC.Ifng4h
FAIRE.UNC.OPEN
HIST.BI.CTCF
HIST.BI.H2az
HIST.BI.H3k04me1
HIST.BI.H3k09me3
HIST.BI.H3k27ac
HIST.BI.H3k27me3
HIST.BI.H3k36me3
HIST.BI.H3k4me2
HIST.BI.H3k4me3
HIST.BI.H3k79me2
HIST.BI.H3k9ac
HIST.BI.H4k20me1
HIST.BI.Pol2b
HIST.UW.H3k27me3
HIST.UW.H3k36me3
HIST.UW.H3k4me3
REPL.UW.01.G1b
REPL.UW.02.S1
REPL.UW.03.S2
REPL.UW.04.S3
REPL.UW.05.S4
REPL.UW.06.G2
TFBS.HAIB.Gabp
TFBS.HAIB.Nrsf
TFBS.HAIB.Pol2
TFBS.HAIB.Taf1
TFBS.SYDH.Ap2alpha
TFBS.SYDH.Ap2gamma
TFBS.SYDH.Baf155Iggmus
TFBS.SYDH.Baf170Iggmus
TFBS.SYDH.Brca1a300Iggrab
TFBS.SYDH.CebpbIggrab
TFBS.SYDH.Cfos
TFBS.SYDH.Chd2Iggrab
TFBS.SYDH.CjunIggrab
TFBS.SYDH.Cmyc
TFBS.SYDH.E2f1
TFBS.SYDH.E2f4
TFBS.SYDH.E2f6
TFBS.SYDH.Elk4Ucd
TFBS.SYDH.Gcn5
TFBS.SYDH.Gtf2f1ab28179Iggrab
TFBS.SYDH.Hae2f1
TFBS.SYDH.Ini1Iggmus
TFBS.SYDH.Irf3Iggrab
TFBS.SYDH.JundIggrab
TFBS.SYDH.Max
TFBS.SYDH.Mxi1af4185Iggrab
TFBS.SYDH.NfyaIggrab
TFBS.SYDH.NfybIggrab
TFBS.SYDH.Nrf1Iggmus
TFBS.SYDH.P300sc584sc584Iggrab
TFBS.SYDH.Pol2
TFBS.SYDH.Pol2s2Iggrab
TFBS.SYDH.Prdm19115Iggrab
TFBS.SYDH.Rad21Iggrab
TFBS.SYDH.Rfx5200401194Iggrab
TFBS.SYDH.Rpc155
TFBS.SYDH.Smc3ab9263Iggrab
TFBS.SYDH.Stat1Ifng30
TFBS.SYDH.Stat3Iggrab
TFBS.SYDH.TbpIggrab
TFBS.SYDH.Tcf7l2Ucd
TFBS.SYDH.Tf3c110
TFBS.SYDH.Tr4
TFBS.SYDH.Usf2Iggmus
TFBS.SYDH.Znf143Iggrab
TFBS.SYDH.Zzz3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Bias score threshold (p-val<=1e-5)
Score = 0.25
μ = 2.405%
Percent of haplotype biased peaks
20
10
0

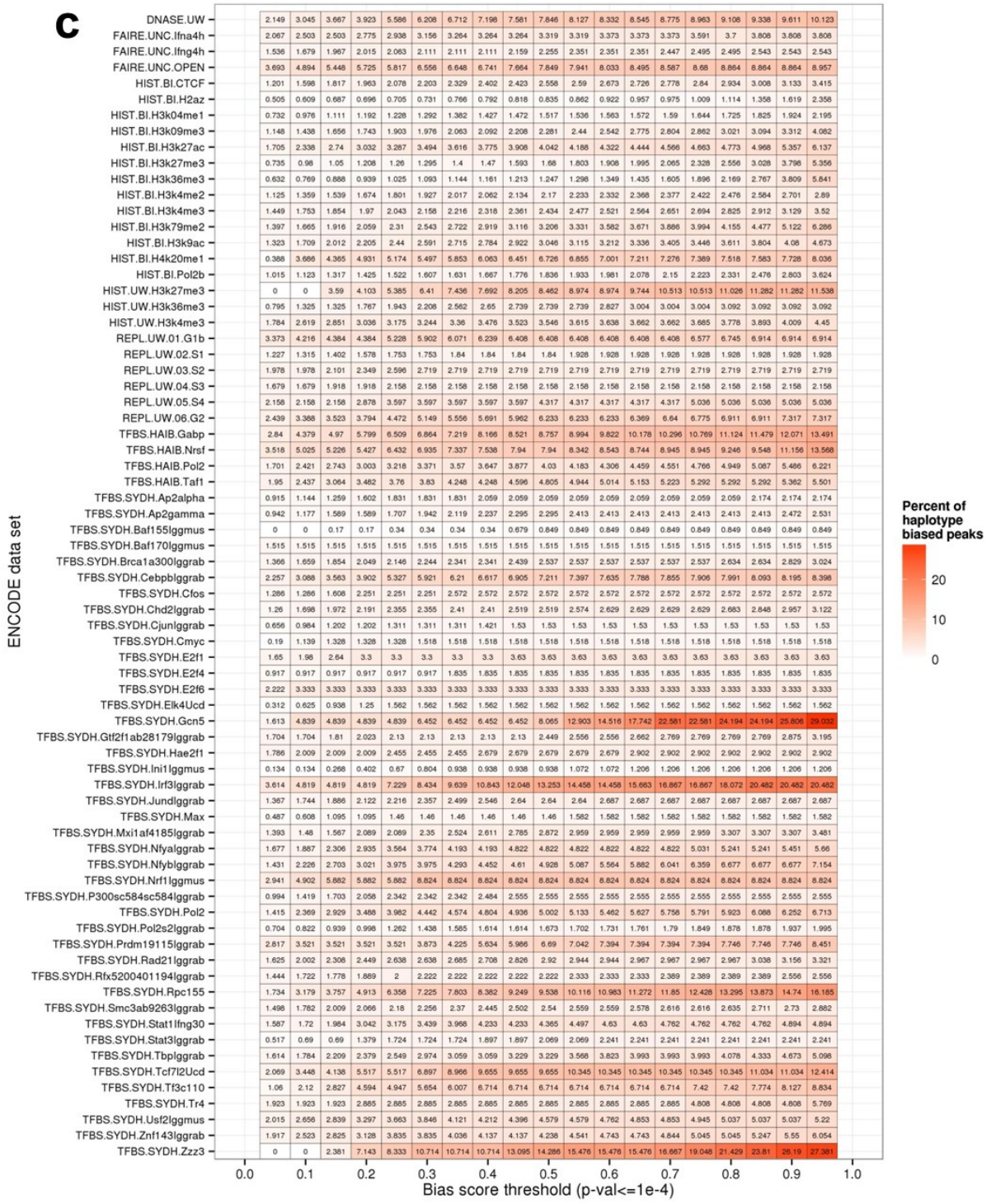


Figure D.4.42. Haplotype imbalanced ENCODE peak percentages.
Percentages of peaks within each ENCODE enrichment data set called as outliers at three thresholds (a. P<1e-6, b. P<1e-5, and c. P<1e-4) with respect to normalized Gaussian haplotype imbalance score. The dashed box in b. represents the scoring threshold used of a p-value of 1e-5 and normalized imbalance score of 0.25.

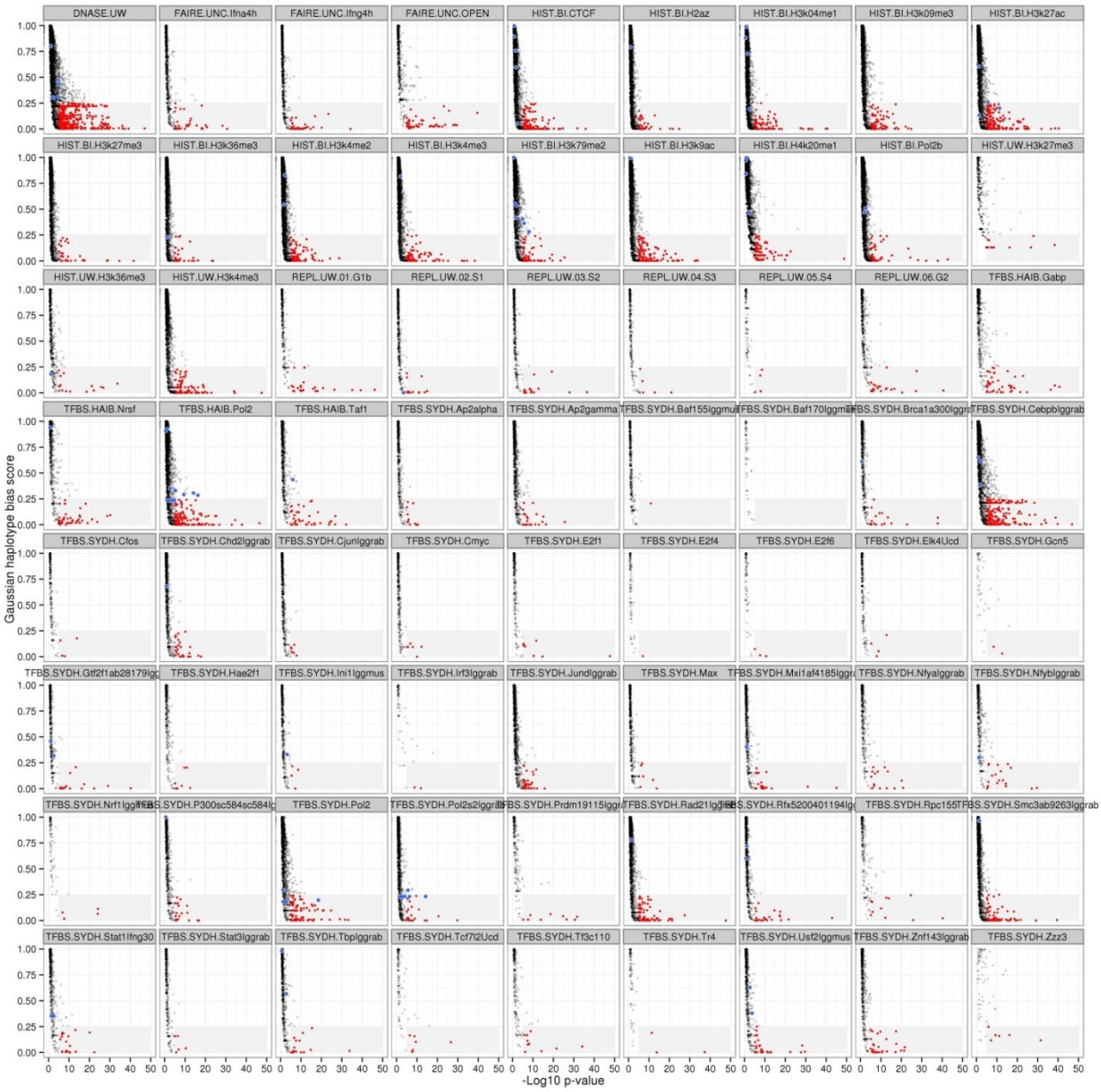


Figure D.4.43. ENCODE peak haplotype imbalance scoring.
For each peak with an ENCODE data track, the normalized haplotype imbalance score is plotted against the $-\log_{10}$ *p*-value (the degree of significance against the null hypothesis of haplotype-balanced signal). Gray boxes with red points represent 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.

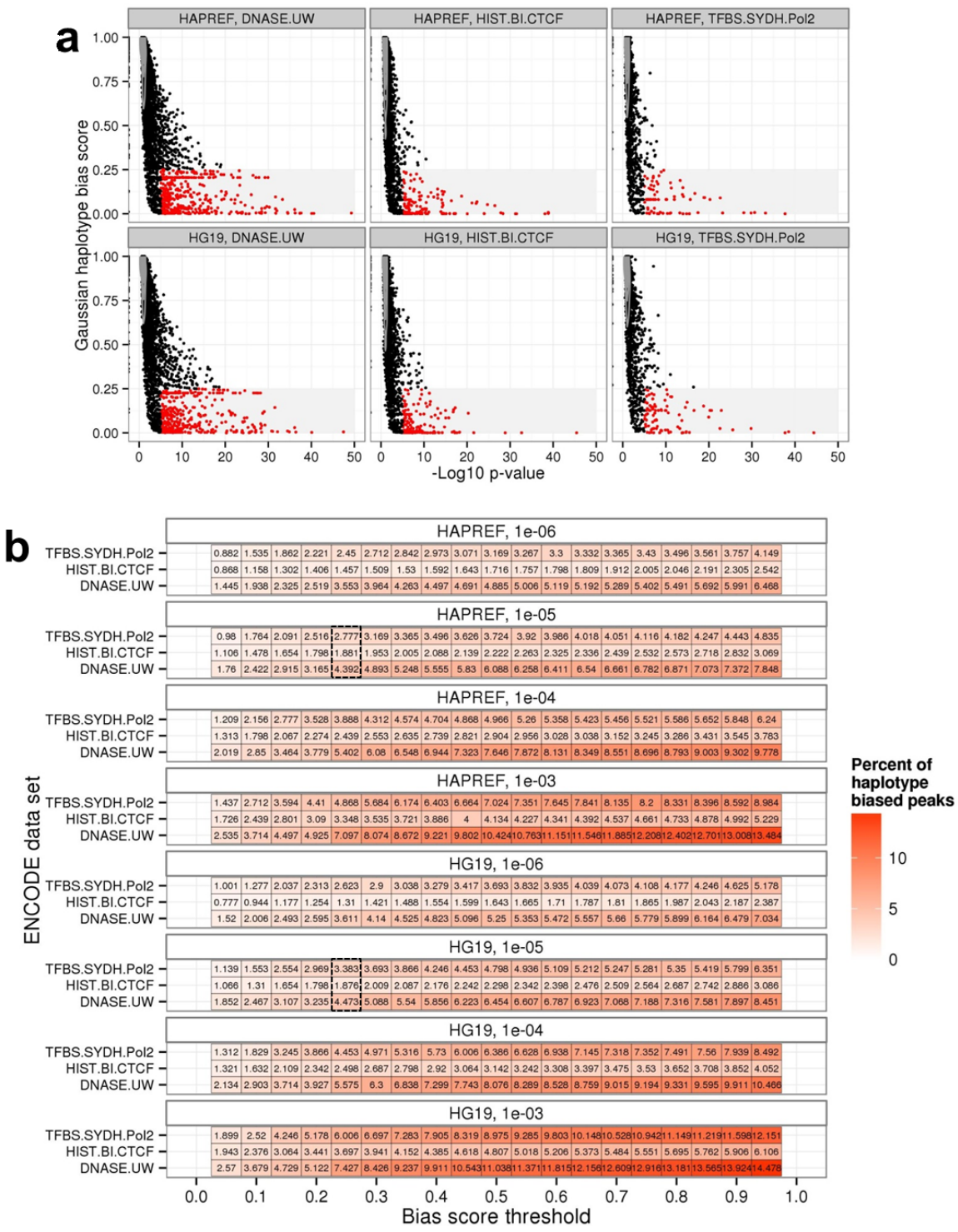


Figure D.4.44. ENCODE peak reference bias effects on outlier calling.
The use of a HeLa-specific haplotype-resolved reference eliminates the reference bias, but does not substantially change the set of peaks called as outliers. a. Haplotype imbalance scores when aligning to a haplotype-resolved HeLa reference (top) or hg19 (GRCh37, 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.

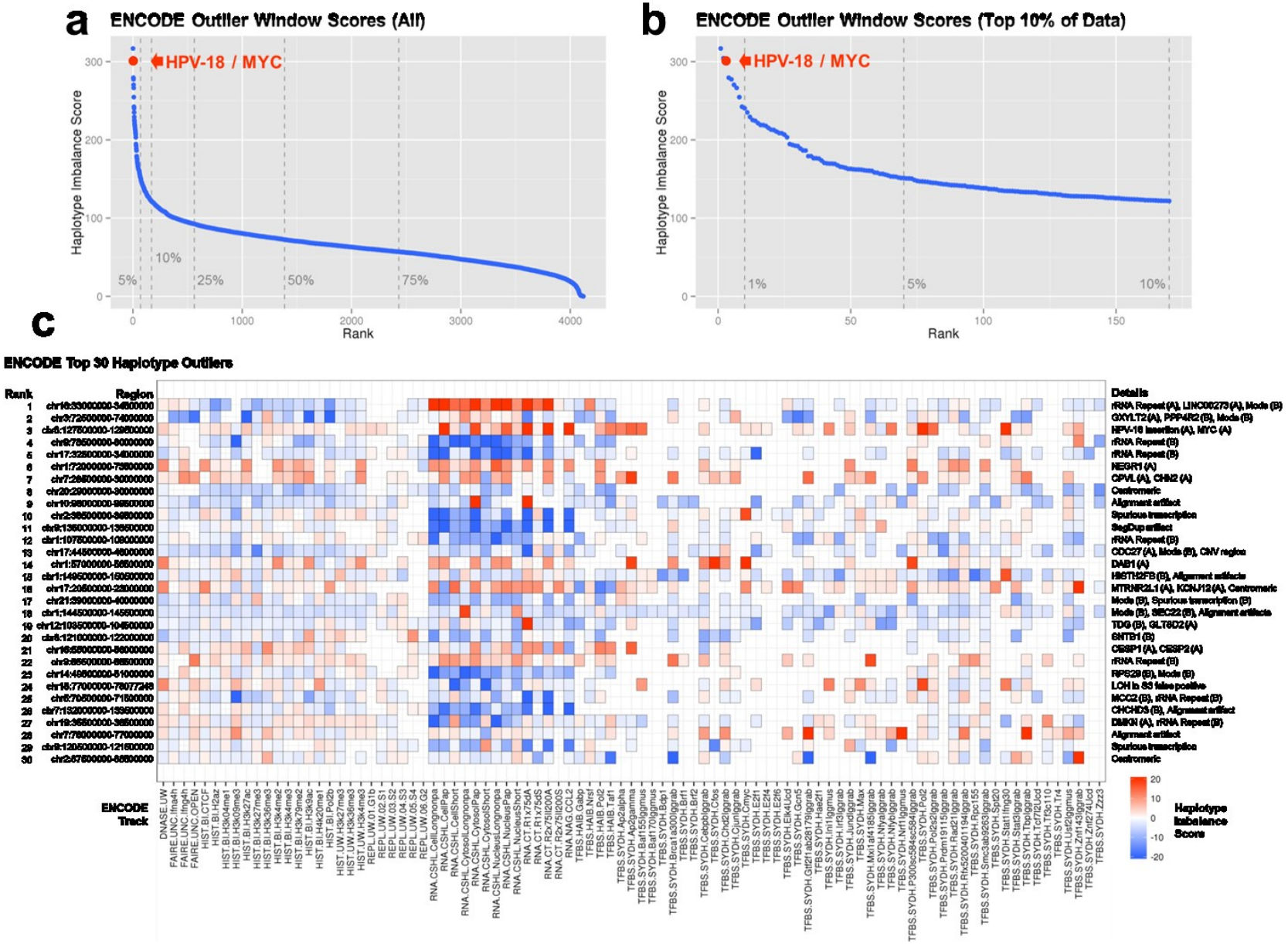

a
ENCODE Outlier Window Scores (All)
HPV-18 / MYC
Haplotype Imbalance Score
Rank
5%
10%
25%
50%
75%
b
ENCODE Outlier Window Scores (Top 10% of Data)
HPV-18 / MYC
Haplotype Imbalance Score
Rank
1%
5%
10%
c
ENCODE Top 30 Haplotype Outliers
Rank
Region
Details
ENCODE Track
Haplotype Imbalance Score

Figure D.4.45. 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.

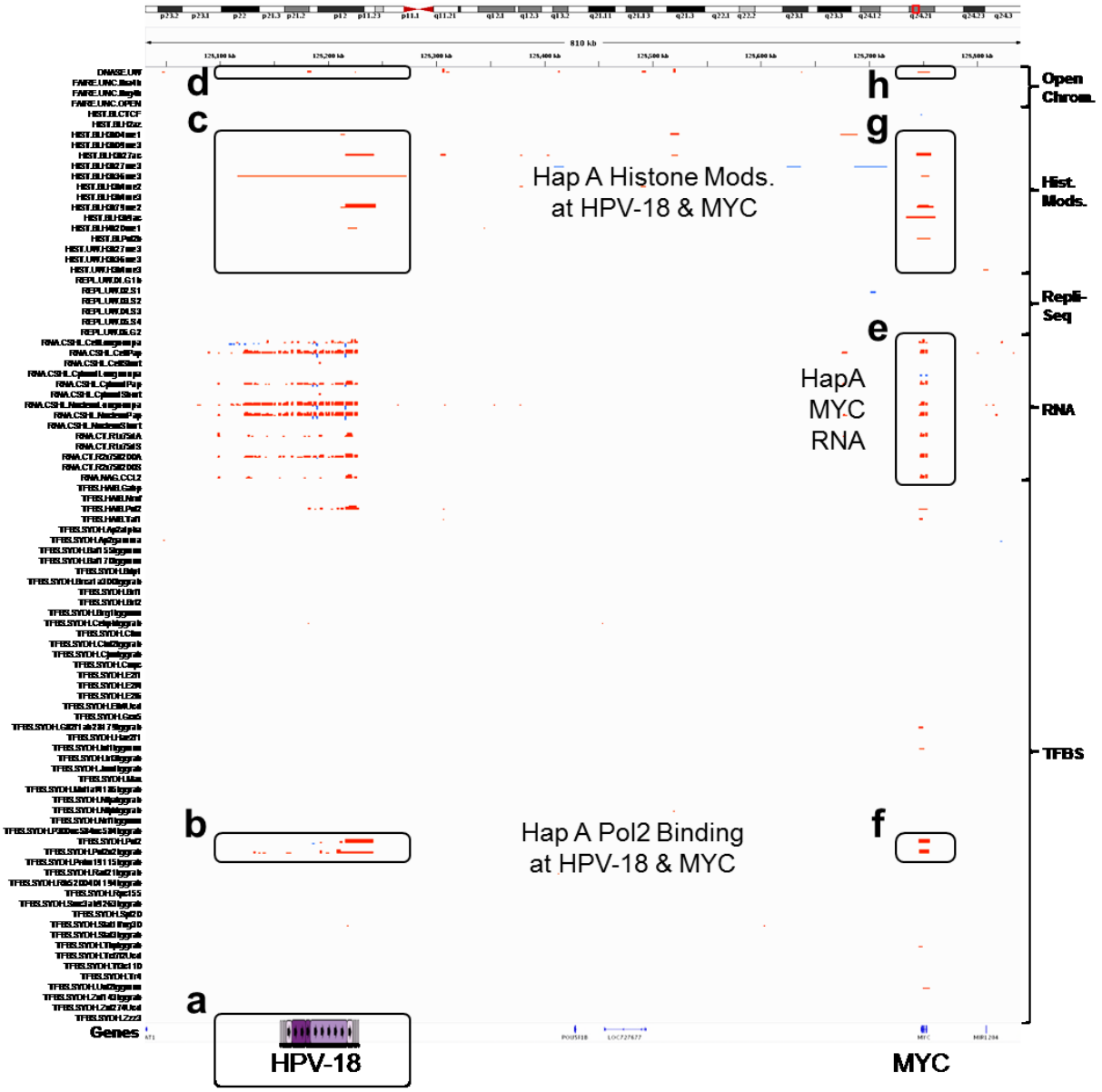


Figure D.4.46. ENCODE haplotype imbalances for HPV-18 and *MYC*.
Red peaks indicate haplotype A imbalance, blue peaks represent haplotype B for copy number normalized haplotype imbalance scores. Letters represent regions corresponding to items in the schematic presented in Figure 6.4a.

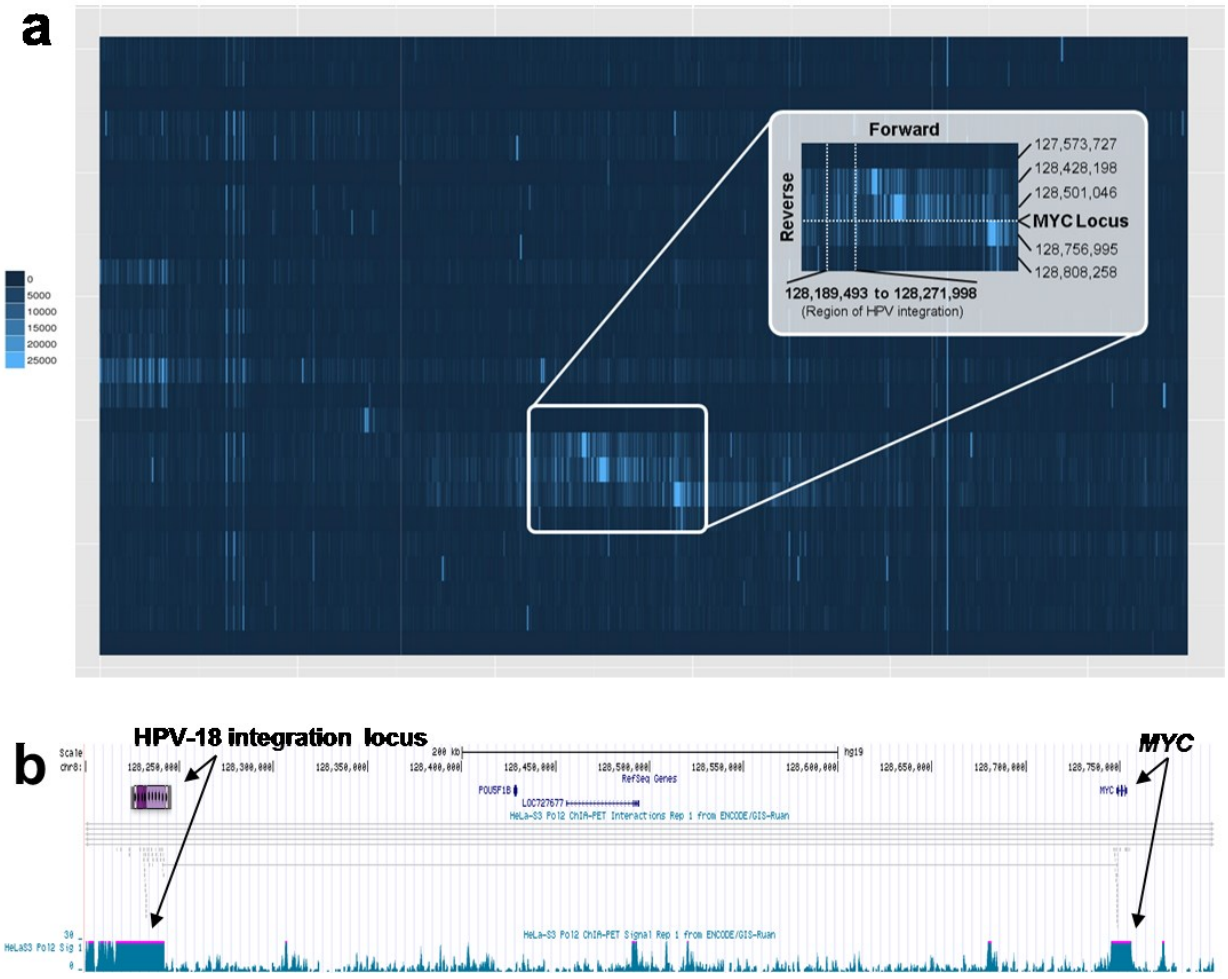


Figure D.4.47. Long-range interaction with *MYC* from 5C and ChIA-PET data.
a. ENCODE 5C chromatin interaction data (available only for the GM12878 cell line) demonstrates long-range interactions between *MYC* and distal upstream sites. The highlighted region includes the site of HPV-18 integration (into 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.

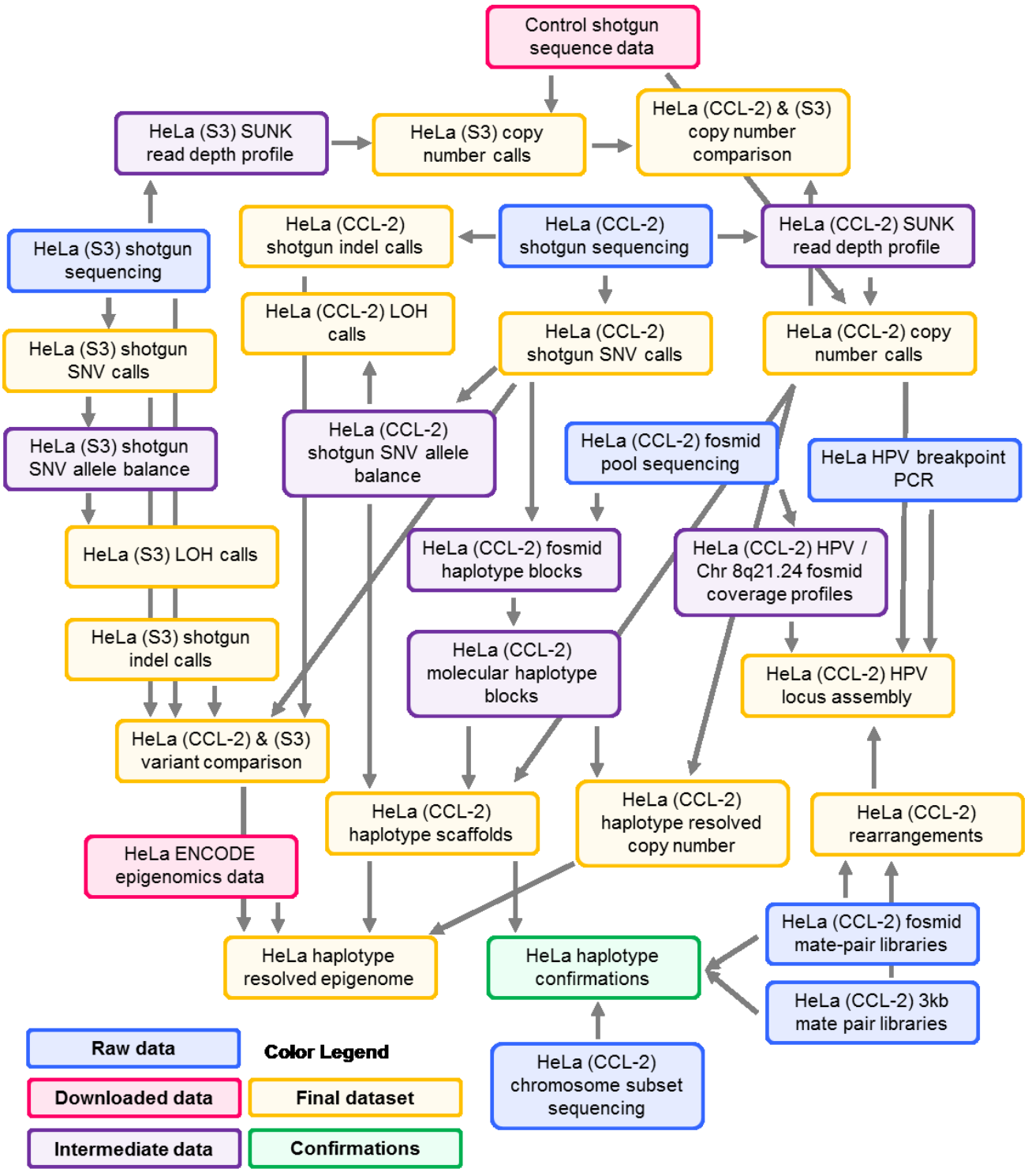


Figure S.4.48. Datasets and analyses for HeLa CCL-2 and HeLa S3.

References

1.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).

2.Watson, J. D. & Crick, F. H. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology* **18,** 123-31 (1953).

3.Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 5463-7 (1977).

4.Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921 (2001).

5.Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931-45 (2004).

6.Jou, W. M., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237,** 82-8 (1972).

7.Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94,** 441-8 (1975).

8.Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 560-4 (1977).

9.Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687-95 (1977).

10.Kasper, T. J., Melera, M., Gozel, P. & Brownlee, R. G. Separation and detection of DNA by capillary electrophoresis. *Journal of chromatography* **458,** 303-12 (1988).

11.Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674-9 (1986).

12.Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8,** 175-85 (1998).

13.Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8,** 186-94 (1998).

14.Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242,** 84-9 (1996).

15.Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik & Church, G. M. Fluorescent in situ sequencing on polymerase colonies. *Analytical biochemistry* **320,** 55-65 (2003).

16.Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* **309,** 1728-32 (2005).

17.Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376-80 (2005).

18.Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53-9 (2008).

19.Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348-52 (2011).

20.Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)* **323,** 133-8 (2009).

21.Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research* (2014). doi:10.1101/gr.168450.113

22.Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome research* **23,** 121-8 (2013).

23.Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature biotechnology* **30,** 1232-9 (2012).

24.Derrington, I. M. *et al.* Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 16060-5 (2010).

25.Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 18904-9 (2013).

26.Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621-8 (2008).

27.Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)* **330,** 641-6 (2010).

28.Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* **316,** 1497-502 (2007).

29.Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133,** 523-36 (2008).

30.Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology* **27,** 1173-5 (2009).

31.Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology* **30,** 265-70 (2012).

32.Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 139-44 (2010).

33.Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* **326,** 289-93 (2009).

34.DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science (New York, N.Y.)* **278,** 680-6 (1997).

35.Blat, Y. & Kleckner, N. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98,** 249-59 (1999).

36.Wu, J., Smith, L. T., Plass, C. & Huang, T. H. ChIP-chip comes of age for genome-wide functional analysis. *Cancer research* **66,** 6899-902 (2006).

37.Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics* **10,** 669-80 (2009).

38.Gitan, R. S., Shi, H., Chen, C., Yan, P. S. & Huang, T. H. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome research* **12,** 158-64 (2002).

39.Clark, S. J., Harrison, J., Paul, C. L. & Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic acids research* **22,** 2990-7 (1994).

40.Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* **33,** 5868-77 (2005).

41.Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452,** 215-9 (2008).

42.Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology* **26,** 779-85 (2008).

43.Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology* **28,** 1097-105 (2010).

44.Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57-74 (2012).

45.Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311-22 (2008).

46.Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75-82 (2012).

47.Jayaseelan, S., Doyle, F. & Tenenbaum, S. A. Profiling post-transcriptionally networked mRNA subsets using RIP-Chip and RIP-Seq. *Methods (San Diego, Calif.)* (2013). doi:10.1016/j.ymeth.2013.11.001

48.Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods* **10,** 361-5 (2013).

49.Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature methods* **9,** 1107-12 (2012).

50.Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature genetics* **14,** 450-6 (1996).

51.Payen, C. *et al.* The Dynamics of Diverse Segmental Amplifications in Populations of Saccharomyces cerevisiae Adapting to Strong Selection. *G3 (Bethesda, Md.)* (2013). doi:10.1534/g3.113.009365

52.Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994-1007 (2012).

53.Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154,** 888-903 (2013).

54.Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research* **23,** 555-67 (2013).

55.Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41,** 1061-7 (2009).

56.Singer, M. F. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28,** 433-4 (1982).

57.Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291,** 1304-51 (2001).

58.Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nature methods* **8,** 61-5 (2011).

59.Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11,** R119 (2010).

60.Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome research* **22,** 1139-43 (2012).

61.Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature biotechnology* **29,** 59-63 (2011).

62.Amini, S. *et al.* Haplotype-resolved Whole Genome Sequencing using a Contiguity Preserving Transposition Assay. *Under Review*

63.Adey, A. *et al.* In vitro, long-range sequence information for de novo genome assembly via transposase contiguity.

64.Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207-11 (2013).

65.Kahvejian, A., Quackenbush, J. & Thompson, J. F. What would you do if you could sequence everything? *Nature biotechnology* **26,** 1125-33 (2008).

66.Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature biotechnology* **26,** 1135-45 (2008).

67.McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research* **19,** 1527-41 (2009).

68.Lennon, N. J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome biology* **11,** R15 (2010).

69.Craig, D. W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nature methods* **5,** 887-93 (2008).

70.Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature methods* **5,** 1005-10 (2008).

71.Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* **6,** 291-5 (2009).

72.Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature methods* **7,** 130-2 (2010).

73.Reznikoff, W. S. Tn5 as a model for understanding DNA transposition. *Molecular microbiology* **47,** 1199-206 (2003).

74.Reznikoff, W. S. Transposon Tn5. *Annual review of genetics* **42,** 269-86 (2008).

75.Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60-5 (2008).

76.Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754-60 (2009).

77.Goryshin, I. Y., Miller, J. A., Kil, Y. V., Lanzov, V. A. & Reznikoff, W. S. Tn5/IS50 target recognition. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 10716-21 (1998).

78.Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078-9 (2009).

79.Zilberman, D. & Henikoff, S. Genome-wide analysis of DNA methylation patterns. *Development (Cambridge, England)* **134,** 3959-65 (2007).

80.Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature biotechnology* **27,** 353-60 (2009).

81.Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature protocols* **6,** 468-81 (2011).

82.Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27,** 361-8 (2009).

83.Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315-22 (2009).

84.Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics* **7,** e1002389 (2011).

85.Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS biology* **8,** e1000533 (2010).

86.Gertz, J. *et al.* Transposase mediated construction of RNA-seq libraries. *Genome research* **22,** 134-41 (2012).

87.Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research* **40,** D571-9 (2012).

88.Shemesh, M., Pasvolsky, R., Sela, N., Green, S. J. & Zakin, V. Draft Genome Sequence of

Alicyclobacillus acidoterrestris Strain ATCC 49025. *Genome announcements* **1,** (2013).

89.Rong, X. & Gardener, B. B. Draft Genome Sequence of Cryptococcus flavescens Strain OH182.9_3C, a Biocontrol Agent against Fusarium Head Blight of Wheat. *Genome announcements* **1,** (2013).

90.Wang, D., Wu, R., Xu, Y. & Li, M. Draft Genome Sequence of Rhizopus chinensis CCTCCM201021, Used for Brewing Traditional Chinese Alcoholic Beverages. *Genome announcements* **1,** e0019512 (2013).

91.Blanco-Ulate, B., Rolshausen, P. E. & Cantu, D. Draft Genome Sequence of the Grapevine Dieback Fungus Eutypa lata UCR-EL1. *Genome announcements* **1,** (2013).

92.Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 1513-8 (2011).

93.Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490,** 49-54 (2012).

94.Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology* (2013). doi:10.1038/nbt.2727

95.Adey, A., Morrison, H., Xun, X. & Kitzman, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome* (2010).

96.Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* (2013). doi:10.1038/nmeth.2688

97.Schwartz, J. J., Lee, C., Hiatt, J. B., Adey, A. & Shendure, J. Capturing native long-range contiguity by in situ library construction and optical sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 18749-54 (2012).

98.Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 5552-7 (2013).

99.Kidd, J. M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature methods* **7,** 365-71 (2010).

100.Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic acids research* **40,** 2041-53 (2012).

101.Gey, G., WD, C. & MT, K. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Research* 264-265 (1952).

102.Gartler, S. M. Apparent Hela cell contamination of human heteroploid cell lines. *Nature* **217,** 750-1 (1968).

103.Skloot, R. *The Immortal Life of Henrietta Lacks*. (Random House LLC, 2010).

104.Macville, M. *et al.* Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer research* **59,** 141-50 (1999).

105.Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7,** 548 (2011).

106.Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)* **338,** 222-6 (2012).

107.Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56-65 (2012).

108.Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45,** 81-94 (2008).

109.Goodwin, E. C. *et al.* Rapid induction of senescence in human cervical carcinoma cells. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 10978-83 (2000).

110.Rosty, C. *et al.* Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. *Molecular cancer* **4,** 15 (2005).

111.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,** 2252-63 (2002).

112.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,** 13174-86 (2012).

113.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,** 10333-40 (2008).

114.Wingo, S. N. *et al.* Somatic LKB1 mutations promote cervical cancer progression. *PloS one* **4,** e5137 (2009).

115.Wistuba, I. I. *et al.* Deletions of chromosome 3p are frequent and early events in the pathogenesis of uterine cervical carcinoma. *Cancer research* **57,** 3154-8 (1997).

116.Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature biotechnology* **29,** 51-7 (2011).

117.Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519-25 (2012).

118.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 cell to supply conditioning factors. *Proceedings of the National Academy of Sciences of the United States of America* **41,** 432-7 (1955).

119.Nelson-Rees, W. A., Daniels, D. W. & Flandermeyer, R. R. Cross-contamination of cells in culture. *Science (New York, N.Y.)* **212,** 446-52 (1981).

120.Wentzensen, N., Vinokurova, S. & Doeberitz, von, M. K. 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,** 3878-84 (2004).

121.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,** 4305-10 (1989).

122.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,** 422-7 (2000).

123.Peter, M. *et al.* MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. *Oncogene* **25,** 5985-93 (2006).

124.Ahmadiyeh, N. *et al.* 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 9742-6 (2010).

125.McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science (New York, N.Y.)* **342,** 632-7 (2013).

126.Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* **11,** 41-6 (2014).

127.Yamaguchi, S. *et al.* Tet1 controls meiosis by regulating meiotic gene expression. *Nature* **492,** 443-7 (2012).

128.Wang, Q. *et al.* Tagmentation-based whole-genome bisulfite sequencing. *Nature protocols* **8,** 2022-32 (2013).

129.Shendure, J., Schwartz, J. & Adey, A. Massively parallel contiguity mapping. *WO Patent* (2012).

130.Kitzman, J. O. *et al.* Noninvasive whole-genome sequencing of a human fetus. *Science translational medicine* **4,** 137ra76 (2012).

131.Snyder, M. W. *et al.* Noninvasive fetal genome sequencing: a primer. *Prenatal diagnosis* **33,** 547-54 (2013).

132.Fan, H. C. *et al.* Non-invasive prenatal measurement of the fetal genome. *Nature* **487,** 320-4 (2012).

133.Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43-9 (2014).

134.Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90-4 (2011).

135.Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science (New York, N.Y.)* **338,** 1622-6 (2012).

136.Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (New York, N.Y.)* **343,** 193-6 (2014).

137.Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* (2013). doi:10.1038/nmeth.2772

138.Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome research* (2013). doi:10.1101/gr.161034.113

139.Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153,** 1134-48 (2013).

140.Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148,** 816-31 (2012).

141.Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (New York, N.Y.)* **341,** 1237905 (2013).

142.Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research* (2013). doi:10.1101/gr.161679.113

143.Manoil, C. & Beckwith, J. TnphoA: a transposon probe for protein export signals. *Proceedings of the National Academy of Sciences of the United States of America* **82,** 8129-33 (1985).

144.Blin, N. & Stafford, D. W. A general method for isolation of high molecular weight DNA from eukaryotes. *Nucleic acids research* **3,** 2303-8 (1976).

145.Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *Journal of molecular biology* **188,** 415-31 (1986).

146.Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18,** 6097-100 (1990).

147.Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)* **329,** 75-8 (2010).

148.Adey, A., Burton, J., Kitzman, J., Hiatt, J. & Lewis, A. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* (2013).

149.Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26,** 589-95 (2010).

150.McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297-303 (2010).

151.Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and

population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)* **27,** 2987-93 (2011).

152.Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome research* **22,** 1154-62 (2012).

153.Da Huang, 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).

154.Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods* **7,** 576-7 (2010).

155.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,** 469-81 (2011).

156.Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25,** 1105-11 (2009).

157.Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)* **27,** 2325-9 (2011).