Molecular tagging to overcome limitations of massively parallel sequencing

Joseph B Hiatt

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

Doctor of Philosophy

University of Washington

2012

Reading Committee:
Jay Shendure, Chair
Phil Green
Robert Waterston

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Molecular tagging to overcome limitations of massively parallel sequencing

Joseph B Hiatt

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

Massively parallel technologies have dramatically reduced the cost and increased the throughput of DNA sequencing, transforming the study of patterns of genetic variation in human and model organisms, the genetic basis of disease, and the organization, regulation, and function of genomes. However, the cost and throughput gains of the most widely used massively parallel sequencing platforms are offset by substantial drawbacks with respect to read length and base-calling accuracy, limiting their utility for many exciting and potentially important research and clinical applications. These include *de novo* assembly of genomes and metagenomes, diversity profiling of metagenomic communities and viral populations, synthetic functional assays to interrogate long genetic elements, and sensitive and highly multiplexed cancer-related gene sequencing. This dissertation describes a paradigm, "molecular tagging," which I developed to overcome the read length and error rate limitations of massively parallel sequencing, in the context of its use for three distinct applications. I first describe the development of a molecular tagging-based method called "Subassembly" and its application to *de novo* genome and metagenome assembly. Complexity-bottlenecked libraries of DNA fragments at least ~350 nucleotides in length were amplified, re-fragmented, and subjected to massively parallel sequencing such that we could identify groups of reads derived from the same longer progenitor fragment. Groups of reads were then locally assembled to generate highly accurate haploid consensus sequences that effectively corresponded to single input molecules, and these long consensus sequences were used to improve *de novo* genome assembly for single bacterial and metagenomic samples. Next, I describe the development and application of a massively parallel assay to dissect transcriptional enhancer elements with single nucleotide resolution. We applied this method to determine the functional consequences of all possible single nucleotide

changes in three mammalian enhancers; this assay involved an optimized version of the Subassembly experimental protocol and analytical strategy. Finally, I describe the characterization and application of a multiplex assay for ultra-sensitive targeted sequencing of cancer-related genes in clinical tumor samples. I integrated the molecular tagging paradigm with the molecular inversion probe multiplex capture strategy to develop a method that is simultaneously simple, rapid, cost-effective, and ultra-sensitive for sub-clonal variation in genetically heterogeneous tissue samples. Molecular tagging is therefore a broadly applicable strategy to overcome key limitations of massively parallel sequencing that is expanding the utility of this already transformative group of technologies to a number of important research and clinical applications.

ACKNOWLEDGMENTS

I am deeply indebted to many for unwavering support, both academic and personal, over the past several years. Here I attempt to briefly (and inadequately) acknowledge their significance to my graduate career.

Jay Shendure, my mentor, has been nothing short of ideal. His creativity, enthusiasm, optimism, and curiosity are infectious and were a constant source of motivation and inspiration to me. Jay was always available, interested, and involved without being heavy-handed. I am also very grateful to him for building the Shendure lab into what it is today: a fun, exciting, creative, intellectually stimulating and enriching, and incredibly productive leader in genomics technology development.

I am very grateful to the entire Shendure lab for being a fantastic group of colleagues, and I owe special gratitude to several members of the lab. Rupali Patwardhan has been an invaluable collaborator and mentor on several projects. I am grateful to her for her critical contributions to the development of the Subassembly method and for allowing me to participate in her paradigm-establishing enhancer bashing project. She has been an endlessly patient source of advice on all matters related to analytical design and implementation and is largely responsible for any proficiency in bioinformatics that I have developed over my graduate career. Emily Turner was also an essential contributor to Subassembly, having conceived the original method, and was a generous scientific and career mentor throughout. Jacob Kitzman was a model colleague and was especially patient and generous with his time as I attempted to learn more sophisticated programming strategies. Choli Lee defies description; his patience, enthusiasm, and willingness to accommodate the absurd requests of frantic scientists underlie the success of the Shendure lab, Genome Sciences, and the University of Washington as a whole where Illumina sequencing is concerned. Ruolan Qiu has been a remarkably diligent and patient teacher of traditional molecular biology at its highest levels. She has also been a uniquely generous and genuine colleague and friend. The Shendure lab is a warmer, kinder place with Ruolan, not to mention infinitely more rigorous. Jerrod Schwartz, Brian O'Roak, Akash Kumar, and Steve Salipante have all been valuable collaborators on various projects. Finally, Sarah Ng, Andrew Adey, Martin Kircher, and the rest of the Shendure lab have been invaluable sources of advice, friendship, and camaraderie.

I have also been extremely fortunate to engage in a number of productive and enriching collaborations with scientists outside the Shendure lab. For contributing substantially to my smMIP project, I am very grateful to Colin Pritchard in Laboratory Medicine. Furthermore, I am very grateful to: Lea Starita in the Fields lab; Xinxian Deng in the Disteche lab; Russell Berg in the Ramakrishnan lab; Mike Schmitt, Jesse Salk, and Scott Kennedy in the Loeb lab; and Gabriel Loeb in the Rudensky lab (at Memorial Sloan-Kettering Cancer Center) for including me in their exciting efforts. Working with and learning from these talented researchers was an honor.

I would like to thank Larry Loeb and Peter Rabinovitch for their support and guidance on a fellowship application. Lisa Border in Genome Sciences has also been a very patient and helpful guide through the intricacies of the funding process.

Throughout my career I received innumerable forms of support from the MSTP administration, including Marshall Horwitz, Larry Loeb, Mary Claire-King, Maureen Holstad, Marcie Buckner, and Julia Lawrence. I was also extremely fortunate to receive support from the Barbo family through the ARCS program. Their generous financial support has been extremely helpful, but I am also especially grateful to Chuck, Linda, Julie, Will, and the rest of the Barbo clan for their friendship and personal support, and their continued generosity towards UW graduate students.

I am also particularly grateful to my MSTP E-07 classmates, who have been a constant and unwavering source of friendship, support, and academic excitement for the past five years, and to my Genome Sciences and medical school classmates as well.

Several great scientists and generous mentors gave me early opportunities far more substantial than I merited, helped cultivate my interest in science, and taught me fundamental principles of scientific research in general and molecular biology in particular. To Eric Green, my primary mentor in the Dolmetsch lab, I cannot possibly express sufficient gratitude. Eric is an incredibly generous and patient teacher, talented and brilliant scientist, and great friend. Ricardo Dolmetsch and Steve Quake gave me invaluable inopportunities as a novice scientist. Members of their respective labs, including Natalia Gomez-Ospina, Jocelyn Krey, Rafael Gomez-Sjoberg, and Anne Leyrat, were always patient and

generous mentors. Angela Wu was a willing and able collaborator and I was very fortunate to work with her during and after my time in the Quake lab.

Finally, I would like to thank my thesis committee (Tony Blau, Stan Fields, Phil Green, and Bob Waterston) for gently and firmly helping to shape my graduate career, and for sage wisdom and helpful comments throughout.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Opening comments

DNA is the critical information storage medium underlying life; as such, many areas of biomedical research and clinical medicine involve the manipulation of DNA in some capacity. Unlike some classes of RNA and nearly all proteins, DNA has a relatively stereotyped three-dimensional structure[1] such that the majority of its biological function is encoded in its primary sequence. Methods for the determination of the sequence of a given DNA fragment have therefore been the focus of intensive efforts over the course of decades; the effectiveness, cost, and scalability of such methods have widespread implications for basic and translational research as well as clinical practice.

Methods for determining the sequence of a clonal population of DNA fragments that could be considered practical were first described by Maxam and Gilbert[2] and Sanger[3] in 1977. The Sanger method eventually gained wider use and was aggressively optimized with respect to a number of critical parameters (*e.g.* cost, read length, base-calling accuracy, parallelization, and automation) in a broad effort across the scientific community. Especially important advances included the shift from slab gel to capillary electrophoresis[4] and from radioactive to fluorescent detection[5] of ddNTP-terminated fragments and the development of sophisticated algorithms to call nucleotides and estimate associated error probabilities from the raw instrument readings[6,7]. These methodological improvements resulted in routine acquisition of reads more than 500 nucleotides (nt) long with per-base substitution error rates below $1x10^{-3}$ (*i.e.* ≥Q30). Methods for the relatively cost-effective, routine and automated collection of long and highly accurate sequencing reads facilitated landmark achievements in the genomics community; most notably, these include the complete sequencing of several model organism[8-10] genomes as well as the human genome[11]. While improvements in sequencing cost and throughput using the Sanger method were substantial, the cost of sequencing individual eukaryotic genomes by the early 2000s remained extremely high (in the tens of millions of dollars). Beginning in 2005, a new generation of technologies began to emerge[12-15] that promised to dramatically reduce sequencing costs and increase throughput. These platforms are often referred to collectively as “massively parallel” or “next-generation” sequencing; in this dissertation, I will use the term massively parallel sequencing (abbreviated MPS). While the details of

implementation are quite different across platforms, they share many important features, including the parallel manipulation and interrogation of a very large number of molecular "features" with a single reagent volume and detector. This stands in stark contrast to even the highest throughput Sanger implementations, where sequence reads are derived from individual reaction volumes, and leads to reductions in reagent use and instrument occupancy of several orders of magnitude[16].

The cost and throughput advances of MPS have led to an ongoing transformation of many areas of genomics, including the study of genetic variation[17], gene expression[18] and its regulation[19,20], the genetic basis of rare[21] and common[22] disease, and numerous other areas. However, these advances are offset by substantial drawbacks with respect to read length and base-calling accuracy[23], and for many exciting potential uses of MPS, long and/or accurate reads are indispensable. These areas include *de novo* genome assembly[24,25], shotgun and amplicon metagenomics[26], structural variation in larger genomes[27], pathogen diversity, detection of genetic variation amongst a heterogeneous population of genomic equivalents (sub-clonal variation), and accurate ascertainment and phasing of variation in synthetic nucleic acid populations. Methods are therefore needed that combine the low cost and extreme throughput of MPS with read lengths and accuracies approaching or even exceeding those of Sanger sequencing.

1.2 Organization

This dissertation describes the development and application of a flexible and broadly useful paradigm, termed "molecular tagging," that is capable of extending the effective read length, decreasing the error rate, and increasing the quantitative precision of MPS platforms. Molecular tagging substantially expands the potential application space of MPS. In Chapter 2, I review existing (*i.e.* those available prior to the widespread use of MPS) methods for the detection of sub-clonal genetic variation. In Chapter 3, I highlight selected potential applications of MPS that would benefit from longer reads and/or higher accuracy base calling. In Chapter 4, I describe the initial development of molecular tagging and its application to *de novo* bacterial and metagenomic genome assembly. In Chapter 5, I describe the development of a method, termed Massively Parallel Functional Dissection (MPFD), to interrogate the functional impact of all possible substitution mutations in mammalian transcriptional regulatory elements;

further developments to molecular tagging were indispensable for MPFD. In Chapter 6, I describe the development of a targeted capture method involving molecular tagging that is sensitive to sub-clonal genetic variation and its application to rapidly, cost-effectively, and deeply sequencing the complete coding sequence of actionable disease genes in clinical cancer samples. In Chapter 7, I survey advances in molecular tagging due to my work and the work of other laboratories, describe alternate approaches to achieve long reads and/or highly accurate base-calls available as of August 2012, and speculate on future opportunities and challenges for this paradigm.

Chapter 2 Methods to detect sub-clonal genetic variation

In this chapter, I describe established methods to sensitively and accurately detect sub-clonal genetic variation. With one obvious exception, these methods are not sequencing-based, limiting their generality to detect arbitrary variation and their ability to phase variation located on the same haplotype. The importance of a combined method, *i.e.* a sequencing method that is capable of both long reads and detecting sub-clonal genetic variation, is described in more detail in the following chapter. However, a discussion of existing approaches also serves to motivate the need for the development of new methods. This discussion focuses primarily on methods that existed prior to the widespread adoption of massively parallel sequencing (MPS) platforms; while describing my work in later chapters (Chapters 4-6) and in a concluding chapter (Chapter 7), alternative MPS-based methods will also be discussed.

Methods for the detection of sub-clonal genetic variation have been actively developed and applied since the earliest days of genetics and molecular biology. Here I will describe a relevant subset of these methods that remain in use today. These are phenotypic screening[28,29], cloning and sequencing (including COLD-PCR[30]), single molecule PCR (including digital PCR[31] and BEAMing[32]), allele-specific PCR[33,34], the Random Mutation Capture assay[35], and mass spectrometry[36,37]. My discussion focuses on three important performance criteria: sensitivity, throughput and scalability, and generality to detect arbitrary mutations.

2.1 Phenotypic screening

Phenotypic screening may be considered the first genotypic discrimination method; for example, a full decade before the discovery of the structure of DNA, Luria and Delbruck famously used the emergence of resistance to phage infection in bacterial cultures to gain insight into the mechanism of genetic mutation[38]. By the 1970s the central importance of DNA was fully appreciated and more explicit methods had been developed to detect certain types of mutations based on the emergence of a particular phenotype in a living host. These methods fell into two general classes: reversion assays and forward mutation assays. Reversion assays select for a spontaneous or mutagen-induced mutation that restores the wild-type sequence of a mutant gene and thus its function, for which a selection pressure is being applied. A

particularly well-known application of reversion assays is the so-called Ames test[28] to establish the mutagenic and thus carcinogenic potential of environmental chemicals, which works by subjecting various Histidine auxotrophs of *Salmonella typhimurium* to a given agent and screening for the emergence of prototrophs. To test for various mechanisms of mutagenesis, multiple auxotrophic strains are used with different mutations (*i.e.* point mutations, insertions or deletions in the histidine biosynthetic pathway). Another type of phenotypic screening is the forward mutation assay, which selects for the loss of a gene that is conferring a conditionally toxic phenotype, *e.g.* an enzyme that converts an added metabolite into a toxin, thereby killing cells that harbor a functional copy of the gene[29]. Advantages of these technologies include: (a) sensitivity and quantitative precision often limited only by the number of organisms that can be cultured and the mutation rate of the polymerase used to replicate that sequence in the experimental host; (b) selection for mutations independent of sequence characterization such that these methods can be performed without cost- and labor-intensive sequencing. Indeed, these methods were developed prior to the description of Sanger or Maxam-Gilbert sequencing. However, they have several critical drawbacks related to generality, including: (a) need for the sequence of interest to encode an element whose function is easily assessed; (b) need for manipulation of the sequence into an appropriate host for functional assessment (such as a plasmid in a bacterial cell); and (c) restriction to the study of mutations that alter the response of the organism harboring the sequence to the selection pressure being applied, which may render certain positions in a given element as well as certain mutations at a given position invisible to the assay. These methods have therefore been used largely to study general properties of mutation, including the mutagenic properties of environmental chemicals[28] and the mutation rates of various polymerases[39].

2.2 Cloning and Sanger sequencing

Another method of detecting sub-clonal variation is direct cloning and Sanger sequencing of sequences of interest. Advantages of this approach include direct sequence characterization (such that each assay is not specific to a single point mutation or indel) and the collection of long sequence reads for phasing variation. However, cloning and Sanger sequence is extremely slow, labor-intensive, and costly[16]. According to Poisson statistics, observing a given variant with 95% probability requires ~3-fold

oversampling relative to its frequency; for example, ~300x coverage is required to observe a variant present at one percent in 95% of trials. Therefore, the detection of sub-clonal variation in the absence of any enrichment for variant-harboring sequences using cloning and sequencing is not practical. Furthermore, an amplification step (typically PCR) is almost always used in order to enrich for a given region of interest, such that detection of sub-clonal variation by cloning and sequencing is limited by the error rate of this step (for non-proofreading PCR polymerases, on the order of one error in ten thousand[40]). Alternatively, the PCR product of a region of interest can be directly sequenced (*i.e.* without effective single molecule separation by way of cloning). In this approach, substitution variation is detected as the superposition of two peaks at a single position in the electropherogram. However, the sensitivity of this approach is very poor. To address this limitation, an enrichment method for variant sequences, termed COLD-PCR, was recently developed[30]. By reducing the denaturing temperature of the PCR, COLD-PCR selectively amplifies heteroduplex sequences, and as such, is capable of enriching sub-clonal substitution mutations by as much as ~13-fold and deletion mutations even more dramatically, *i.e.* ~300-fold. Initial descriptions of COLD-PCR focused on the enrichment of sub-clonal variation such that it became detectable by Sanger sequencing; more recently it has also been applied as a step in a MPS workflow[41]. While COLD-PCR is an effective enrichment method for sub-clonal variation, it is not quantitative, limiting its utility. Furthermore, assay conditions must be tuned to a particular target amplicon and it remains unclear whether the approach will multiplex well.

2.3 Single molecule PCR

Limiting dilution single molecule or “digital” PCR is another powerful approach for the detection and quantitation of sub-clonal variation[31]. By diluting PCR templates to an expected frequency of approximately one per two PCR reactions, most reactions that productively amplify will have been seeded by a single template molecule. This in turn enables one-by-one interrogation of clonal populations of molecules. In an early description, a fluorescent probe was used in the course of PCR that discriminated wild-type from mutant sequences[31], which is a much more cost-effective read-out for mutation status than Sanger sequencing. The number of PCR reactions harboring a variant amplicon population is compared to the number harboring a wild-type population, directly yielding the variant frequency. Compared to

phenotypic screening, digital PCR is far more general as it does not rely on function encoded by the sequence of interest. Compared to cloning and Sanger sequencing, it can be much more cost-effective; however, if fluorescent probes or other methods are used as a read-out in lieu of sequencing, the method is necessarily restricted to one or a small number of expected variants. Digital PCR is similarly limited by the error rate of the PCR polymerase. Finally, the throughput of digital PCR as described by Vogelstein and Kinzler was very limited, as reactions were carried out in a 96-well plate and two to four plates were used in each experiment. To overcome this throughput limitation, Vogelstein's group subsequently extended the method to leverage the parallelization afforded by water-in-oil emulsions and flow cytometry with a technology called BEAMing (for Beads, Emulsion, Amplification, Magnetics)[32]. While BEAMing enables substantial improvements in throughput and sensitivity (to at least 1% variant frequency and likely much lower), the assay remains sensitive to PCR polymerase mutation rates, specific to one or a small number of expected variant alleles, and is complex and difficult to carry out.

2.4 Allele-specific PCR

Allele-specific PCR is based on the selective amplification of one allele relative to another by designing one or both primers to select for an allele of interest. Since its initial description in 1989[33,34] and thanks to a wide variety of subsequent efforts to optimize the method, allele-specific PCR has achieved widespread adoption in both research and clinical settings as a reliable method for sub-clonal variant detection. Indeed, recent studies have established sensitivity for sub-clonal variants present at frequencies below 0.1%[42], although this level of variant discrimination is not routinely achieved in the clinical diagnostic setting. Advantages of allele-specific PCR include a simple workflow, low input requirement, and because the read-out is not sequencing-based, low cost. However, the assay requires prior knowledge of the variant, development and optimization for each variant of interest, does not scale well to interrogating multiple variants in a single assay, and only exhibits modest sensitivity.

2.5 Random Mutation Capture

Many of the approaches discussed above are susceptible to the mutation rate of PCR polymerases because the genotypic discrimination event takes place downstream of amplification. One recently developed approach[35] circumvents this limitation by performing genotypic discrimination prior to

amplification. This method, termed Random Mutation Capture (RMC), is effectively an extension of longstanding genotyping methods based on the creation or destruction of a restriction site by a genomic variant. In the case of the RMC assay, a restriction site is destroyed by a variant, resulting in an amplifiable template whose frequency can be quantified by serial dilution digital PCR. This method is potentially highly sensitive (to at least ~$2x10^{-8}$ as estimated by the authors), but is restricted to interrogation of variants that disrupt restriction sites, and can be technically challenging.

2.6 Mass spectrometry-based genotyping

Finally, mass spectrometry has been used for multiplexed genotyping of specific sites[37] and detection of sub-clonal genetic variants[36]. In this approach, MALDI-TOF mass spectrometry is used to discriminate single nucleotides added to the end of a primer that is hybridized adjacent to the site to be genotyped. Primers are partially complementary to the target of interest and contain non-complementary 5' tag sequences to enable locus discrimination; genotyping is performed by resolving the mass of the single dideoxynucleotide incorporated onto the 3' end of the primer. In contrast to all of the methods described this far, this method scales well with respect to both the number of sites genotyped and the number of samples processed. For example, Thomas *et al* report genotyping 238 sites in 1,000 tissue specimens[36]. However, the assay is restricted to pre-specified variants, limited by the error rate of the whole genome amplification process used to prepare samples, and has not been shown to be sensitive to variation present at below ~5%.

2.7 Conclusion

In summary, several methods for the detection of sub-clonal variation were in use at the time the work described herein began. Some, like the Random Mutation Capture assay, can be exquisitely sensitive, but are not generally applicable to arbitrary variation and are not easily multiplexed to a large number of genomic targets. Most are not sequencing-based at all due to the high cost and low throughput of Sanger sequencing; those that do use Sanger sequencing are typically cost- and throughput-limited, resulting in poor sensitivity for sub-clonal variation at low frequencies.

Chapter 3 Applications for high-throughput, high-accuracy sequencing with the potential for long reads

Beginning in 2005 with the initial descriptions of commercially viable massively parallel sequencing approaches[12,13] and continuing with subsequent descriptions of other commercial platforms over the past several years[14,15,43,44], MPS platforms have transformed many areas of biomedical research. Key applications include but are by no means limited to: (a) resequencing genomes[45] and their protein-coding complement (*i.e*. "exomes")[17] to gain insight into patterns of genetic variation with respect to both single nucleotide variation as well as larger scale copy number and structural variation in the human lineage[46-50] and other organisms; (b) discovering the genetic etiology of rare Mendelian diseases[21,51]; (c) genetic studies of ancient hominins[52,53]; (d) genome- and transcriptome-wide profiling of protein-DNA[54,55] and protein-RNA[56-58] interactions; (e) high-resolution functional dissection of regulatory elements[59] and proteins[60]; (f) elucidation of the genetic underpinnings of various cancers[61-64]; (g) abundance measurements of mRNAs[18] and other RNA species, including microRNAs; and (h), transcriptome-wide transcription[65,66] and translation[67] rate measurements. In all of these applications, MPS has rapidly replaced existing technologies due to its extremely low cost, high throughput, and digital read-out for variation[68].

However, for a wide variety of important applications that could potentially be transformed by MPS technologies, the cost and throughput gains of the most cost-effective MPS platforms are rendered moot by the associated error rate and read length limitations. In this chapter, I will highlight selected applications that would benefit substantially from a cost-effective, high-throughput sequencing approach that lacks these limitations. As was just highlighted, many applications of MPS do not suffer substantially from these drawbacks, and many applications that would benefit from a technology capable of either long reads or sensitive sub-clonal variant detection would not necessarily require both. In the course of this discussion, I will describe the extent to which one or both of these obstacles must be circumvented for each application. The applications to be discussed are: (a) *de novo* genome and metagenome assembly; (b) long amplicon sequencing for metagenomic taxonomic classification, viral gene and genome diversity, and functional genomics; and (c) deep sequencing of cancer-related genes in clinical samples.

3.1 De novo genome and metagenome assembly

De novo genome assembly is one of the longstanding challenges in genomics because of the fundamental role played by genome sequences in diverse areas of basic biological and biomedical investigation[69]. Genome sequences yield new genes, reveal patterns of evolutionary history and sequence constraint, and enable much more sophisticated manipulation of experimental organisms. Additionally, the importance of germline[70] and somatic[71] structural variation in human health and disease and the difficulty of detecting these events when aligning sequencing data to a reference genome suggests a use for *de novo* assembly even in "resequencing" projects. The *de novo* genome assembly field has made several milestone achievements, including: (a) the first shotgun genome sequence (the bacteriophage lambda) in 1982[72]; (b) proposal of a method to map the human genome using restriction fragment length polymorphisms (RFLPs) in 1980[73]; (c) construction of the first genome-wide human genetic map in 1987[74]; (d) the first non-viral genome (the human bacterial pathogen *Haemophilus influenzae*) in 1995[75]; (e) the first eukaryotic genome (*Saccharomyces cerevisiae*) in 1997 followed by other eukaryotic model organisms including *Caenorhabditis elegans*[10], *Drosophila melanogaster*[76], and *Arabidopsis thaliana*[77] by the year 2000; and (f) the draft[78] and nearly complete[11] sequence of the human genome by 2004.

3.1.1 Shotgun sequencing using Sanger reads

With the exception of the phiX genome, which was sequenced using the "plus-minus" method, the landmark genomes described above were all sequenced using some form of "shotgun" Sanger sequencing, where random fragments of genomic DNA are cloned, sequenced, and assembled using sophisticated computational tools. Important differences exist between the approaches used for the various organisms described above, illustrating the complexity of *de novo* genome assembly. One such difference is the extent to which "hierarchical shotgun" as opposed to naïve shotgun sequencing is used, depending on the size and repeat content of the genome under study[78]. However, all major genome assemblies have been based on highly accurate, long Sanger reads. Notably, the accuracy and read lengths of Sanger sequencing was substantially improved during the scale-up for the Human Genome Project thanks to such innovations as the *phred* base-calling algorithm[6,7] and novel dye chemistry. Still,

the very high per-base cost and low throughput of the Sanger technology rendered the sequencing of new genomes a slow and expensive process.

3.1.2 The advent of massively parallel sequencing

Short read MPS technologies (*i.e.* Illumina, AB SOLiD) rapidly transformed the *de novo* assembly problem. By removing cost and throughput barriers for all but the largest genomes, sequencing a genome to high fold-coverage became routine. However, short read lengths and high per-base error rates prevented application of the same computational approaches[79] (namely, the overlap-layout-consensus strategy employed by the most popular assemblers for Sanger data). As an aside, it should be noted that one MPS technology (the Roche/454 platform[12]) offers substantially longer read lengths (~200-400 nt at the time) and the most popular assembler designed for this platform also uses the overlap-layout-consensus approach. However, the Roche/454 platform is at least ten-fold more expensive per-base than the short read platforms and has a high error rate and highly biased error profile[16], motivating the development of computational approaches to use short read data. Therefore, given the impracticality of the overlap-layout-consensus strategy for short read data and the high cost of the Roche/454 platform, an approach using de Bruijn graphs (initially proposed by Waterman and colleagues[80,81]) and designed for short read data was adopted in several implementations[82-86]. Generally speaking, this approach works by decomposing short reads into even substrings of length *k* (*i.e. k*-mers) and constructing a directed graph where nodes are individual *k*-mers observed in reads and edges connect adjacent *k*-mers within reads. Compared to the overlap-layout-consensus method, the de Bruijn graph approach scales far more favorably with increasing amounts of sequence data because, in the absence of sequencing errors, the size of the graph depends only on *k*-mer length and the multiplicity of *k*-mers in the true sequence. Reconstruction of the genome sequence using the de Bruijn graph is then reduced to the polynomial time Eulerian path problem[81].

3.1.3 Parameters influencing *de novo* assembly contiguity and accuracy

For a given genome to be assembled and in the absence of sequencing errors, the contiguity of a *de novo* assembly using a de Bruijn graph algorithm (with sufficient coverage) can be explicitly calculated for a given *k*. As *k* is increased, *k*-mer multiplicity is reduced, resulting in more contiguous assemblies.

Therefore, holding all other factors constant, increased read length leading to increased *k* would yield more contiguous assemblies. However, the situation is substantially complicated by several additional factors. Foremost amongst these is sequencing error, which rapidly complicates the graph by introducing false nodes and edges. To address this challenge, many de Bruijn graph assemblers perform pre-assembly "error correction" of reads as well as editing of the graph during assembly to remove structures characteristic of sequencing error; stand-alone error correction packages have also been developed[87]. Still, these approaches add complexity to the assembly process[87], and are problematic when applied to metagenomic samples. Reducing sequencing error is therefore also expected to improve assembly contiguity and accuracy (again, holding all other factors constant). However, the multitude of other factors that influence the contiguity and accuracy of a *de novo* assembly should not be ignored. These include the size and repeat content of the genome being sequenced, coverage non-uniformity due to library construction and sequencing biases, sequencing error types and patterns, and the use of so-called "paired-end" or "mate-pair" reads[88] (*i.e.* where two reads are collected from opposite ends of a longer fragment). Paired-end or mate-pair reads can yield highly contiguous assemblies even with short reads (and therefore small *k*), which reduces the need to attain longer read lengths genome-wide. A new method that is capable of improving read length or accuracy but also negatively affects another parameter, such as coverage, sequence context bias, or paired-end capability, may in fact have a net negative impact on *de novo* assembly quality. It is therefore difficult to generally quantify the relationship between read length and sequencing accuracy, and assembly quality; nonetheless, these are important factors influencing assembly quality that, if improved without detracting from other parameters, can facilitate improved *de novo* assembly overall.

3.2 Long amplicon sequencing for taxonomic classification, viral gene diversity, and functional genomics

A high-throughput sequencing technology capable of long and highly accurate reads would also be extremely useful for sequencing a highly complex population of long and very similar (but not identical) nucleic acid species. Important examples of such applications include metagenomic community diversity profiling and organismal classification, viral gene and genome diversity characterization, and the sequencing of populations of highly related synthetic nucleic acid molecules being subjected to functional

analysis. While each of these applications is itself a broad grouping of different experimental schema, they all share a similar analytical challenge. Each requires the ability to accurately characterize, in high throughput, a population of nucleic acids that is at once highly diverse with respect to the number of distinct molecular species but also highly similar in that the sequence divergence of any two species may be as low as or even below 1%[59,89]. Furthermore, these sequences are often at least ~300 nucleotides in length, and, for synthetic functional studies, read length may directly limit the size of the domain under study[60]. Therefore, a reasonable target is a massively parallel sequencing method capable of Sanger-like performance, *i.e.* 500 nucleotide reads with a substitution error rate below 0.1%.

3.2.1 Taxonomic classification by rRNA sequencing

Since the 1970s, evolutionary classification of organisms has relied heavily on sequence analysis of the 16S (in prokaryotes) or 18S (in eukaryotes) ribosomal RNA (rRNA) gene due its deep conservation and slow but consistent divergence over evolutionary time[90]. This method is especially popular in the field of microbiology, both as a research tool to explore microbial community diversity[91] but also as a practical method for species classification in more applied settings, including clinical medicine[92]. As mentioned above, this gene is deeply conserved, facilitating its use to reconstruct extremely ancient relationships, but also necessitating that relatively long segments (*i.e.* at least 400 bp) of the gene must be analyzed to accurately classify organisms[91]. (Incidentally, high sequence identity across species also facilitates sensitive and inclusive isolation of the gene from diverse microbial communities using simple experimental protocols.)

From a methodological standpoint, classification of a small number of putatively homogeneous microbial cultures is straightforward: PCR and Sanger sequencing is well suited to this task. However, recent interest in the large-scale characterization of diverse and complex populations of unicellular organisms, *i.e.* metagenomics, has necessitated the development of larger-scale methods for 16S rRNA sequencing. In clinical and environmental metagenomics, a population of organisms is sampled and directly subjected to sequence analysis in order to avoid biases in community structure that could be introduced by intermediate culture steps. These populations may be highly diverse with relative abundances of different operational taxonomic units (OTUs) varying over multiple orders of magnitude[93]. Therefore, an effective

high-throughput sequencing method must be both capable of long reads in order to phase variation present across variable domains of the 16s rRNA gene and accurate base-calling in order to confidently assign reads to OTUs and especially to detect low-abundance OTUs. Since 2005, the Roche/454 MPS platform has been widely applied for this purpose[26]; however, as described above, this is one of the least cost-effective MPS platforms and has a highly systematic error profile. Alternate methods capitalizing on the most cost-effective platforms, and with lower error rates, would be broadly useful to this community.

3.2.2 Viral gene diversity

Viruses employ a wide variety of mechanisms to evade the host immune system[94]. RNA viruses, which are replicated via RNA polymerase or reverse transcriptase, often rely on direct mutation of immunogenic epitopes to thwart the production of broadly effective antibodies. One particularly poignant and medically important example of this phenomenon is the Human Immunodeficiency Virus[95] (HIV). Owing to the high error rate of reverse transcriptase (approximately 3 per 100,000[96]), a given HIV infection is composed of a complex population of closely related but genetically distinct viruses. In this population lie rare sub-clones that have the potential to expand and evade the host immune response in the event that a specific neutralizing antibody is produced against the dominant sub-clone[97]. Furthermore, the *env* gene encoding gp160 (which is cleaved by host proteases to yield gp120 and gp41) is long, comprising nearly one third of the viral genome, or ~3 kb. As was the case for 16S sequencing, analysis of viral diversity has typically relied on cloning or single molecule PCR, followed by Sanger sequencing. Furthermore, as with 16S rRNA sequencing, recent efforts have been directed at adaptation of the Roche/454 MPS platform to HIV diversity profiling[98,99]. However, pyrosequencing is even more problematic for this application, as the error rate and pattern renders detection of low-frequency variants very difficult. Research and clinical characterization of viral diversity would therefore benefit substantially from a high-throughput and cost-effective sequencing method capable of long, highly accurate reads.

3.2.3 Synthetic functional genomics

A holy grail of genetics is the accurate prediction of phenotype from genotype. With the advent of MPS technologies and the resultant possibility of routine whole-exome or whole-genome sequencing in clinical medicine, this capability has never been more urgently needed. However, methods to interpret variation

remain rudimentary even for protein coding regions[100], and are essentially non-existent for non-coding sequences, despite their potential phenotypic importance[19]. While a variety of observational data are available to aid in interpreting variation, including evolutionary conservation (for both coding and non-coding variation) and protein-DNA interaction datasets[19] (specifically for non-coding variation), these methods are best suited to identifying functional elements and nucleotides (in the case of non-coding sequence) and protein domains and amino acids (for coding variation). They are not substantially informative when it comes to predicting the phenotypic impact of a given variant. To improve predictive power for specific variants, direct measurement of the functional impact of specific individual mutations is needed. Indeed, such methods for high-resolution dissection of regulatory elements[101] and coding sequences[102] were described more than twenty years ago. However, these methods are extremely labor-intensive, and, as such, have not been applied to a sufficiently broad extent. In the last several years, the Shendure and Fields groups at the University of Washington have dramatically scaled these assays by harnessing the throughput of MPS as a method both of determining genotype and quantifying functional impact[59,60], with impressive results. Nevertheless, these methods were still limited in an important dimension: the size of functional element or coding region being studied. This limitation was in large part imposed by the short reads and high error rates associated with MPS. Therefore, studies of the functional impact of coding[60] and regulatory[59] variation, and other types of synthetic functional genomics experiments (*e.g.* mutagenesis of non-coding RNA genes) would benefit substantially from a high-throughput sequencing method offering long and highly accurate reads.

3.3 Deep sequencing of cancer-related genes in clinical samples

A fundamental goal of oncology is to increase the precision of cancer care, both with respect to prognosis and therapy. Considering therapy first, researchers have long sought treatments that selectively kill cancer cells and spare normal cells. However, for nearly fifty years after Sidney Farber and Louis Diamond first showed that a metabolic antagonist (aminopterin, a folate analog) could induce remission of pediatric leukemia in some patients[103], clinicians were almost entirely limited to using broadly cytotoxic chemotherapeutic agents whose only selectivity was for rapidly dividing and/or metabolically active cells[104]. (One exception to this rule is the use of estrogen receptor antagonists such as tamoxifen in

breast cancer[105]; still, these agents are not uniquely tropic for cancer cells.) While effective and sometimes curative chemotherapeutic regimens have been developed using these compounds, their non-specific nature generally causes systemic toxicity and thus severe side effects, limiting dosing and compliance. The late 1990s saw a paradigm shift with the development of the targeted therapeutic imatinib, which is a tyrosine kinase inhibitor that inhibits growth of chronic myelogenous leukemia cells harboring the *bcr-abl* translocation mutation[106]. Subsequently, several genotype-specific therapies have been developed, and many more are in clinical trials[107]. The advent of genotype-specific therapies is an important dimension of the personalization of cancer care; for individuals whose tumors harbor the mutations that render the neoplasm sensitive to a particular targeted agent, that agent can have dramatic therapeutic impact including long-term remission and even cure. Even in the absence of an effective targeted therapy, genotype-specific prognosis (*e.g.* monosomy of chromosome 3 in uveal melanoma[108]) is an emerging and potentially transformative aspect of personalized cancer care.

A critical step in the development and implementation of targeted therapies is effective genotyping of clinical tumor samples for mutations in genes of interest. Massively parallel sequencing has great promise to facilitate and even accelerate this process as a cost-effective and broad genotyping method; however, there are several barriers to the application of MPS in this area. First, cancer is fundamentally a disease of mutation and clonal expansion, and clinically relevant mutations necessarily arise in single cells before expanding to high frequency due to selective advantage[109]. It is therefore of particular interest to be able to detect very low-frequency mutations, which may predict the development of resistance to a given targeted therapy (*e.g. EGFR* T790M in lung adenocarcinomas treated with *abl* kinase inhibitors[110] or *KRAS* mutations in colorectal cancers treated with anti-EGFR antibodies[111,112]). Second, further complicating this problem is the fact that in clinical samples, cancer cells are often admixed with non-neoplastic stromal cells, reducing the apparent allele frequency even of somatic mutations that are fixed in the population of neoplastic cells. Third, despite the plummeting cost of MPS, whole-genome deep sequencing of research and clinical samples remains time- and cost-prohibitive. Instead, sequencing resources are typically focused on regions of interest that are likely to harbor interpretable (and therefore clinically significant) mutations[113] at both low and high frequencies. Fourth, the vast majority of both newly

isolated and archival clinical tumor specimens have been formalin-fixed and paraffin-embedded (FFPE), limiting the quality and quantity of available DNA.

Therefore, an ideal method for clinical cancer gene sequencing would have the following properties: (a) a simple and rapid workflow such that it is easily scalable from single samples to hundreds of samples, and easily implemented in clinical and research settings; (b) high sensitivity and positive predictive value to detect and precisely quantify sub-clonal variation at very low frequencies; (c) compatibility with genomic DNA derived from FFPE tissues; (d) modular, flexible, and scalable targeting to regions of interest; and (e), low per-sample cost. As with previous applications discussed in this chapter, the error rates of MPS platforms have limited their utility for this field. However, in contrast to the previously discussed applications, MPS read lengths are sufficient for cancer gene resequencing.

3.4 Conclusion

In summary, a number of exciting and important areas of biomedical research and clinical medicine would benefit substantially from a cost-effective and high-throughput sequencing platform that produces highly accurate, long reads. These areas include (but are not limited to) *de novo* genome assembly, metagenomic diversity profiling, deep characterizing of populations of viral genes and genomes, synthetic functional genomics, and deep sequencing of cancer-related genes from clinical tumor samples. While the most widely available massively parallel sequencing platforms meet the throughput and cost requirements, error rate and read length limitations have largely precluded their implementation towards these important applications. Methods are therefore needed that leverage MPS but retain the read lengths and error rates of capillary sequencing.

Chapter 4 Parallel, tag-directed assembly of locally derived short sequence reads

This chapter is based on the following published paper:

Joseph B Hiatt, **Rupali P Patwardhan**, Emily H Turner, Choli Lee and Jay Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods*, 7, 119 - 122 (2010).

Bold face indicates equal contributors.

Emily Turner and Jay Shendure conceived the initial approach. I led the development of the subassembly method to its published form and performed the majority of experimental work, with contributions from Emily Turner. Rupali Patwardhan developed the computational framework to perform data analysis. Rupali Patwardhan, Emily Turner, and I performed data analysis. Choli Lee performed Illumina sequencing. Jay Shendure and I wrote the manuscript with contributions from Rupali Patwardhan and Emily Turner.

4.1 Summary

"Subassembly" is an *in vitro* library construction method that extends the utility of short-read sequencing platforms to applications requiring long, accurate reads. A long DNA fragment library is converted to a population of nested sublibraries, and a tag sequence directs grouping of short reads derived from the same long fragment, enabling localized assembly of long fragment sequences. Subassembly can be applicable in a variety of contexts such as accurate *de novo* genome assembly, metagenome sequencing, rare variant detection and sequencing of long, randomly assembled synthetic DNA molecules.

4.2 Introduction

The cost and throughput advantages of massively parallel sequencing are offset by large tradeoffs with respect to read length and accuracy[16]. Although the availability of reference assemblies renders short reads sufficient for genomic re-sequencing and digital profiling[18,114], other areas such as metagenomics[115], *de novo* assembly of complex genomes[85], immunoglobulin diversity profiling[116] and molecular haplotyping[117] are more challenging. In metagenomics, for example, sequences are derived from a population of related and unrelated genomes with highly varying abundances and a potentially enormous effective complexity. For identifying new open reading frames and for resolving related sequences within such a population, long reads remain indispensable[115].

As a means to deliver long reads using existing short-read massively parallel platforms, we developed a multiplex, *in vitro* strategy, termed subassembly, which is conceptually analogous to hierarchical shotgun genome assembly (Figure 4.1). In this approach, one of the two reads from a paired-end read serves as a sequence tag that identifies groups of short reads sharing a clonal origin, that is, deriving from the same longer DNA fragment (~500 bp). Each group of short, locally derived reads is then collapsed to a long, subassembled (SA) read. To evaluate performance, we applied this method to two samples: genomic DNA from a (G+C)-rich organism, *Pseudomonas aeruginosa* strain PAO1, and a previously characterized metagenomic sample from lake sediment[93].

4.3 Method overview

For subassembly, we sheared DNA to relatively long lengths (for example, ~500 bp), ligated 'tag-adjacent' adaptors to the fragments and then diluted and PCR-amplified these fragments (Figure 4.1 and 0: Methods). The dilution step before PCR imposed a complexity bottleneck, such that a limited number (~10^5–10^7) of long fragments were amplified to high abundance (0: Methods). The PCR amplicons were concatemerized and then sonicated, and a single 'breakpoint-adjacent' adaptor was ligated to the sheared fragments. We performed a second round of PCR in which one primer corresponded to a tag-adjacent adaptor and the other primer corresponded to the breakpoint-adjacent adaptor. The resulting amplicons effectively comprise a population of nested sub-libraries derived from the original long-fragment library. The tag-adjacent adaptor provides access to genomic sequence that corresponds to the ends of the long fragments. As this end sequence will be consistent across amplicons derived from the same long fragment, it can serve as a tag to identify molecules that are clonally derived. After paired-end sequencing, the read primed by the tag-adjacent adaptor identifies the original long DNA fragment, and the read primed by the breakpoint-adjacent adaptor represents sequence from a shearing-determined breakpoint in that fragment. As a relatively short read could serve as a unique tag identifier, we obtained paired-end reads of unequal length (20-bp 'tag read' and 76-bp 'breakpoint read'). In the analysis, we used tag reads to group breakpoint reads and separately subjected each tag-defined read group (TDRG) to local assembly with *phrap*[6].

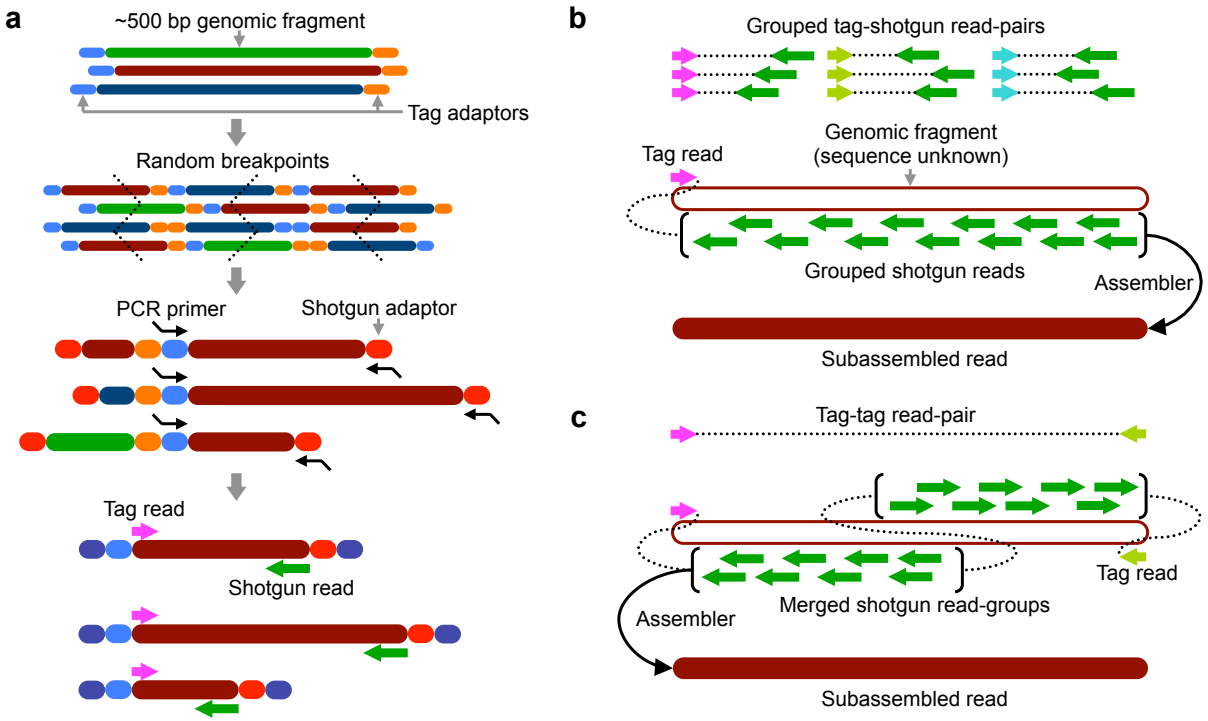


Figure 4.1. Subassembly method overview.

(a) Long DNA fragments are ligated to tag-adjacent adaptors, diluted and PCR-amplified. Dilution imposes a complexity bottleneck so that a limited number of long fragments are amplified. Concatemerized PCR products are then sheared by sonication and ligated to a breakpoint-adjacent adaptor. A second PCR amplification prepares amplicons for sequencing; one end of these amplicons corresponds to an end of a long fragment and the other end corresponds to a shearing breakpoint internal to that fragment. (b) Breakpoint reads are grouped *in silico* based on the sequence of the corresponding tag read. Breakpoint reads within a group, which derive from positions internal to the same parent long fragment, are subjected to local assembly to generate a subassembled read. (c) The metagenomic bottlenecked long-fragment library is subjected directly to paired-end Illumina sequencing to identify pairs of tag reads that were derived from opposite ends of the same original fragment. Two groups of breakpoint reads defined by distinct tag reads are merged and assembled together to generate one or more subassembled reads. In this study, this step was only applied to the metagenomic sample.

4.4 Application of subassembly to P. aeruginosa genome assembly

To rigorously assess performance, we applied subassembly to *P. aeruginosa* strain PAO1. After fragmenting genomic DNA, we size-selected it to ~550 bp (Figure A.1a) and processed the sample as illustrated in Figure 4.1. We used Illumina Genome Analyzer II (GA-II) to generate 56.8 million read pairs. We grouped the read pairs into TDRGs by the 20-bp tag (0: Methods) and separately subjected 76-bp breakpoint reads in each TDRG to local assembly with *phrap* to produce SA reads (Table A.1). We discarded SA reads not derived from identically oriented breakpoint reads (1.2%) and those failing

subassembly entirely (2.7%). For subsequent analyses, we considered only the longest SA read from TDRGs with ≥10 members.

This subset comprised 1.03 million SA reads with a median length of 338 bp (Figure 4.2a; Table A.2). The bimodal distribution may be due to uneven coverage of the original fragment secondary to imperfect size selection (Figure A.2). To assess quality, we mapped the SA reads to the *P. aeruginosa* strain PAO1 reference[118] and found that 99.82% had significant ($p < 10^{-6}$) alignments with basic local alignment search tool (BLAST)[119], with 98% of SA reads aligning along ≥95% of their full lengths. Although the contributing Illumina reads had an error rate of 2.4%, the substitution error rate of aligning SA reads was 0.25%. The longest correct SA read was 680 bp, likely an outlier from the gel-based size selection but nonetheless an indicator of the method's potential. We also estimated quality scores for bases in SA reads from the quality scores of contributing breakpoint reads (0). The 85% of bases in SA reads with the highest estimated quality scores were >99.99% accurate with respect to substitution errors when compared to the P. aeruginosa strain PAO1 reference (Figure 4.2b). Finally, we calculated the substitution error rate as a function of position along the SA read. The low overall error rate of one per 400 bp was maintained for hundreds of bases in the SA reads (Figure 4.2c).

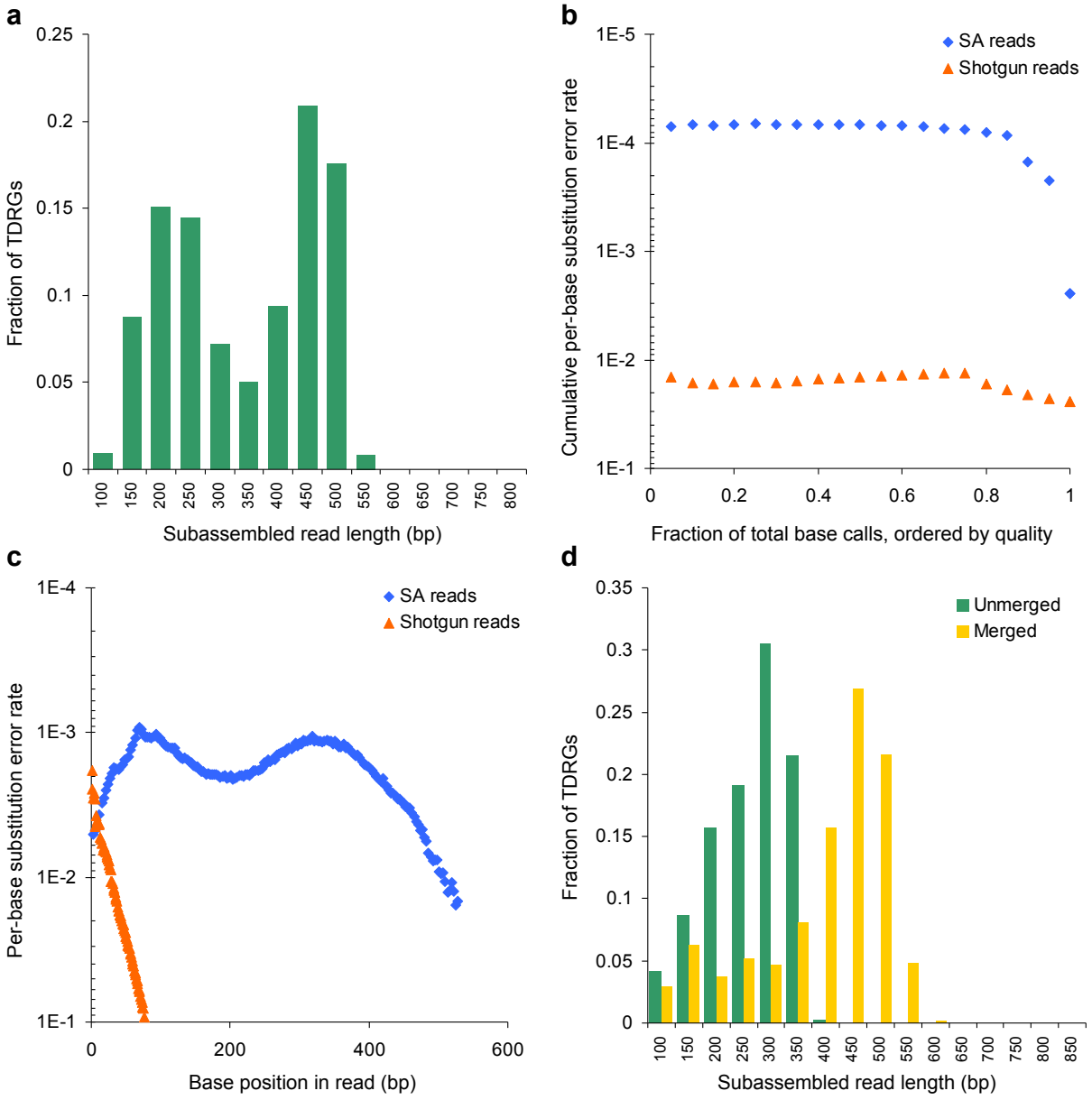


Figure 4.2. Evaluation of subassembly performance.

(a) Distribution of subassembled (SA) read length for P. aeruginosa sample and for methylamine metagenomic sample for unmerged and merged pairs of tag-defined read groups. (b) Cumulative per-base substitution error rate of base calls binned as a function of descending base quality in raw and SA reads, or the error rate of the x% of bases with the highest quality scores, after using BLAST to define the corresponding sequence in the reference. (c) Substitution error rate of base calls as a function of base position in raw and SA reads (binned every 3 bases). (d) Total length in sequences longer than a variable cutoff produced from SA reads compared to a standard shotgun library for the 100–1,000 bp range in which metagenomic analyses become possible. SA reads and assembled SA reads were compared to assembly of 48-bp or 76-bp paired-end reads from a standard Illumina shotgun library using Velvet with optimized parameters and an equivalent amount of raw sequence. Assembled SA reads refers to contigs produced by CABOG from SA reads.

Based on alignment with BLAST, SA reads covered 98.85% of the reference at a mean coverage of 63-fold. We observed bias against regions of extremely high G+C content (>70%) relative to shotgun sequencing (Figure A.3), which could be mitigated by optimizing PCR conditions. We also observed slight systematic bias in the distribution of SA read quality scores across the reference that we conclude is unlikely to compromise accuracy at positions with adequate coverage (Figure A.3).

To explore the utility of subassembly for de novo genome assembly, we assembled all filtered SA reads using CABOG[120], resulting in 708 contigs ≥1 kilobase (kb) with an N50, or the length x such that 50% of the genomic length is in sequences at least x long, of 15 kb (Table 4.1). The substitution error rate was ~1/14,000, and there was a total of 65 bp of inserted or deleted sequence across 31 contigs. Contigs ≥20 kb, which comprised 2.3 Mb, were more accurate, with a substitution error rate of ~1/250,000 and 20 bp of insertion-deletions across eight contigs. BLAST alignment predicted 11 contigs ranging in size from 1 to 18 kb to contain local misassemblies, but four of these were related to differences between the strain used here and the reference (Note A.2), leaving only seven true misassemblies. Six of these were very local deletions or expansions of <400 bp (within contigs <20 kb long), and one 1,125 bp contig displayed a more complex BLAST alignment.

Table 4.1. De novo assembly of *P. aeruginosa* genome using subassembled (SA) reads

Input	Assembly strategy	# of contigs / scaffolds	Contig / scaffold N50	Longest contig / scaffold	Total sequence	Reference coverage
SA reads	Celera	708	15,070 bp	160,221 bp	6.07 Mb	96.2%
SA reads + PE fragment + jumping mate-pair	Celera + scaffolding	32	444,483 bp	915,353 bp	6.11 Mb	99.3%

Assembly of SA reads from *P. aeruginosa* using the Celera assembler produces long and accurate contigs and can be further extended by scaffolding contigs with short (~200 bp) and long (~2.5 kb) mate-pairing data. Listed is the data used for assembly, the assembly strategy (we used a custom scaffolding algorithm), the number of contigs (for SA reads, ≥ 1 kb) or scaffolds (for SA reads with shotgun data, ≥ 5 kb), the contig or scaffold N50, the longest contig or scaffold, and the coverage of the reference genome. Physical coverage (sequence covered by contigs and N's spanning contigs) is shown for the assembly derived from SA reads supplemented with paired-end and mate-pair data.

Shotgun assembly of SA reads therefore resulted in long and highly accurate sequences with contiguity likely limited by sequence content biases. To facilitate scaffolding, we included sequencing data from one

lane of a paired-end fragment library (2 × 36 bp; insert size ~200 bp) and one lane of a mate-paired jumping library (2 × 36 bp; insert size ~2.5 kb). Using a custom iterative scaffolding algorithm (0), we generated 32 scaffolds ≥5 kb, with scaffold N50 of 445 kb, longest scaffold of 915 kb and 99.3% physical coverage of the reference (Table 4.1). Notably, scaffolding introduced only one misassembly, likely because of the presence of multiple nearly identical phage-like insertions (Note A.2). Our results, which were generated from a single platform, compare favorably to summary statistics of a published *de novo* assembly from a related organism that had been generated by combining long-read 454 and short-read Illumina data[121] (Note A.3).

To evaluate subassembly on a complex metagenomic sample, we used total DNA isolated from lake sediment and enriched for methylamine-fixing microbes [93]. We started with a slightly shorter long-fragment library (~450 bp; Figures A.1b, Figure A.4) and imposed a more stringent complexity bottleneck by diluting the long-fragment library to ~105–106 molecules before PCR (0: Methods). We obtained 21.8 million read pairs, which resulted in 262,298 TDRGs, in which the median length of the longest SA read in filtered TDRGs was 256 bp (Table A.2; Figure 4.2a).

In addition to the nested breakpoint reads that we used to produce SA reads, we also obtained 1.8 million paired-end reads from the original long-fragment library (2 × 20 bp), allowing us to merge TDRGs whose tags were observed as a read pair (Figure 4.1). We merged ~68% of the metagenomic TDRGs in this fashion. Subjecting breakpoint reads from merged TDRGs to local assembly yielded SA reads with a median length of 408 bp (Figure 4.2a; Table A.2).

4.5 Application of subassembly to metagenomics

We hypothesized that localized, tag-directed assembly would be particularly useful in the context of metagenomics, for which the highly non-uniform representation of organisms complicates de novo assembly from short reads. To test this, we generated a standard Illumina shotgun paired-end library from the same metagenomic sample and assembled reads from this library with Velvet[82] using optimized parameters (Table A.3; Figure A.5). We evaluated shotgun assemblies from both paired-end 76-bp reads and paired-end 48-bp reads. For both assemblies, we used 2.2 Gb of raw sequence, which was equal to the amount of data used for subassembly.

CABOG assembly of SA reads yielded considerably more total sequence data in longer contigs than direct assembly of shotgun reads, generating greater than twice as much sequence in contigs ≥200 bp (Figure 4.2d; Table A.3). Unassembled SA reads comprised greater than five times as much sequence ≥200 bp. Notably, shotgun assemblies did achieve greater contiguity at the longest lengths (Table A.3; Figure A.5). These long contigs may be due to deep sampling of the most abundant genomes. However, many are likely to represent misassemblies, as we did not observe long BLAST alignments to the available Sanger sequence data 93 or to any sequence in the GenBank nt or env_nt databases.

To conservatively estimate each method's effective coverage, we compared assembled contigs to 37.2 Mb of Sanger sequence data recently reported for the same sample[93] (0: Methods; Figure A.6). Although the complexity of the metagenomic sample likely remains undersampled, subassembly covered at least 45% more of the Sanger sequence reference when compared to contigs assembled from the paired-end short-read library. In addition, subassembly generated a comparable amount of total sequence as compared to Sanger sequencing data (39.5 Mb versus 37.2 Mb) in somewhat shorter contigs (median of 390 bp versus 835 bp) but with considerably less effort (three Illumina sequencing lanes versus hundreds of Sanger sequencing runs). In summary, subassembly produced substantially more sequence at lengths necessary for accurate phylogenetic classification[122] and gene discovery[123] than direct assembly from shotgun short reads and did so in better agreement with the available Sanger sequencing data, suggesting that the quality of assembled data may also be higher.

4.6 Discussion

Given that we observed accurate SA reads of nearly 700 bp, optimization of this method in concert with the tag-pairing approach (Figure 4.1) could potentially extend the effective length of SA reads to ~1 kb, that is, approaching the maximum length of Sanger sequencing data. One potential concern about the method as described is that tag sequences from different long DNA fragments can occasionally be identical by chance, especially if samples contain repetitive elements at high abundance. A simple modification would be to use a tag-adjacent adaptor containing an embedded degenerate sequence (for example, a randomized 20-bp segment), as this would completely decouple the tag sequence from the sample composition.

Finally, we note that subassembly offers a fundamental advantage in the way that a low error rate is achieved with a second-generation sequencing platform. Accurate assembly of short shotgun reads can be successful, provided that these reads are derived from relatively random sequence and that deep, uniform coverage can be otained[82]. Platforms such as Roche/454 offer long reads at a cost that is likely similar to subassembly (Note A.4) but have error profiles comparable to those of other second-generation sequencing platforms. Therefore, achieving high consensus accuracy also depends on the assumptions of uniform sampling and of a common origin for nearly identical reads. In contrast, because subassembly samples individual long DNA fragments and separately reconstructs a consensus sequence for each one, the production of long, accurate SA reads is insulated from nonuniform representation and sequence relatedness in the sample of interest.

4.7 Notes

4.7.1 Data availability

Raw Illumina sequence reads have been deposited to the NCBI Short Read Archive under the accession number SRA010316.

4.7.2 Acknowledgments

We thank L. Chistoserdova and M.G. Kalyuzhnaya (University of Washington) for the gift of the methylamine-enriched metagenomic DNA sample, C. Manoil (University of Washington) for the gift of *P. aeruginosa* strain PAO1 genomic DNA and P. Green for helpful discussions. J.B.H. was supported by US National Institutes of Health grant T32GM007266 and an Achievement Rewards for College Scientists fellowship.

Chapter 5 Functional dissection of enhancers

This chapter is based on the following published paper:

Rupali P Patwardhan, **Joseph B Hiatt**, Daniela M Witten, Mee J Kim, Robin P Smith, Dalit May, Choli Lee, Jennifer M Andrie, Su-In Lee, Gregory M Cooper, Nadav Ahituv, Len A Pennacchio and Jay Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30, 265–270 (2012).

Bold face indicates equal contributors.

Rupali Patwardhan performed the majority of experimental work. I developed experimental protocols and analytical algorithms for Subassembly to accurately reconstruct synthetic enhancer haplotypes, and contributed to designing other analytical phases. Specifically, I simplified and optimized experimental protocols (replacing physical with enzymatic fragmentation), and redeveloped the analysis pipeline from scratch (replacing *phrap* with a purpose-built, reference-guided consensus caller). Rupali Patwardhan, Daniela Witten and I performed all data analysis, with one significant exception noted below. Rupali Patwardhan, Jay Shendure and I wrote the manuscript. Mee J Kim, Robin Smith and Dalit May performed tail vein injections and RNA extraction from mouse livers. Mee J Kim and Robin Smith performed luciferase assays for validation of six individual ALDOB mutants. Choli Lee performed Illumina sequencing. Gregory Cooper contributed the evolutionary conservation analysis.

5.1 Summary

The functional consequences of genetic variation in mammalian regulatory elements are poorly understood. Here we report the *in vivo* dissection of three mammalian enhancers at single-nucleotide resolution through a massively parallel reporter assay. For each enhancer, we synthesized a library of >100,000 mutant haplotypes with 2–3% divergence from the wild-type sequence. Each haplotype was linked to a unique sequence tag embedded within a transcriptional cassette. We introduced each enhancer library into mouse liver and measured the relative activities of individual haplotypes *en masse* by sequencing the transcribed tags. Linear regression analysis yielded highly reproducible estimates of the effect of every possible single-nucleotide change on enhancer activity. The functional consequence of most mutations was modest, with ~22% affecting activity by >1.2-fold and ~3% by >2-fold. Several, but not all positions with higher effects showed evidence for purifying selection, or co-localized with known liver-associated transcription factor binding sites, demonstrating the value of empirical high-resolution functional analysis.

5.2 Introduction

Patwardhan et al[59] recent described a method called 'synthetic saturation mutagenesis' in which programmable microarrays were used to synthesize variants of core promoters, each in cis with a downstream tag sequence. The population of core promoter variants was subjected to a cell-free in vitro assay, after which sequencing of the transcribed tags was performed to quantify the relative activity of specific core promoter variants. This method is very effective in the context of core promoters, and potentially other small elements. However several aspects limit its broader application and scalability: (i) when each regulatory element variant is synthesized as a separate array feature, the overall cost of synthesis remains high; (ii) the separate synthesis of individual variants also limits how many combinations of mutations can be simultaneously programmed; (iii) the maximum length of array-synthesized oligonucleotides is currently 200–300 bp, whereas mammalian enhancers can be 1 kb or longer; (iv) access to array-derived oligonucleotide libraries remains restricted to a few groups; and (v) the cell-free, in vitro assay that we used poorly captures biological context.

To overcome these limitations and facilitate the high-resolution dissection of mammalian enhancers, we developed an improved method, termed massively parallel functional dissection (MPFD) (Figure 5.1). We then used MPFD to assess the extent to which all possible single-nucleotide variants (SNVs) affect the activity of three mammalian enhancers that are active in the liver, designated here ALDOB (hg19:chr9:104195570-104195828)[124-126], ECR11 (hg19:chr2:169939082-169939701)[127] and LTV1 (mm9:chr7:29161443-29161744).

5.3 Method overview

To apply the MPFD method (Figure 5.1) to the three enhancers of interest, each enhancer was synthetically constructed by polymerase cycling assembly using overlapping oligonucleotides (~90 bp) containing a programmed level of degeneracy. At each position, 97% of molecules were expected to be synthesized correctly with 1% doping of each possible single-nucleotide substitution (Appendix B: Methods). Therefore, each synthetic enhancer molecule contained, on average, three mutations per 100bp, randomly distributed along its length. The population of molecules was inherently complex, both with respect to representation of all possible SNVs of the wild-type enhancer as well as myriad unique combinations. Because nearly all synthetic enhancers contained multiple substitutions, they are referred to here as 'enhancer haplotypes'.

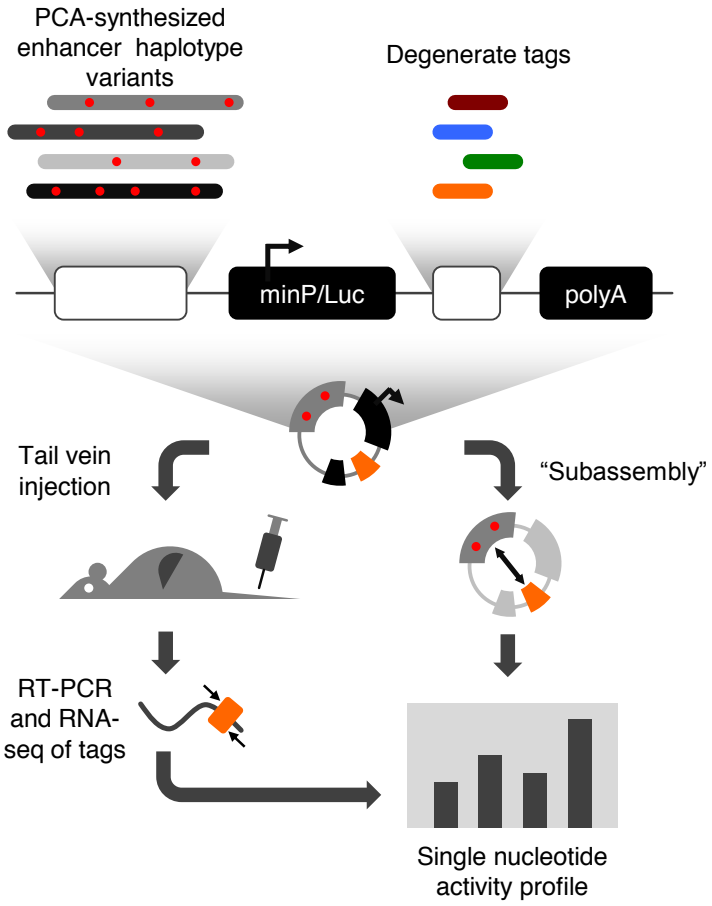


Figure 5.1. Overview of MPFD.

We used doped oligonucleotide synthesis and polymerase cycling assembly (PCA) to generate a highly complex library of enhancer haplotypes for each enhancer studied. On average, each enhancer haplotype diverged from wild type by ~2–3% (red circles represent mutations). These mutant enhancers, along with 20-bp degenerate tags, were cloned into an expression vector (pGL4.23) containing a minimal promoter driving transcription of luciferase (minP/Luc). We performed 'subassembly' on each library to determine the full sequence of each enhancer haplotype and to identify the 20-bp tag to which each haplotype was cloned in *cis*. Each library was then introduced into two mice through hydrodynamic tail vein injection, livers were harvested after 24 h and sequencing was performed to quantify abundance of transcribed 20-bp tags. These data were used to estimate the effect of each possible mutation on transcriptional activation.

Next, a library for assessing the activity of each enhancer haplotype was created by cloning the synthetic enhancers into a plasmid (Promega pGL4.23), which contains a minimal promoter upstream of the luciferase gene. In order to uniquely tag each enhancer haplotype, we cloned an oligonucleotide containing a 20-bp, fully degenerate subsequence to a separate site in the 3′ untranslated region (UTR) of the luciferase gene. The sequences of specific 20-bp tags cloned in *cis* with specific enhancer haplotypes were determined by massively parallel sequencing. As the enhancer haplotypes were highly related

sequences with lengths that exceeded the maximum read-length of the Illumina platform, we used tag-guided subassembly[128] to enable full-length, high-accuracy sequencing of individual enhancer haplotypes in association with their downstream tags. Each resulting library included >100,000 fully sequenced enhancer haplotypes, with nearly all containing multiple substitutions, and each associated with one or more unique tags.

The library was then subjected to what was effectively a massively parallel *in vivo* reporter assay. For the experiments described here, we used the hydrodynamic tail vein injection assay[127,129] to assess *in vivo* enhancer activity in the mouse liver. Mice were euthanized 24 h after injection, at which time total RNA was extracted from each liver, followed by RT-PCR and massively parallel sequencing of cDNA from transcribed tags.

5.4 Results

We studied three mammalian enhancers identified by diverse methods (Figure 5.2). ALDOB (259 bp) is a human intronic enhancer of the aldose B gene[124-126]. ECR11 (620 bp) is a human enhancer located in an intron of dehydrogenase/reductase SDR family member 9 (DHRS9)[127]. LTV1 (302 bp) is a candidate mouse enhancer located on the 3′ side of zinc-finger protein 36 (Zfp36) (Figure B.1a,b). The activity of each wild-type enhancer was confirmed using a conventional hydrodynamic tail vein injection assay, in which luciferase activity in liver tissue was measured 24 h after injection (Figure B.1c).

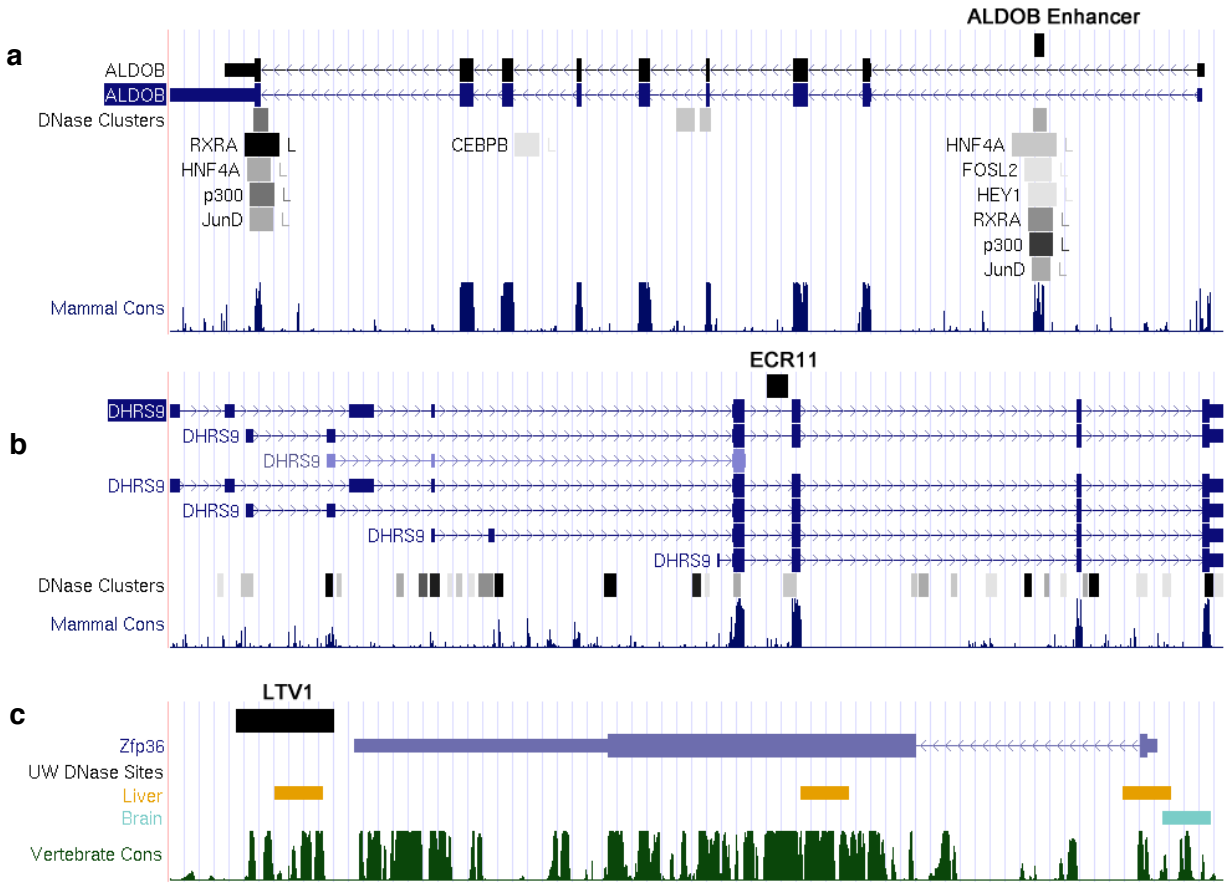


Figure 5.2. Schematics of candidate enhancer loci.

UCSC genome browser snapshots depicting the genomic locations of each of the three enhancers. Each enhancer was identified by diverse methods. (a) The human ALDOB enhancer is located in the first exon of ALDOB. It was identified and characterized by conventional transgenic assays 124-126, and overlaps extensively with heavily conserved clusters of ENCODE DNase hypersensitivity sites 130 and HepG2 ChIP-Seq peaks. (b) Human enhancer ECR11 is located in the fifth intron of DHRS9 in a region that overlaps with an ENCODE DNase hypersensitivity cluster on its 3` end and is conserved to mice. It was identified by comparative genomics and liver-specific transcription factor binding site analyses 127. (c) Mouse enhancer LTV1 is located immediately downstream (3') of Zfp36 in a conserved region and overlaps with DNase hypersensitivity sites 130 from liver tissue but not brain. This enhancer was first identified by p300 ChIP-Seq on early adult mouse liver [L.A.P., unpublished data]. Using deletion experiments, we isolated a functionally equivalent 302bp core element that was used for mutagenesis (Figure B.1a,b).

We applied MPFD to systematically dissect the functional consequences of all possible SNVs in these three enhancers. Sequencing with subassembly confirmed that the resulting libraries were complex, with a total of 641,135 distinct haplotypes associated with 1,186,696 tag sequences (Table 5.1). The observed number of mutations per haplotype approximated expectations, with ~2–3 substitutions per 100 bp (Figure B.2) and were well distributed (Figure B.3). All possible substitution variants of each enhancer were represented in ≥42 uniquely tagged haplotypes. On average, each position was disrupted on ~4,000

distinct enhancer haplotypes. Furthermore, all possible pairs of positions were disrupted in ≥1 haplotype with the exception of a single pair of positions in LTV1.

Table 5.1. Enhancer haplotype library characteristics

Library	Number of haplotypes	Number of tags	% of possible substitutions in at least one haplotype	% of possible pairs of sites in at least one haplotype	Per-base mutation rate per haplotype (mean ± s.d.)
ALDOB	378,450	406,071	100% (777 of 777)	100% (33,411 of 33,411)	0.021 ± 0.010
ECR11	105,795	105,832	100% (1860 of 1860)	100% (191,890 of 191,890)	0.023 ± 0.006
LTV1 rep. 1	119,950	403,869	100% (906 of 906)	99.99% (45,449 of 45,451)	0.031 ± 0.010
LTV1 rep. 2	105,188	270,924	100% (906 of 906)	99.99% (45,449 of 45,451)	0.031 ± 0.010

For each library of enhancer haplotypes, we list the number of distinct haplotypes, the number of tags with which those distinct haplotypes are associated in *cis*, the percentage of possible single nucleotide substitutions that are present in at least one haplotype, the percentage of possible pairs of positions where both positions contain mutations together in at least one haplotype, and the per-base mutation rate in each library.

We introduced each library (one each for ALDOB and ECR11, and two independently constructed libraries for LTV1) into two mice by hydrodynamic tail vein injection (Figure B.1d). Total RNA from each mouse liver was split into several aliquots (ALDOB: N = 39; ECR11: N = 69; LTV1-1: N = 10; LTV1-2: N = 10), with each aliquot separately subjected to RT-PCR with primers flanking the 20-bp tag located in the 3′ UTR of the luciferase transcriptional cassette, and then to massively parallel sequencing on an Illumina GAIIx. Because target RNA was very scarce relative to cellular RNA, a modest number of target RNA molecules contributed to each RT-PCR, leading to a complexity bottleneck. In other words, within each sequencing library, all reads corresponding to any single tag appeared to have been derived from amplification of a single RNA molecule. We therefore used the number of RNA aliquots in which a

particular tag was observed, and not the total number of reads associated with a tag, as a measure of the relative transcriptional activity of its associated enhancer haplotype.

For each position in each enhancer, we constructed a linear model to assess the extent to which the presence of a mutation at that position is predictive of a change in the number of RNA aliquots in which an enhancer haplotype was observed, which is effectively a proxy for its effect on transcriptional activation, that is, 'effect size' (Appendix B: Methods). Specifically, we use the term 'effect size' to describe the log2-fold change in the predicted transcriptional activity, as measured by the number of RNA aliquots in which a tag-associated haplotype appeared, relative to the wild type. We first sought to assess reproducibility, so we calculated effect sizes separately for the two independently constructed LTV1 libraries (combining data from the two mice subjected to each of these libraries). For ALDOB and ECR11, we calculated effect sizes separately on the data from each mouse. For these two types of biological replicates, the effect sizes were highly correlated (r = 0.96 for LTV1, r = 0.93 for ALDOB, r = 0.96 for ECR11). Because reproducibility was high and to increase resolving power, we performed all subsequent analyses after combining data across mice for each enhancer haplotype library (data for one of the two LTV1 replicate libraries is shown in Figure B.4).

We next recalculated effect sizes in two ways (Figure 5.3; Figure 5.4; Figure 5.5). First, as for the reproducibility analysis, we constructed separate linear models for each position where mutational status was encoded as a single binary variable representing whether an enhancer haplotype was wild type or mutant at that position (Figure 5.3a; Figure 5.4a; Figure 5.5a). Second, we constructed separate multiple linear regression models for each position with three variables, each corresponding to a particular nucleotide substitution at that position (Figure 5.3b; Figure 5.4b; Figure 5.5b). For each enhancer, we also constructed a multiple linear regression model incorporating all positions. These models were also significantly predictive ($p < 0.01$) (Note B.1; Table B.1), and yielded effect-size profiles similar to models constructed independently for each position (Figure B.5). As the coefficients from models constructed independently for each position are more naturally interpreted as position-specific effects, we used these models for subsequent analyses.

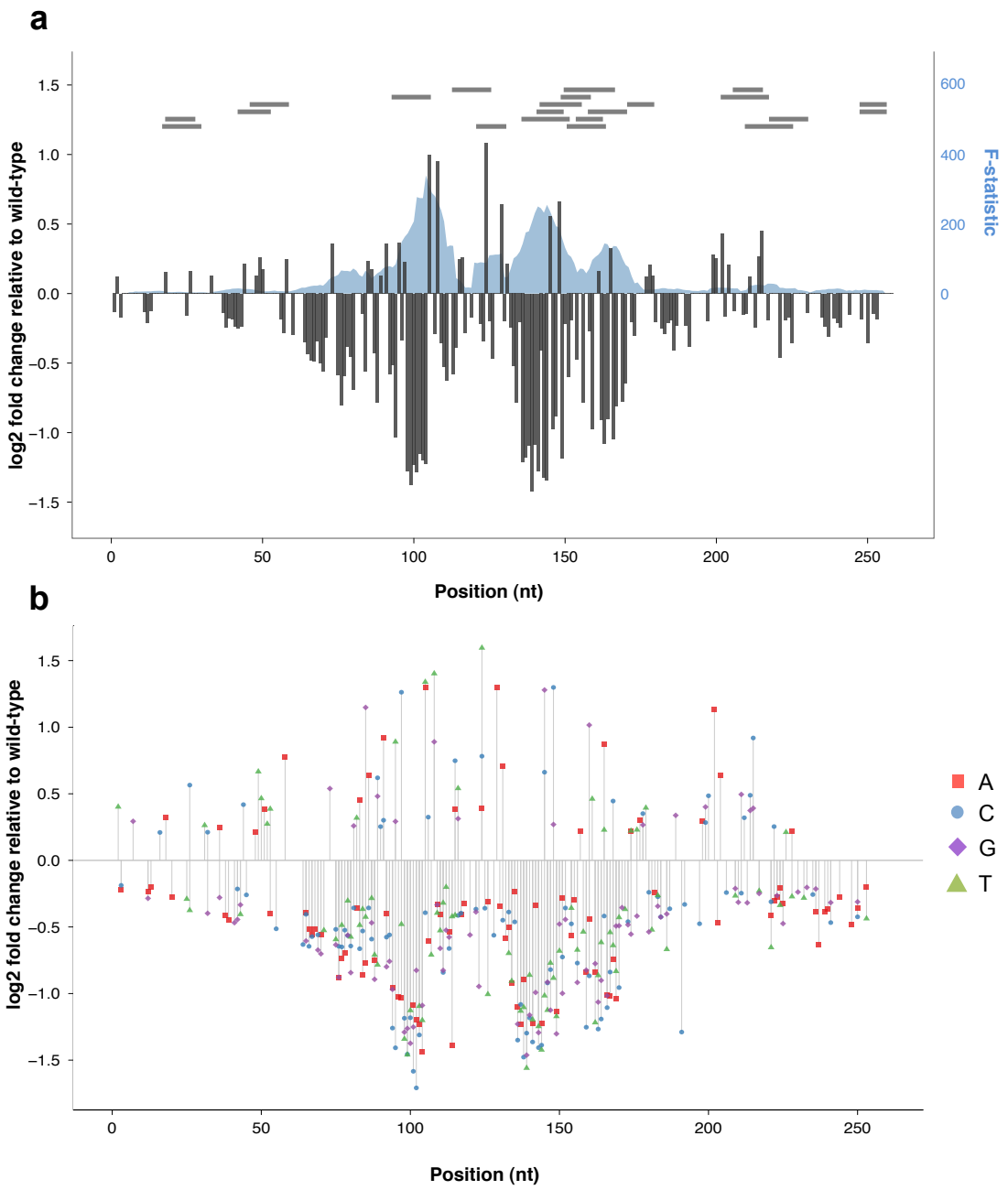


Figure 5.3. Effect size on transcriptional activity of all possible substitution mutations in ALODB enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for ALDOB ((a) and (b), respectively), Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 39). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel (a), blue shadow, right axis). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in (a).

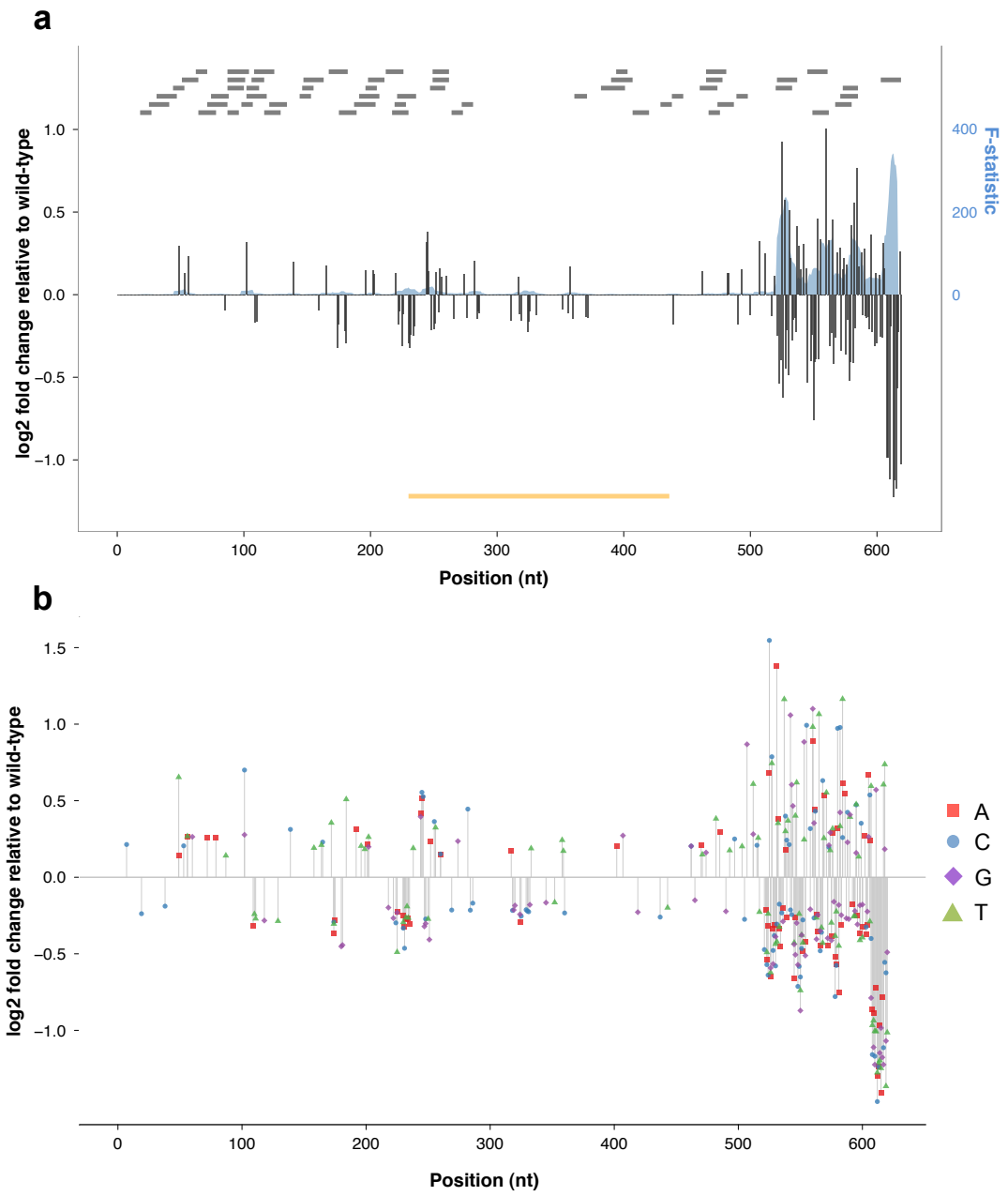


Figure 5.4. Effect size on transcriptional activity of all possible substitution mutations in ECR11 enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for ECR11 ((a) and (b), respectively). Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 69). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel (a), blue shadow, right axis). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in (a). The location of a partial LINE element in ECR11 is shown as an orange bar at the bottom of (a).

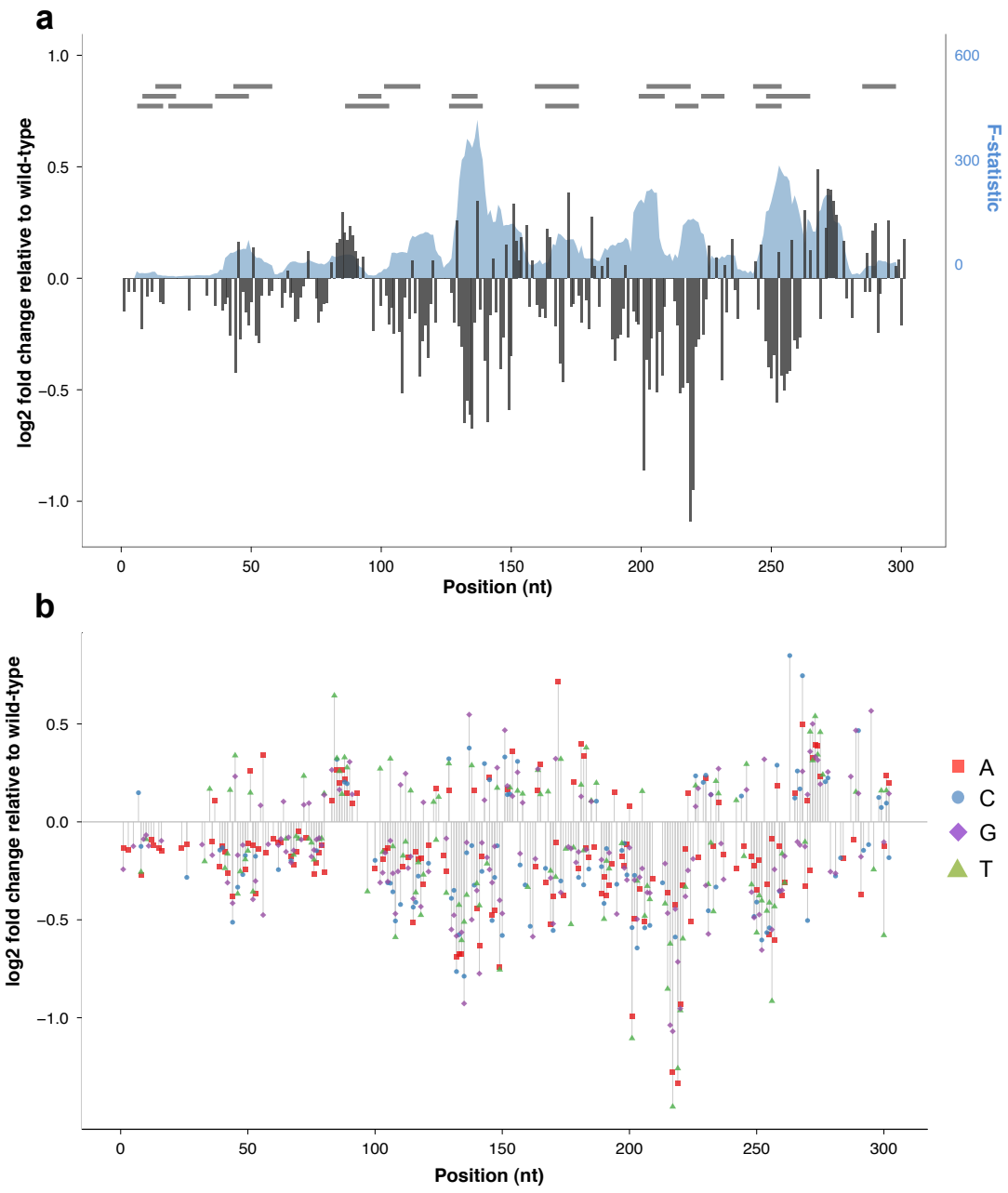


Figure 5.5. Effect size on transcriptional activity of all possible substitution mutations in LTV1 enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for LTV1 ((a) and (b), respectively). Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 10). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel (a), blue shadow, right axis). The locations of TFBS

predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in (a).

To provide further validation, we also performed site-directed mutagenesis to individually introduce the six mutations in ALDOB that were predicted to have among the largest effect sizes (three increasing activity and three decreasing activity), and tested these individually using the hydrodynamic tail vein luciferase assay (Figure 5.6). Observed luciferase fold-changes were highly correlated with effect-size predictions from the models (R = 0.985).

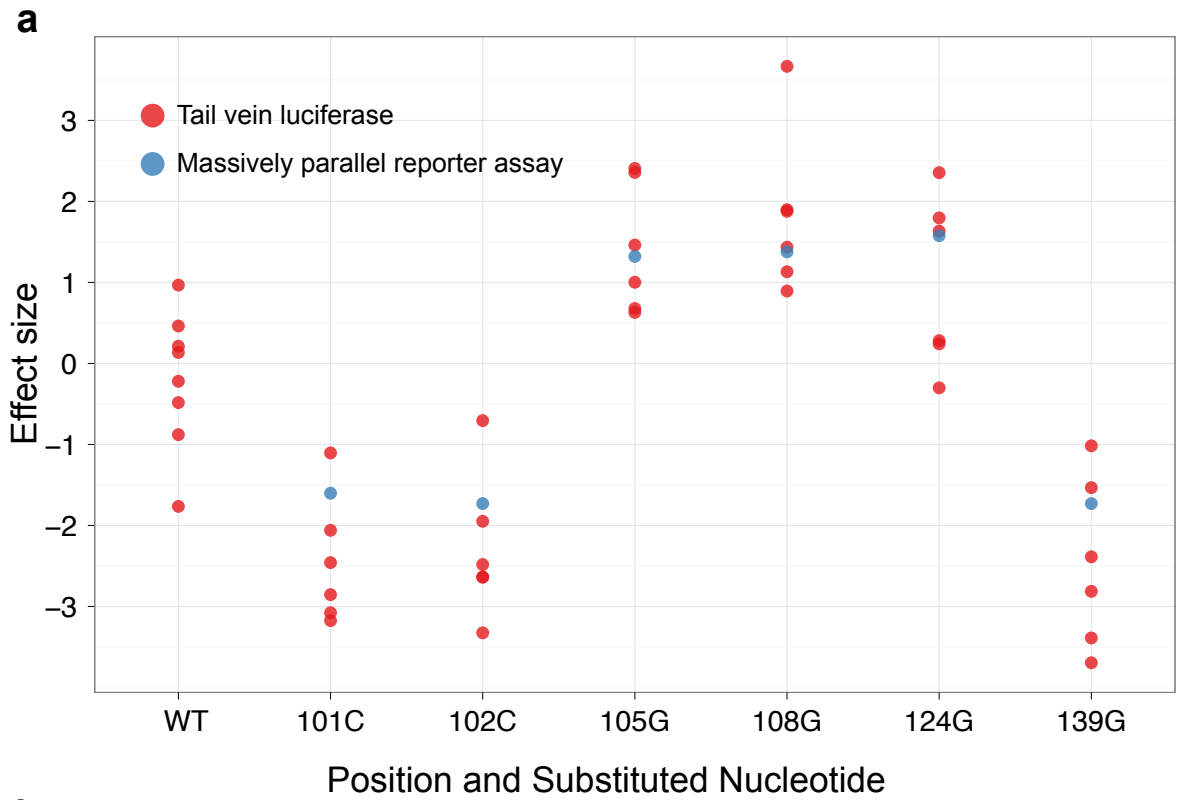


Figure 5.6. Validation of MPFD predictions using the hydrodynamic tail vein luciferase assay.

Shown are mutation effect sizes (log2 fold-change in expression of mutant compared to wild-type) for six single nucleotide ALDOB enhancer variants compared to the wild-type sequence. Each mutant was injected individually in at least six mice, and luciferase activity was measured at 24 hours post injection. Measurements from individual mice are shown. Effect sizes determined by the massively parallel reporter assay described here are shown for comparison. Effect sizes calculated via hydrodynamic tail vein luciferase assay were highly correlated with luciferase activity (R=0.985). On average, the observed fold-change in luciferase for individually tested mutations was ~25% greater in magnitude than the effect size predictions from our massively parallel reporter assay, although the predicted effect size based on the massively parallel reporter assay always fell within the range of effect sizes observed in the individual luciferase replicates. This may reflect differences between the assays or, alternatively, systematic but modest underestimation by our current methods.

5.4.1 Co-localization of high-impact positions and known TFBSs

Across each enhancer, the effect-size profiles exhibited spatial structure—that is, a clustering of positions with larger effect sizes. Positions separated by less than ~6 nucleotides had significantly correlated effect sizes ($p < 0.01$) (Figure B.6). To further explore this, we performed multiple linear regression using mutational status at ten adjacent positions (that is, a binary variable for wild-type or mutant) at a time (Appendix B: Methods). These models remained predictive of transcriptional activity in a spatially resolved pattern (Figure 5.3a; Figure 5.4a; Figure 5.5a). We suspected that these clusters of correlated positions might represent transcription factor binding sites (TFBSs). Indeed, when we predict TFBSs[131] (Figure 5.3a; Figure 5.4a; Figure 5.5a; Table B.2), we observe striking overlap between predicted binding sites and clusters of highly predictive positions (Figure 5.3a; Figure 5.4a; Figure 5.5a). For example, a predicted binding site for HNF4 in the ALDOB enhancer (bases 94-105) coincides with a highly predictive localized model (Figure 5.3a). Furthermore, all mutations in this region had negative effects on activity, with the notable exception of mutations that increased identity with the consensus HNF4 binding site, which were activating (e.g., 95A→G and 105T→A) (Figure 5.7a). The same pattern was observed for other predicted sites as well, for example, a predicted HNF1 binding site at bases 135-148 in ALDOB (Figure 5.7b). Notably, independent experiments have established that these two transcription factors drive this element *in vivo*[126]. The spatial patterns may also reveal or refine broader features of activity—for example, the boundaries of functional elements. For example, in ECR11, computational prediction yielded a large number of predicted liver-specific TFBSs in the proximal 300 bases[127], but we observed that the highest impact SNVs were largely confined to the distal 160 bases (Figure 5.4a; Figure B.7).

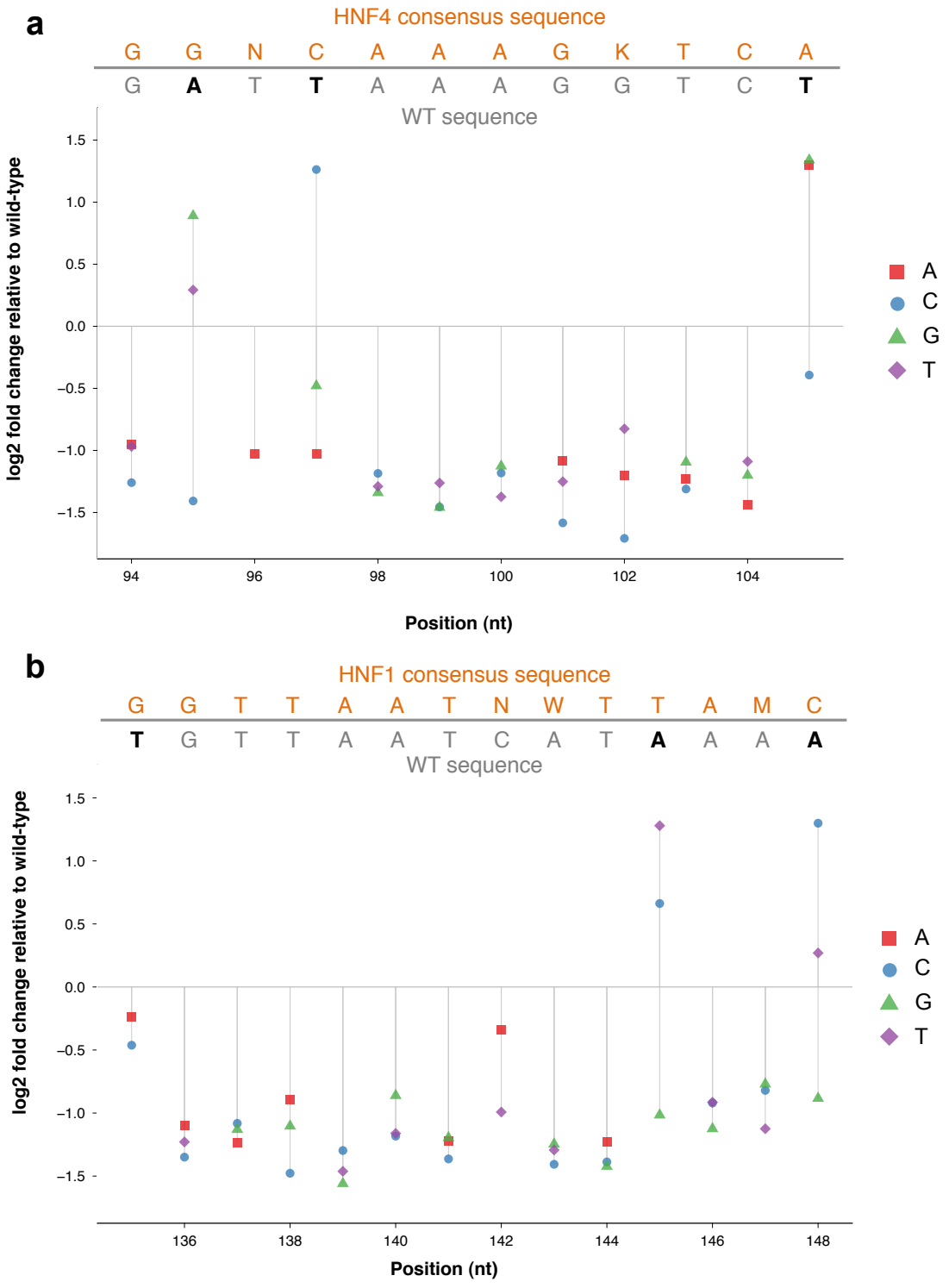


Figure 5.7. Profiles of mutation effect size in TFBSs.

For a predicted HNF4 site (positions 94–105) (a) and a predicted HNF1 site (positions 135–148) (b) in ALDOB, the effect size for each possible substitution, with the consensus TF binding sequence (orange) and the enhancer sequence (gray for consensus, black for nonconsensus) is plotted. Nonconsensus positions where rescue is observed after mutating to consensus are shown in boldface. HNF4 binding to

the ALDOB enhancer region in human liver has been previously demonstrated[132], whereas *in vivo* occupancy data for HNF1 at this region is not yet available.

5.4.2 Relationship between evolutionary and functional constraint

Evolutionary constraint in noncoding, regulatory DNA has frequently served as a proxy for functional constraint [133-135]. However, recent studies have shown that many enhancers are evolving rapidly and that mammalian genomes contain large numbers of evolutionarily young, sometimes species-specific, enhancers [132,136]. All three enhancers studied here are grossly conserved between human and mouse (Figure 5.2). We therefore investigated the relationship between functional constraint and evolutionary constraint at single-nucleotide resolution. For two of three enhancers, linear models, constructed to assess whether evolutionary constraint (that is, Genomic Evolutionary Rate Profiling (GERP)[137]) was predictive of functional constraint (that is, the absolute value of univariate model coefficients that we obtained), were significantly predictive with modest explanatory power (ALDOB: R^2 = 0.1232, P = 6.31e-9; LTV1: R^2 = 0.03911, P = 5.47e-4). For both enhancers, positions with the highest functional effect sizes were significantly associated with elevated evolutionary constraint scores (p < 0.01) (Figure B.8). However, not all positions with high GERP scores (≥4) had functional effect sizes in the top quartile for each enhancer (ALDOB: 33 of 61, 54%; ECR11: 5 of 25, 20%; LTV1: 0 positions with GERP≥4). These positions might have functions unrelated to the enhancer activity assayed here or might be of greater functional relevance in other contexts, for example, other tissues or developmental time points. On the other hand, a small set of highly functional positions, for example, most nucleotides within the distal-most C/EBP motif in ECR11, have lo GERP scores, consistent with lineage or species-specific activity.

5.4.3 Effect-size spectrum of single-nucleotide variants

A substantial proportion of polymorphisms and new mutations in mammalian genomes are single-nucleotide substitutions[138]. However, the functional dissection of regulatory elements has historically relied on introducing nested or scanning deletions, limiting the extent to which they inform the interpretation of naturally occurring variation. Our results provided an opportunity to examine the distribution of effect sizes of SNVs in mammalian enhancers on the magnitude of transcriptional activation (Figure 5.8). Notably, we observed that the majority of SNVs result in only a modest change in transcription relative to the wild-type enhancer. Overall, <25% of the mutations alter transcriptional activity

by >1.2-fold. Furthermore, only a few mutations, mostly in ALDOB, altered activity by a factor of >2. These results suggest that these enhancers are highly robust to the vast majority of potential SNVs. Further application of this method will be needed to assess whether this is a general property of mammalian enhancers.

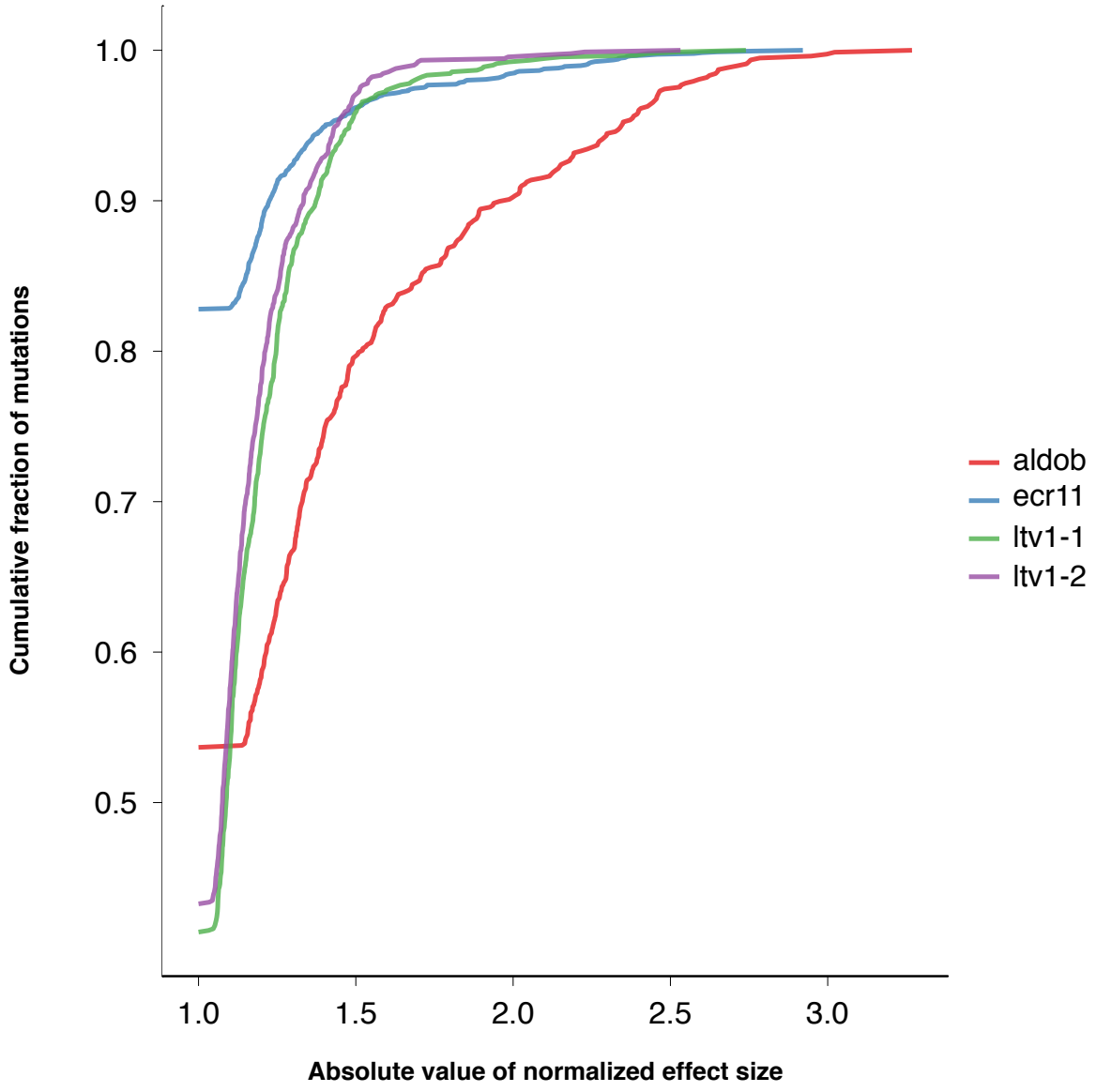


Figure 5.8. Distribution of effect sizes for all possible substitution mutations in three mammalian enhancers.

For the three enhancers studied (two replicate libraries for LTV1), the cumulative fraction of substitutions possessing a given effect size is expressed as the absolute value of the effect size of a given substitution. For example, across the three enhancers, between ~80% and ~95% of substitutions influence transcriptional activity by less than a factor of 1.5.

Perhaps as expected, the majority of functionally important mutations decreased activity (70% or 850/1,211). In general, only one substitution at a given position was activating, for example, substitutions that render a motif more like the consensus sequence (Figure 5.7). However, we observed some notable exceptions, including positions 83–93 and 272–278 in LTV1, where all or almost all substitutions were activating, consistent with binding of a repressive transcription factor. Positions 83–93 harbor a predicted binding site for NF-1, whereas there are no predicted sites in the immediate vicinity of positions 272–278, highlighting the value of experimental assessment of mutational impact.

5.4.4 Epistatic interactions

Finally, we sought to leverage the fact that our enhancer libraries contain multiple mutations on each haplotype to assess the degree of epistasis, or interaction, between positions in the enhancer. To obtain adequate power, we restricted our analysis to pairs of positions that were both mutated in at least 20 haplotypes. For each pair of positions that passed this cutoff, we built a multiple linear regression model consisting of three binary variables where the first two variables encoded mutation status (wild type or mutant) at each position independently and the third encoded whether both are mutant in a particular haplotype. With a false-discovery rate (FDR) cutoff of 0.05, we observed few pairs with a significant interaction term (ALDOB: 82 of 33,389, 0.25%; ECR11: 199 of 184,206, 0.10%; LTV1: 45 of 43,975, 0.10%), suggesting that the effects of multiple SNVs on the same haplotype are generally additive, or that our study lacked power to identify subtle interactions. Interacting pairs were significantly enriched for proximity (that is, pairs within 10 bp of each other versus pairs further apart, ALDOB: $P < 1e\text{-}4$; ECR11: $P < 1e\text{-}3$; LTV1: $P < 1e\text{-}4$), and we observed several different classes of interacting pairs with respect to the signs of the individual position effects and the sign of the interacting term (Table B.3).

5.5 Discussion

We developed a strategy to construct complex libraries of mammalian enhancers that contain all possible single-nucleotide substitutions and hundreds of thousands of distinct haplotypes. This method surpasses its predecessor[59] in terms of cost effectiveness, tunability, applicability to full-length regulatory elements and integration with an in vivo assay. We applied this method to empirically measure the distribution of effect sizes of all possible SNVs in three mammalian enhancers in an *in vivo* model. A key finding is that

the vast majority of SNVs in these enhancers have highly reproducible yet remarkably modest effects on transcriptional activation. The distribution suggests that enhancers are highly robust to single-nucleotide changes. We also find that most combinations of single-nucleotide changes have additive effects on function. As expected, there is a clear relationship between the magnitude of functional impact and the location of predicted TFBSs, although not all predicted TFBSs are functional, and not all functional motifs are associated with predicted TFBSs. Similarly, evolutionary constraint, although clearly correlated with the magnitude of functional impact, does not predict it well on a nucleotide-by-nucleotide basis.

There remain some limitations of the method. First, although we exploited a mouse tail vein assay to assess function *in vivo*, the regulatory elements are episomal and therefore may not be subject to the same mechanisms governing elements residing on chromosomes. For example, because of the size of the synthetic construct, we were unable to assess the effects of mutations that may influence long-range interactions between regulatory elements. This might be addressed in part by transitioning to a lentiviral system, which would facilitate use in additional tissues and may also enable the application of other assays, for example, ChIP-Seq, to enhancer variant libraries. Furthermore, our results must also be considered specific to the minimal promoter used here until other promoter classes are tested. Second, we have assayed these enhancers in a single tissue and at a single time point. The activity profile of specific positions could well be different in other tissues; this is the long-standing context problem[139]. Third, because of the scarcity of the target transcript relative to total RNA, we observed complexity bottlenecking, limiting the precision of our estimates of the effect size. This can be addressed by optimization of the RNA isolation step, for example, by hybridization-based enrichment. Fourth, we restricted our analysis to enhancer haplotypes containing only substitutions, as this was the dominant form of variation introduced during synthesis. To facilitate simultaneous dissection of the functional consequences of small insertions and deletions (indels), one could use reduced-fidelity oligonucleotide synthesis conditions, or polymerase cycling assembly with oligonucleotides containing programmed indels. Current efforts are directed at implementing these improvements, scaling this method to more enhancers and applying it to other classes of noncoding regulatory elements.

A fundamental goal of modern biology is to understand the human genome at single-nucleotide resolution. Single-nucleotide differences between genomes are causative for, or affect susceptibility to, a host of diseases, and single-nucleotide mutations are a primary source of raw material for evolution. We anticipate that the high-throughput, empirical measurement of the functional impact of single-nucleotide variants in enhancers will substantially facilitate the analysis of noncoding variants in genome-wide association study hits, the study of the mechanistic basis for enhancer activity and the engineering of enhancers with desired properties. Furthermore, with cost-effective, massively parallel methods for functional analysis, it may soon be realistic to empirically measure the functional effects of all possible single-nucleotide changes in all noncoding regulatory elements in the human genome.

5.6 Notes

5.6.1 Data availability

Raw sequencing reads available in the NCBI Short Read Archive under accession number SRA049159. A full list of mutations interrogated for this work, along with the associated effect sizes and P values, are provided as Supplementary Data on the Nature Biotechnology website.

5.6.2 Acknowledgments

We thank R. Qiu and J. Kitzman for advice on experimental strategies, and B. Cohen and D. Pe'er for helpful discussions. This work was supported in part by grants HG003988 from the National Human Genome Research Institute (L.A.P.), US National Institutes of Health (NIH) grant DP5OD009145 (D.M.W.), National Institute of General Medical Sciences (NIGMS) award number GM61390 (N.A.), National Institute of Child Health and Human Development (NICHD) grant number R01HD059862 (N.A.), the Pilot/Feasibility grant from the University of California, San Francisco Liver Center (P30 DK026743) (N.A.), AG039173 from the National Institute on Aging (J.B.H.) and a fellowship from the Achievement Rewards for College Scientists Foundation (J.B.H.). M.J.K. was supported in part by NIH Training grant T32 GM007175 and the Amgen Research Excellence in Bioengineering and Therapeutic Sciences Fellowship. R.P.S. is supported by a CIHR fellowship in the area of hepatology. Parts of the research were conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. The content is solely the responsibility of

the authors and does not necessarily represent the official views of the NIH, NICHD, NHGRI or the NIGMS.

Chapter 6 Rapid and sensitive multiplex sequencing of actionable genes in clinical cancer samples

This chapter is based on a recently submitted manuscript:

Joseph B Hiatt, Colin C Pritchard, Stephen J Salipante, Brian J O'Roak, Jay Shendure. Rapid and sensitive multiplex sequencing of actionable genes in clinical cancer samples.

Jay Shendure and I conceived and designed the study with input from Colin Pritchard and Stephen Salipante. Brian O'Roak and I designed the smMIPs and developed protocols. Colin Pritchard and Stephen Salipante obtained anonymized clinical samples, coordinated or oversaw single mutation genotyping, and aided with interpretation of results. I performed all other experiments and all data analysis. Jay Shendure and I wrote and revised the manuscript with input from all other authors.

6.1 Summary

Despite recent advances in DNA sequencing technology, we continue to lack practical methodsou for comprehensively detecting actionable cancer mutations in a clinical setting. Here we describe smMIP, an assay that combines single molecule tagging with multiplex targeted capture to enable rapid, cost-effective, accurate, and sensitive resequencing. We validated the method by simultaneously resequencing 33 clinically actionable cancer genes in each of 53 samples (including 45 obtained during routine patient care, of which 40 were formalin-fixed, paraffin-embedded). Single molecule tagging facilitated extremely accurate consensus calling, with an estimated per-base error rate of 8.4×10^{-6} in cell lines and 2.6×10^{-5} in clinical specimens. Altogether, we detected 134 putative somatic non-synonymous variants at frequency greater than ~10%. We replicated 25 of 27 (93%) positive results of single mutation tests from a clinical laboratory and identified 7 low-frequency mutations (0.2% to 4.7%) including *BRAF* p.V600E (melanoma, 0.2% alternate allele frequency), *KRAS* p.G12V (lung, 0.6%), *JAK2* p.V617F (melanoma, colon, two lung, 0.3% to 1.4%), and *NRAS* p.Q61R (colon, 4.7%). We also demonstrate compatibility with a workflow that goes from clinical specimen to analyzed result in less than 72 hours. We anticipate that smMIP will be broadly adoptable as a rapid, cost-effective method for accurately detecting actionable cancer mutations in both research and clinical settings.

6.2 Introduction

The advent of massively parallel technologies has dramatically decreased DNA sequencing costs, which has in turn transformed the study of genetic variation[17], gene expression[18] and its regulation[19,20], the genetic basis of rare[21] and common[22] disease, and other areas. Massively parallel DNA sequencing has also been applied with substantial success to elucidate the genetic underpinnings of a wide variety of cancers in a research setting[61-64]. However, despite the considerable potential for these same findings to directly inform patient care, the molecular characterization of tumors through massively parallel sequencing has yet to be widely adopted as a clinical test. Practical hurdles to the clinical use of a massively parallel sequencing based assay for cancer genome sequencing include cost, accuracy, sample input requirements, workflow simplicity, turnaround time, and data interpretation[140].

Because whole genome sequencing remains cost and time prohibitive from a clinical standpoint, many efforts in this area have instead focused on the multiplex targeted sequencing of cancer-related genes for which mutation status directly informs patient care[141,142]. Emerging methods for multiplex targeted sequencing of clinical cancer specimens include hybrid capture[141,142] and highly multiplexed PCR[143-145]. However, each of these methods has important drawbacks that limit practical utility. Hybrid capture typically entails a complex and time-intensive workflow, high per-sample reagent costs, and limited flexibility to reformulate the protocol as the regions of interest change over time. Highly multiplexed PCR often relies on complex instrumentation[143,145] and may be restricted to a limited number of target genes. Methods for multiplex targeted sequencing must also be robust to relatively small amounts and poor quality of source DNA such as that isolated from formalin-fixed, paraffin-embedded (FFPE) tissue.

A key advantage of multiplex targeted sequencing over whole genome sequencing is that it can provide the high sequence coverage depth necessary to detect clinically relevant mutations present at a low frequency because of non-neoplastic cell admixture or tumor heterogeneity[146-149]. However, the successful detection (or exclusion) of low-frequency mutations is rendered challenging by amplification and sequencing errors and by the variable quantity and damage associated with FFPE-derived genomic DNA. Although it is possible to substantially reduce effective error-rates by carefully modeling the underlying processes[144,150-152], such analytical methods do not actively correct errors. An alternative

approach, developed by us and others, is to perform "single molecule tagging" to mark sequence reads derived from a common progenitor molecule (that is, the same genomic equivalent in source DNA)[128,153-158], and to subsequently use this information to guide consensus calling on a molecule-by-molecule basis[128,155,156]. However, to our knowledge there are no reports integrating single molecule tagging with multiplex targeted sequencing.

Here we sought to develop a method for multiplex targeted sequencing of clinically relevant cancer genes that was: (a) accurate, sensitive, and quantitative for both high- and low-frequency somatic mutations; (b) robust to the variable quality of DNA isolated from clinical FFPE tissue; (c) sufficiently rapid, straightforward, and cost-effective for use in a clinical or high-throughput research setting. The resulting method, termed smMIP (for single molecule Molecular Inversion Probes), combines the MIP strategy for targeted capture[159,160] with single molecule tagging[128,153-158]. MIPs represent an attractive platform for targeted capture because of their very low per-sample cost, workflow simplicity, target set modularity, and low sample input requirements. Single molecule tagging, on the other hand, enables consensus calling for single genomic equivalents present in the input material, thereby facilitating both highly sensitive variant calling and precise quantitation of mutation frequency. The combination of MIPs and single molecule tagging form the basis for an ultra-sensitive, targeted sequencing assay that performs well with respect to attributes relevant for clinical implementation, *e.g.* speed, ease of use, and compatibility with FFPE-treated clinical tumor specimens.

As a proof-of-concept, we designed molecular capture/tagging probes (smMIPs) targeting the coding sequences of 33 cancer genes in which clinically actionable mutations may occur. We applied these probes to the targeted capture, sequencing, and mutational analysis of 53 specimens in parallel, comprising 45 clinical cancer specimens and 8 HapMap DNA mixtures. We demonstrate that smMIPs are capable of ultra-sensitive detection of nearly all types of mutation, while detecting both expected and new variation in these samples. We also describe and validate a rapid smMIP workflow that is capable of going from FFPE specimen to analyzed result in less than 72 hours.

6.3 Results

6.3.1 Multiplex targeted sequencing using smMIPs

We designed and procured a pool of 1,312 smMIP oligonucleotides targeting the coding sequences of 33 cancer-related genes[113] (Table B.1). These smMIPs tiled a total of ~125 kilobases (kb) of genomic sequence, including 80,384 of the 81,190 (99%) coding base pairs (bp) of the 33 targeted genes. Targeted capture with smMIPs involves a standard MIP protocol for "library-free" sequencing[159,160] with slight modifications (Figure 6.1). Following the post-capture PCR amplification, samples are subjected to massively parallel sequencing using the Illumina platform and analyzed using a custom pipeline. Our strategy involves two layers of indexing (Figure 6.1), with one index sequence (the "sample index") resolving capture products from distinct source DNAs and another (the "molecular tag") resolving reads derived from distinct genomic equivalents within individual source DNAs (Note C.1). During analysis, overlapping regions of read-pairs are reconciled to produce ~152 nt "raw reads" that, once molecular tagging information is incorporated, form the basis for highly accurate single molecule consensus reads ("smc-reads"). To our knowledge, this is the first description of molecular tagging integrated with MIPs, and, more generally, with a large-scale multiplex targeted capture strategy.

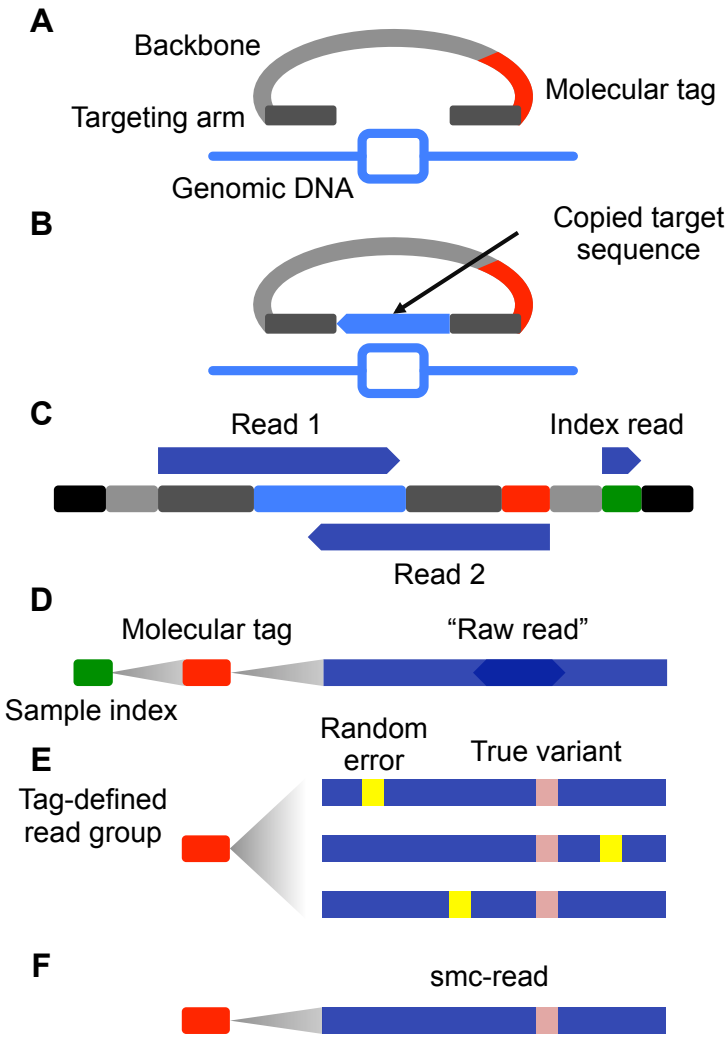


Figure 6.1. Schematic of smMIP method.

(a) Molecular inversion probes (MIPs) consisting of two 16-24 nucleotide (nt) "targeting arms" (dark grey) joined by a constant 28 nt "backbone" sequence (light grey) and a 12 nt degenerate "molecular tag" (red) were designed for the coding exons (light blue rectangle) of 33 cancer-related genes. Targeting arms were complementary to sequences flanking individual regions of interest, each 112 nt in length. (b) Probes are pooled, hybridized to genomic DNA, and polymerase and ligase were added to "gap-fill" the reverse complement of the genomic DNA to which the probe is hybridized (light blue arrow) and ligate the probe into a single-stranded circle. (c) After exonuclease treatment and PCR, sequencing library molecules consist of platform compatibility (black), probe backbone (light grey), targeting arm (dark grey), copied target (light blue), molecular tag (red), and sample-specific index introduced during PCR (green). Massively parallel sequencing is used to collect three reads (dark blue). (d) Overlapping read-pairs are reconciled to form "raw" reads (dark blue), assigned to samples via the sample-specific index sequence (green) and individual capture events via the molecular tag (red). (e) Groups of raw reads assigned to the same probe via alignment to the reference genome and sharing the same molecular tag and sample index form a "tag-defined read group" (TDRG). Random errors (yellow) that occur during library construction and sequencing may be present in some members of the TDRG at some positions. (f) TDRGs are used to call a single molecule consensus sequence (smc-read) for the captured target sequence that is robust to such errors.

To validate the smMIP method, we simultaneously applied it to two sets of samples. First, to assess sensitivity and positive predictive value and the extent to which we could precisely quantify low-frequency variants, we performed smMIP-based targeted sequencing on genomic DNA from two HapMap cell lines (NA12892 and NA19239) and six mixtures of these two gDNAs. Second, to explore the practical utility of the method, we also applied smMIP-based targeted sequencing to a panel of forty-seven genomic DNA isolates from clinical specimens encompassing a wide range of cancers (Table 6.1; Table C.2). All of the non-hematologic DNA isolates were obtained from FFPE-treated tissue (42 of 47 clinical specimens). Importantly, the FFPE specimens were not selected for quality in any way, and included material that had been isolated as long as eight years prior to our experiments and had been process☺ed into genomic DNA as long as seven years after FFPE treatment. These specimens thereby represent a stringent and realistic test of "real-world" method performance. We performed 55 capture reactions in parallel, using ~500 ng of genomic DNA per capture, and carried out sequencing and analysis as outlined above. Two clinical specimens failed to yield sufficient on-target sequence during quality control and were excluded from further analysis, resulting in a success rate of 96% (45 of 47).

Table 6.1. Summary of clinical samples.

Cancer type	Number of samples
Colorectal/rectal adenocarcinoma	18
Non-small cell lung cancer	11
Melanoma	7
Gastrointestinal stromal tumor	4
Myeloproliferative disorder*	3
Acute myeloid leukemia*	2
Urothelial carcinoma	1
Ovarian adenocarcinoma	1

Tissue samples obtained during routine clinical practice and processed by the University of Washington Department of Laboratory Medicine Clinical Molecular Genetics Laboratory or Hematopathology Laboratory. All DNA isolates with the exception of the five total myeloproliferative disorder (*i.e.* polycythemia vera) and acute myeloid leukemia samples were prepared from FFPE tissue.

6.3.2 Method performance

We first sought to assess performance of the smMIP assay with respect to sensitivity, positive predictive value, and uniformity of target enrichment. Because of the heterogeneous nature of the specimens, the number of raw reads obtained per sample varied from 1 to 16 million raw reads (Figure C.1). However,

77% of samples (41 of 53) were within a 3-fold range, and this distribution could likely be improved by automated pooling. Raw reads were aligned to the reference genome (Note C.2; Figure C.1) and processed using a custom analysis pipeline to yield smc-reads.

We first explored coverage of targeted regions, finding that mean smc-read coverage of the targeted coding bases was 3,538x across the HapMap samples and 1,051x across the clinical specimens (Figure 6.2; Note C.3). We then used smc-reads to call genotypes, and, for the HapMap samples, compared our calls to 1000 Genomes ("1KG") pilot project genotypes[161]. For NA12892, we detected 24 of 25 1KG variant sites; the remaining position was not adequately covered in our data. After discarding three positions that were systematically mis-genotyped by our assay, we detected two additional variant positions; these calls were supported by manual inspection of more recent 1KG data. For NA19239, we detected 41 of 44 1KG variant sites; the remaining three positions were not adequately covered in our data. Two additional sites had variant genotypes and were again supported by newer 1KG data. Therefore, based on this limited comparison, our assay is highly accurate at adequately covered positions. Considering all targeted sites, we estimate the sensitivity of our assay to be 93-96% and the positive predictive value to be near 100%.

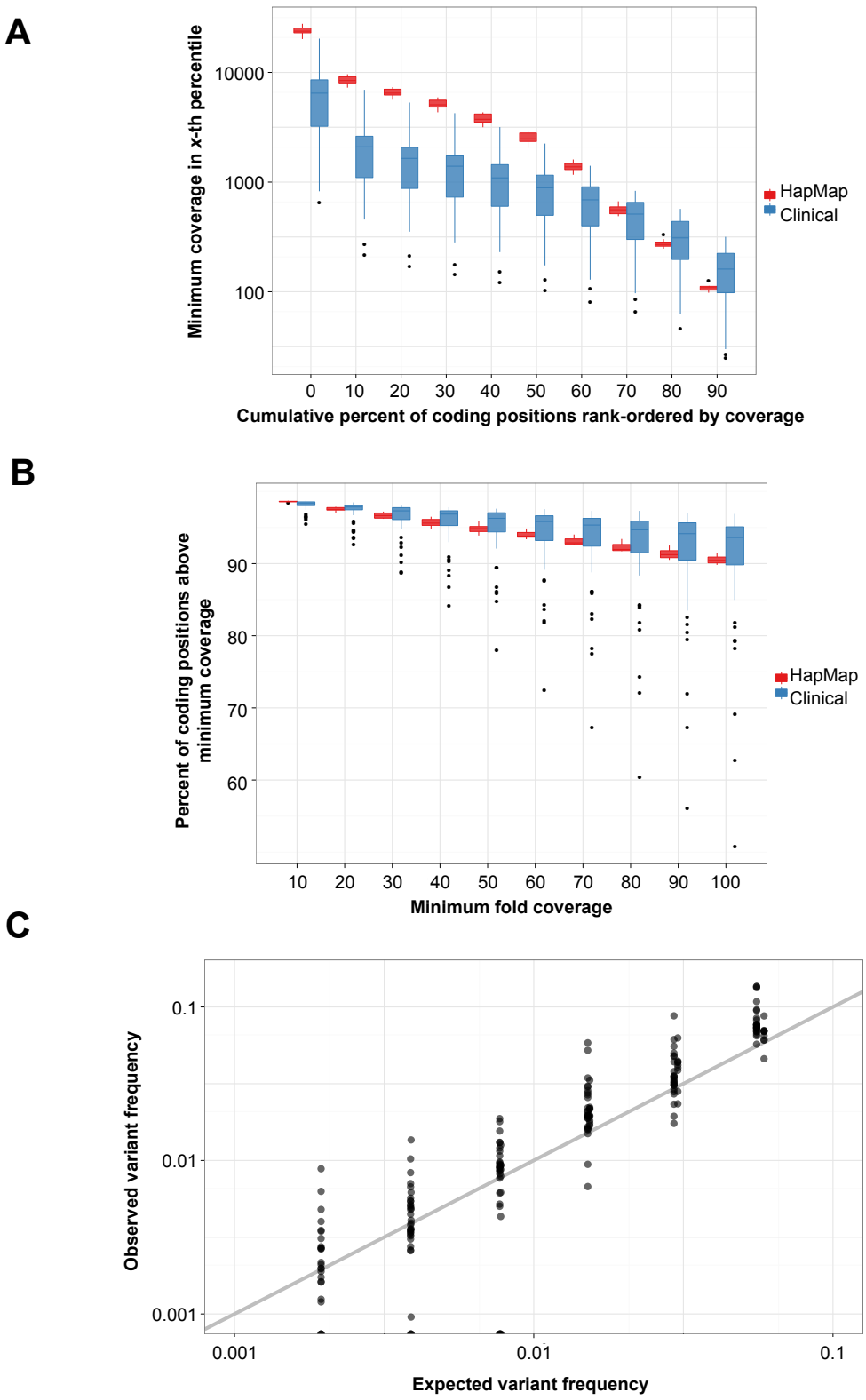


Figure 6.2. smMIP capture performance and detection of low-frequency variation.

(a) Distributions of fraction of coding positions above a given smc-read coverage cutoff across 8 HapMap cell line (red) and 45 clinical cancer (blue) samples (box plot center line: median; upper and lower edges: quartiles; whiskers: farthest data point within 150% of inter-quartile range; dots: outliers). (b) Distributions of minimum coverage in a given percentile of total targeted coding positions, rank-ordered by smc-read coverage. Zeroth-percentile indicates maximum coverage. (c) Observed versus expected variant frequency in smc-read base-calls from mixtures of HapMap genomic DNA samples at known ratios for positions with at least 100x coverage (R=0.94). Ideal performance is shown as grey line (y=x).

6.3.3 Sub-clonal variant detection

To assess whether the smMIP assay was capable of sensitively detecting and accurately quantifying variants present at sub-clonal frequencies, we applied it to six synthetic mixtures of genomic DNA from the two HapMap cell lines combined in a 2-fold serial dilution from 1:8 to 1:256 (resulting in low-abundance genome alternate allele frequencies of ~11% to ~0.2%). We then compared the expected variant frequency to that observed in smc-reads (Figure 6.2c). In general, we observed close agreement between the expected and observed frequency for positions with at least 100x smc-read coverage (R=0.94), with the deviation from expected frequency largely explained by sampling statistics (Figure C.2).

We next sought to quantify the absolute error rate of the smMIP assay and to assess potential sources of error (Note C.4; Figure C.3). In the samples derived from HapMap cell-lines, smc-read base-calls were ~13x more accurate than raw read base-calls, with a substitution rate of 8.4×10^{-6} per base compared to 1.1×10^{-4} per base for raw read calls (Table 6.2). In the clinical samples, the difference was ~5-fold (2.7×10^{-5} per base compared to 1.3×10^{-4} per base). We then explored substitution rates as a function of the expected nucleotide incorporated into the MIP during the gap-fill versus the observed nucleotide in the raw read or smc-read base-call (Figure 6.3a). We observed substantial variation with respect to different pairs of expected/observed nucleotides, as well as an effect of dinucleotide context. Interestingly, the dominant patterns were consistent with pro-mutagenic chemical processes in individual progenitor molecules, namely the oxidatively damaged base 8-oxo-guanine (Figure 6.3b; Figure C.4; Note C.5) and spontaneous deamination of cytosine and 5-methyl-cytosine (Figure 6.3c; Figure C.4; Table C.3; Note C.5).

Table 6.2. Substitution error rates.

		Raw reads		smc-reads		Fold-reduction in sub. rate
		Calls	Sub. rate	Calls	Sub. rate	
HapMap cell lines	All	$1.0x10^{10}$	$1.1x10^{-4}$	$4.6x10^{9}$	$8.4x10^{-6}$	12.8
	No G>A, C>A	$8.8x10^{9}$	$1.0x10^{-4}$	$3.9x10^{9}$	$3.5x10^{-6}$	28.8
Clinical samples	All	$2.2x10^{10}$	$1.3x10^{-4}$	$7.8x10^{9}$	$2.7x10^{-5}$	4.7
	No G>A, C>A	$1.8\ x10^{10}$	$1.0x10^{-4}$	$6.6x10^{9}$	$5.1x10^{-6}$	20.5

Total number of calls, substitution rates, and the fold-reduction in substitution rate comparing smc-reads to raw reads for very high-confidence (≥Q41) base-calls from raw reads (i.e. read-pairs that have been aligned against one another and collapsed into a consensus sequence) and smc-reads (Q60). These data are also shown excluding the G>A and C>A substitutions that are likely caused at least in part by patterns of DNA damage (deamination of C and 5mC and oxidative damage to G resulting in 8-oxo-G). Only positions that were genotyped to sufficient depth that the constitutional genotype of the sample could be confidently determined to be homozygous reference were used to calculate these rates.

When we removed the contribution of these potential sources of false-positive substitution calls, smc-read base-calls were even more accurate, with substitution rates of $3.5x10^{-6}$ per base and $5.1x10^{-6}$ per base for the HapMap and clinical samples, respectively, while the substitution rates of the raw read calls did not decrease substantially (Table 6.2). These patterns are consistent with gap-fill mis-incorporations due to DNA damage and actual sub-clonal heterogeneity constituting the major source of substitutions in smc-read base-calls, and with polymerase errors after the initial gap-fill event constituting a major source of substitutions in raw reads. Smc-reads are therefore at least 5-fold and as much as 30-fold more accurate than the most confident base-calls in the raw reads, with substitution rates as low as $3.5x10^{-6}$ per base when ignoring only two of twelve possible substitutions or $8.4x10^{-6}$ per base when considering all possible single nucleotide substitutions.

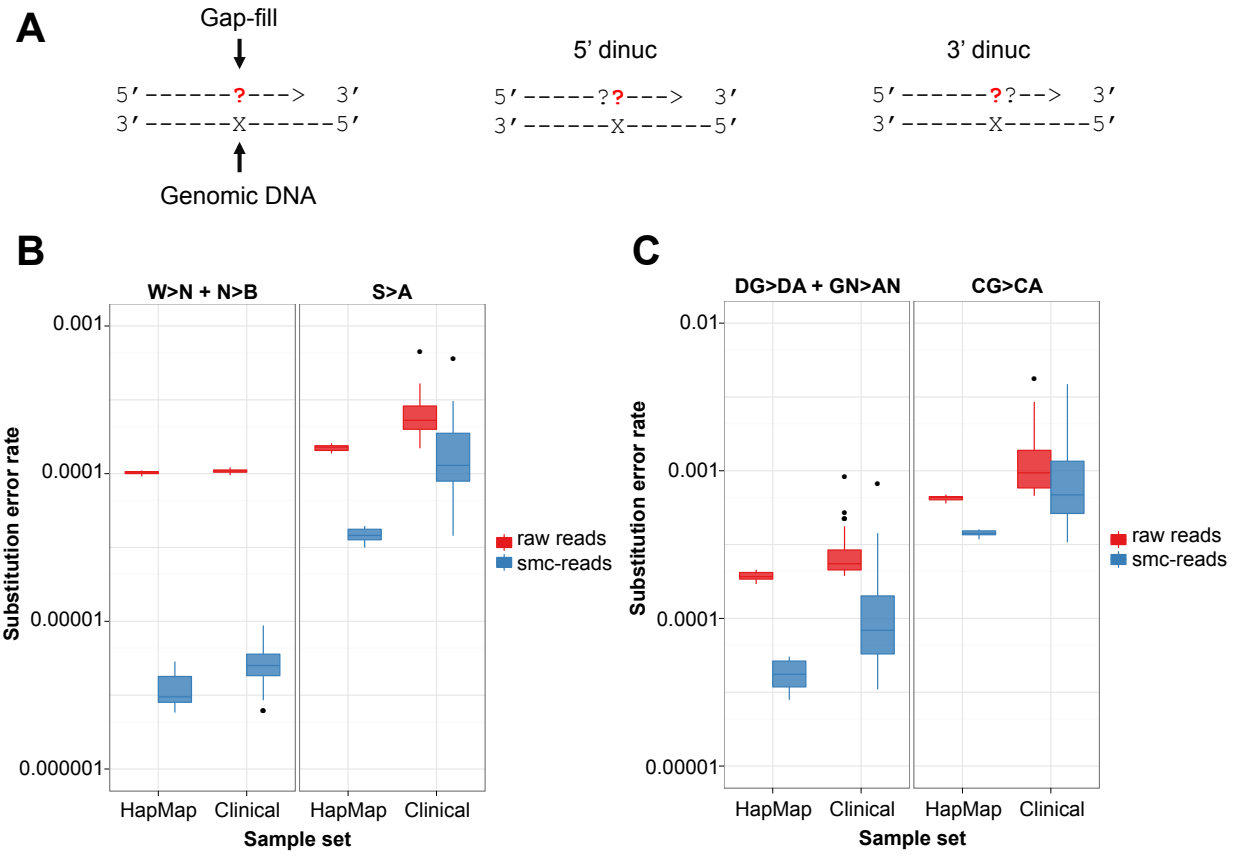


Figure 6.3. Substitution error rates as a function of expected and observed nucleotide during gap-fill.

(a) Schematic illustrating mononucleotide and dinucleotide substitution dependencies being considered. All rates are shown for a given expected gap-fill mono- or dinucleotide, that is the complementary nucleotide(s) to the nucleotide(s) present in the target genomic DNA, considering only ≥Q41 raw read base-calls and Q60 smc-read base-calls at putative homozygous positions based on GATK calls. (b) Distributions of substitution error rates for 8 HapMap cell line and 45 clinical cancer samples, comparing raw reads and smc-reads, and all substitutions other than C>A or G>A (W>N + N>B, left panel) to only C>A and G>A (S>A, right panel). (c) Distributions of substitution error rates comparing raw and smc-reads, and all G>A substitutions occurring in the non-CG dinucleotide context (DG>DA + GN>AN, left panel) to G>A substitutions occurring only in the CG dinucleotide context (CG>CA, right panel).

6.3.4 Detection of somatic variation

To assess whether the smMIP method can potentially replace existing single-gene clinical tests for actionable mutations, we performed a blinded comparison of smMIP results to the results of clinical single gene tests. In particular, a subset of our samples had been previously genotyped for individual actionable substitution and indel mutations in *BRAF*, *EGFR*, *FLT3*, *JAK2*, *KIT*, *KRAS*, *NRAS* and *PDGFRA*. Considering these sites, we detected 25 of 27 (93%) previously identified mutations (Table 6.3; Table C.4; Table C.5). We missed two large (67 and 104 bp) insertions in *FLT3*, although these could in

principle be detected using a more sensitive analysis strategy and/or more densely tiled probes in this region. We further detected two mutations in these sites in two lung cancer samples that had not been previously genotyped at that site (*KRAS* p.G12C in sample 8 and p.G12V in sample 37); these calls were subsequently confirmed using a melt curve-based assay (data not shown).

Table 6.3. Concordance with single mutation tests.

Gene	Mutation	Expected # of events	% detected
BRAF	p.V600E	4	100
BRAF	p.V600K	2	100
EGFR	p.L858R	2	100
EGFR	15 bp deletion (exon 19)	1	100
EGFR	18 bp deletion (exon 19)	1	100
FLT3	67 bp insertion	1	0
FLT3	104 bp insertion	1	0
JAK2	p.V617F	3	100
KIT	6 bp insertion (exon 11)	1	100
KIT	15 bp deletion (exon 11)	1	100
KRAS	p.G12C	2	100*
KRAS	p.G12V	1	100*
KRAS	p.G12D	2	100
KRAS	p.G13D	2	100
NRAS	p.Q61R	1	100
PDGFRA	p.D842V	2	100
Total	All	27	92.6%

smMIP genotypes from clinical samples were compared to single mutation tests previously performed by the University of Washington Department of Laboratory Medicine Clinical Molecular Genetics Laboratory or Hematopathology Laboratory. The smMIP assay detected 25 of 27 expected mutations; two large insertions in *FLT3* were not observed, although the assay is, in principle, capable of detecting these mutations. *smMIP also detected two additional *KRAS* mutations (one p.G12C and one p.G12V) in two lung cancer samples that had not been genotyped for these mutations; these mutations were subsequently confirmed in these samples by the clinical laboratory.

To explore other somatic mutations, and because we did not have access to matched normal tissue, we filtered variant sites identified in the clinical cancer specimens against germline variant sites observed in ~5,400 exomes by the Exome Sequencing Project[162]. We then required at least 30x coverage to remove poorly ascertained sites. Across the 45 clinical samples, this filtering process yielded 134 putative somatic events, of which 74 were found in the Catalogue of Somatic Mutations in Cancer[163] (Figure 6.4a). As expected, several genes were recurrently mutated in specific tumor types (Figure 6.4b), such as 11 of 16 colon cancer samples harboring at least one APC mutation.

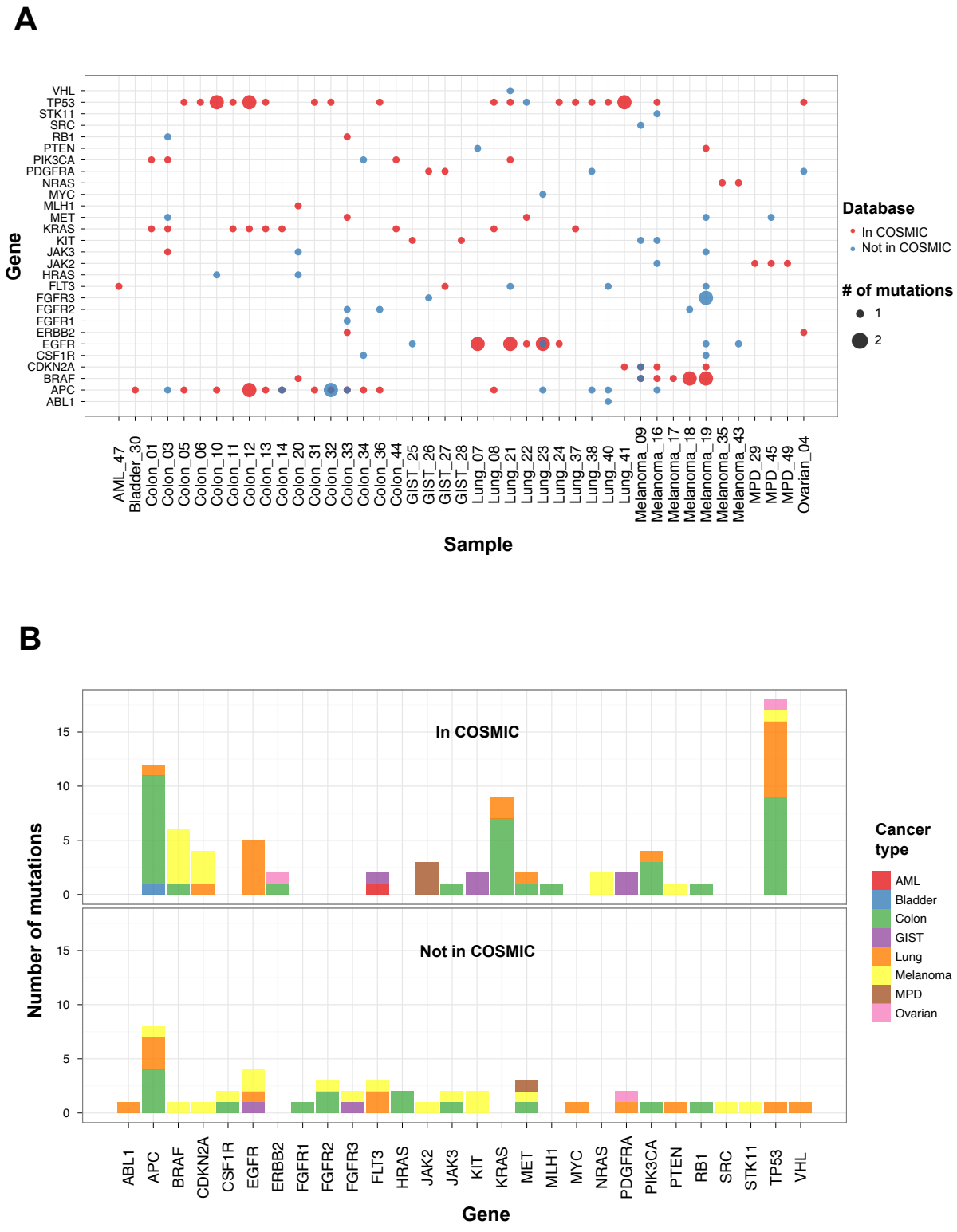


Figure 6.4. Somatic mutations in clinical samples.

(a) Number of coding, non-synonymous point or indel mutations in targeted genes detected in clinical samples (numerical sample ID listed after tissue type) at intra-sample allele frequency ≥~10%, for variants that are candidate somatic events (*i.e.* site not variable in the exomes of 5400 individuals), and as a function of presence in the COSMIC database (red: in COSMIC; blue: not in COSMIC). (b) Number of samples with at least one mutation in a given gene as a function of COSMIC database status and cancer type (AML: acute myeloid leukemia; GIST: Gastrointestinal stromal tumor; MPD: myeloproliferative disorder).

To investigate the possibility of tumor sub-clones in these specimens, we examined the extent to which putative somatic mutations were observed at similar frequencies within individual samples (Figure 6.5a). While some clustering of alternate allele frequencies is apparent, we observed substantial variation of alternate allele frequencies within individual samples, which may reflect the presence of multiple, genetically distinct sub-clones. Copy number gain is another potential source of alternate allele frequencies substantially different from 0.5 or 1. To explore the possible contribution of copy number change to allele frequency variation, we also examined alternate allele frequencies for putatively germline variant sites for all clinical samples and the two pure HapMap cell line samples (Figure 6.5b). Alternate allele frequencies for germline variants appeared substantially more variable in clinical samples compared to the cell line samples, which is consistent with a contribution of copy number gain to the observed variation in allele frequencies.

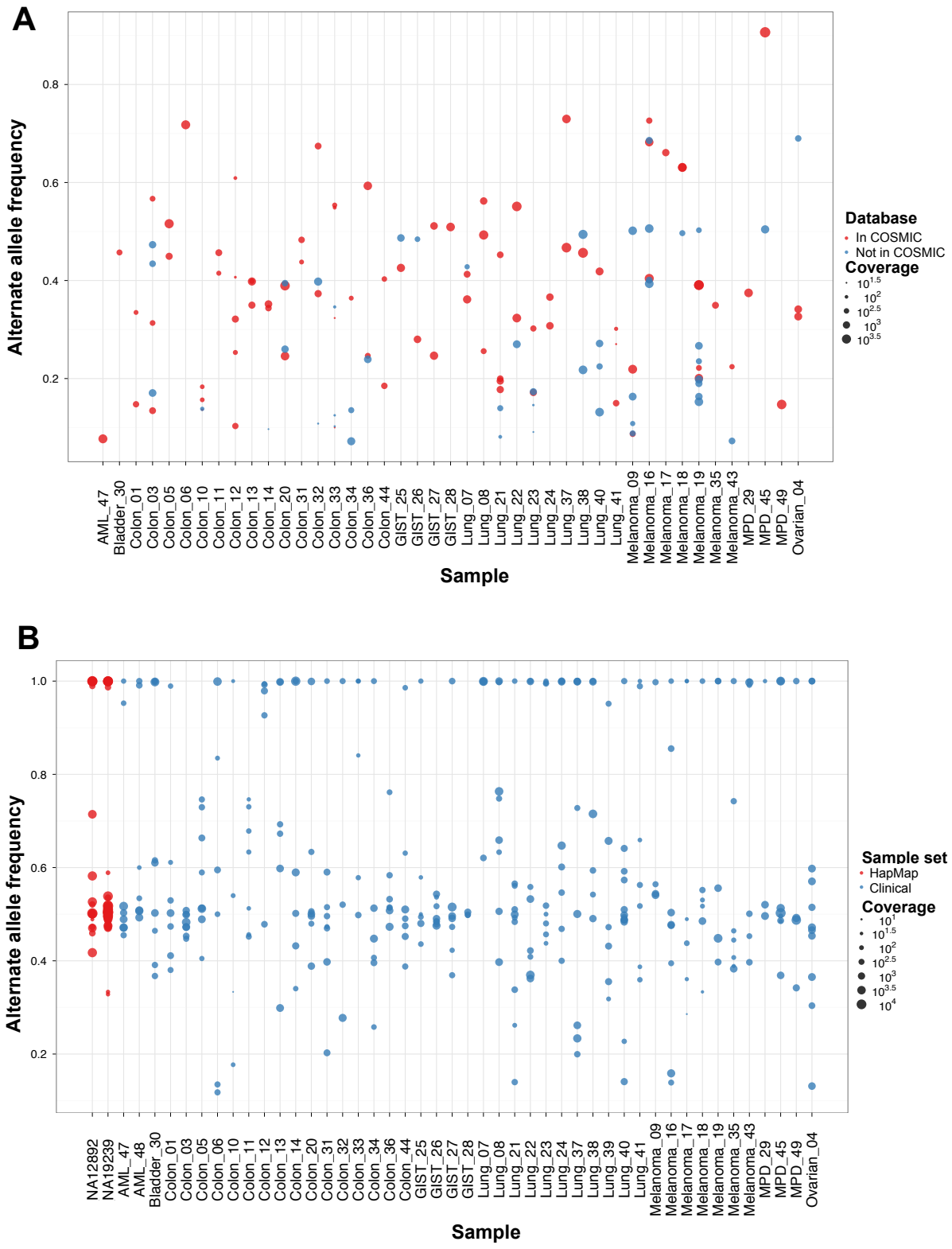


Figure 6.5. Alternate allele frequencies of somatic and germline variants in clinical samples.

(a) Alternate allele frequency for candidate somatic mutations (called by GATK) in clinical samples (numerical sample ID listed after tissue type) as a function of COSMIC database status and with point size scaled to coverage of that site in that sample. (b) Alternate allele frequency for candidate germline variants (based on site having been called as variant in at least one of 5400 exomes) in two unmixed HapMap (red) and 45 clinical cancer (blue) samples with point size scaled to coverage.

6.3.5 Sub-clonal somatic variation at clinically relevant sites

We next explored the potential for the smMIP assay to detect very low-frequency substitutions at clinically relevant sites. Restricting to substitutions found in at least two maximum quality smc-read base-calls, we found seven low-frequency variants at such sites (Table 6.4). These consisted of: low-frequency *JAK2* p.V617F mutations (n=4) in two lung cancers, a melanoma, and a colon cancer; a *BRAF* p.V600E mutation in a melanoma; a *KRAS* p.G12V mutation in one of the lung tumors that also harbored a low-frequency *JAK2* p.V617F mutation; and an *NRAS* p.Q61R mutation in a colon cancer. Considering the substitution rates for the observed mutations in these samples in smc-reads, the least unlikely of these events is expected to occur by chance according to binomial statistics with frequency ~$6x10^{-12}$. To exclude the possibility of artifactual low-frequency variant detection due to sample cross-contamination or index cross-talk, we subjected independent DNA aliquots from the four specimens with low-frequency *JAK2* mutations to confirmatory clinical testing using an allele-specific PCR assay; all four mutations were confirmed. Furthermore, *JAK2* mutations have been reported at low frequency in a previous study in non-small cell lung cancer samples[142]. Additionally, index sequence cross-talk, which would be predicted to give rise to mixed read groups, is not a likely explanation for these low-frequency calls; we required very high quality smc-read base-calls and mixed read groups are not a general phenomenon in our data (Figure C.6). We note that, prior to our study, the melanoma sample was genotyped clinically for *BRAF* p.V600 and this mutation was not detected (as expected given that assay's limited sensitivity); additionally, polyclonality in melanomas with respect to *BRAF* p.V600 mutation status has been previously observed[164-166].

Table 6.4. Low-frequency variation at clinically relevant sites in tumor samples.

Sample	Cancer type	Gene	Chr.	Pos.	Ref. allele	Alt. allele	Sub. in gap-fill	Ref. allele counts	Alt. allele counts	Alt. allele fraction	Mutation in protein
43	Melanoma	*BRAF*	7	140453136	A	T	A>T, T>A	1266	3	0.0024	p.V600E
19	Melanoma	*JAK2*	9	5073770	G	T	G>T, C>A	1465	19	0.0128	p.V617F
34	Colon	*JAK2*	9	5073770	G	T	G>T, C>A	1058	15	0.0140	p.V617F
38	Lung	*JAK2*	9	5073770	G	T	G>T	1229	4	0.0032	p.V617F
41	Lung	*JAK2*	9	5073770	G	T	G>T, C>A	795	6	0.0075	p.V617F
41	Lung	*KRAS*	12	25398284	C	A	G>T	464	3	0.0064	p.G12V
12	Colon	*NRAS*	1	115256529	T	C	A>G	184	9	0.0466	p.Q61R

Numerical sample ID, cancer type, gene name, chromosome, position (hg19), reference allele, alternate allele, substitution occurring during gap-fill process (for comparison with stranded error rate calculations), number of high-quality smMIP (Q60) reference allele calls, number of high-quality smMIP (Q60) alternate allele calls, relative fraction of alternate allele calls, and amino acid substitution for substitution variants detected at low frequencies in tumor samples at clinically relevant sites (*BRAF* p.V600, *EGFR* p.L858, *JAK2* p.V617, *KRAS* p.G12/p.G13, *NRAS* p.Q61, and *PDGFRA* p.D842). We required at least two observations of the alternate allele at smc-read base-call quality Q60. For three of the *JAK2* observations (samples 19, 34, and 41) and the *BRAF* observation, the mutation was observed in independent MIPs targeting both strands.

6.3.6 Rapid workflow characterization

Next, we sought to develop a rapid smMIP workflow using the Illumina MiSeq platform to enable return of results on a clinically useful timescale (Table 6.5). In addition to using the more rapid sequencing instrument, we further streamlined the experimental protocol. To assess performance of this revised workflow, we applied it to eight of the clinical samples that we had already characterized. Five of these samples harbored low-frequency variation at clinically relevant sites as described above; the other three were selected at random.

We subjected the pool of eight clinical samples to a single MiSeq run (152 nt paired-end reads + 8 nt index read). In the absence of a gel-based size-selection, we observed decreased mapping rates compared to the HiSeq data (Figure C.7). Nevertheless, we successfully called genotypes at 85-97% of targeted coding bases (requiring ≥10x smc-read coverage), and covered 92-99% of successfully genotyped sites from high-coverage (*i.e.* HiSeq) data (requiring ≥30x smc-read coverage) (Table 6.6). We observed 100% agreement between rapid workflow (low-coverage) and slow workflow (high-coverage)

genotype calls (605,786 calls across the eight samples). These calls included all five previously ascertained clinically relevant mutations. However, coverage was insufficient to detect any of the newly discovered low-frequency mutations. Such events may be rendered detectable with the rapid workflow by improvements including the reduction of capture-related artifacts and the increasing throughput of bench-top next-generation sequencing platforms. Finally, we explored the extent to which smc-read coverage and base-call substitution rates were affected by reduced sampling of the high-coverage data, finding that coverage and extremely high quality smc-read base-calls were maintained to only 1% (*i.e.* ~$1.5x10^7$ read-pairs) of the total high-coverage data (Figure C.8; Table C.6).

Table 6.5. Rapid workflow timetable.

Step nbr.	Step description	Time (min)	Time (hrs)	Total time (mins)	Total time (hrs)	Start day	End day	Start time	End time
1	Isolate genomic DNA	240	4.00	240	4.00	1	1	9:00 AM	1:00 PM
2	Hybridization, Gap-fill, Ligation*	300	5.00	540	9.00	1	1	1:00 PM	6:00 PM
3	Wait overnight	900	15.00	1440	24.00	1	2	6:00 PM	9:00 AM
4	Exonuclease	65	1.08	1505	25.08	2	2	9:00 AM	10:05 AM
5	PCR	120	2.00	1625	27.08	2	2	10:05 AM	12:05 PM
6	SPRI purification (1.8x)	20	0.33	1645	27.42	2	2	12:05 PM	12:25 PM
7	SPRI purification (0.8x)	20	0.33	1665	27.75	2	2	12:25 PM	12:45 PM
8	Gel	60	1.00	1725	28.75	2	2	12:45 PM	1:45 PM
9	Pooling and quantification	20	0.33	1745	29.08	2	2	1:45 PM	2:05 PM
10	MiSeq	1680	28.00	3425	57.08	2	3	2:05 PM	6:05 PM
11	Analysis**	360	6.00	3785	63.08	3	3	6:05 PM	12:05 AM

Estimated practical timetable (assuming ~8 hour workday) for rapid workflow, including step number, description, time for each step, cumulative time, day of start and finish, and wall clock time of start and finish for each step. *No sample manipulation is required at the end of step 2. **As analysis process is automated, we assumed it could be performed following the end of the MiSeq run (but not during a routine work day).

Table 6.6. Performance of rapid library construction and sequencing workflow.

Sample number	Cancer type	Fraction of targeted bases covered	Fraction of high-coverage sites covered	Concordance with high-coverage data	Clinically relevant mutation	Detected clinically relevant mutation?	Ref. allele / alt. allele counts
13	Colon	0.96	0.98	100% (77378/77378)	*KRAS* p.G12V	Yes	92/56
18	Melanoma	0.85	0.92	100% (68610/68610)	*BRAF* p.V600K	Yes	119/229
19	Melanoma	0.97	0.99	100% (77833/77833)	*BRAF* p.V600K	Yes	412/256
34	Colon	0.96	0.99	100% (77055/77055)	None	NA	NA
37	Lung	0.97	0.99	100% (77645/77645)	*KRAS* p.G12V	Yes	156/418
38	Lung	0.97	0.99	100% (78015/78015)	None	NA	NA
41	Lung	0.91	0.95	100% (73427/73427)	None	NA	NA
43	Melanoma	0.94	0.97	100% (75823/75823)	*NRAS* p.Q61R	Yes	27/6

Numerical sample ID, cancer type, fraction of targeted bases that were successfully genotyped in MiSeq data (requiring at least 10x coverage), fraction of successfully genotyped sites in HiSeq data (requiring coverage at least 30x) that were successfully genotyped in MiSeq data, percent concordance with high-coverage data and number of sites, clinically relevant mutation, whether clinically relevant mutation was accurately genotyped, and allele counts of clinically relevant mutation for eight clinical samples subjected to rapid benchtop sequencing workflow.

6.4 Discussion

Here, we combined technologies for single molecule tagging and molecular inversion probes towards the development of a flexible, cost-effective, rapid, and sensitive method for targeted sequencing of 33 clinically actionable cancer genes. Smc-reads represent the consensus of reads derived from the same progenitor molecule in genomic DNA, and the molecular tagging inherent to the smMIP assay facilitates error-correction down to a substitution rate of 8.4×10^{-6} per base. Furthermore, the smMIP assay is highly quantitative for alternate allele frequencies as low ~0.2%. When we applied this method to a diverse panel of genomic DNAs including 45 clinical cancer specimens (40 of which were FFPE), we observed strong concordance with expected mutations based on clinical single gene tests, and discovered as many new mutations at clinically relevant sites as we missed. Overall, we detected 134 putatively somatic coding mutations across the 45 clinical samples, of which 48 were not found in the COSMIC database, and also identified seven low-frequency variants at clinically relevant sites. Finally, we established and

validated a simple and rapid smMIP workflow that is capable of going from DNA sample to analyzed result in less than 72 hours.

The smMIP assay has important advantages over alternative approaches, including hybrid capture[141,142] and highly multiplexed PCR[143-145]. Compared to hybrid capture, smMIP offers very low per-sample reagent costs and a substantially simpler and more rapid workflow. Furthermore, the capture reagent is modular, meaning new probes can be added "on-the-fly" as clinical practice evolves. Alternatively, the reagent could be split into single gene pools and combined as desired for small batches of samples to most efficiently leverage the rapid turnaround but decreased throughput of the bench-top sequencing platforms. Finally, molecular tagging facilitates single molecule consensus base-calling without relying on pseudo-random fragmentation breakpoints, which may not be informative at the high sequencing depths desired in clinical cancer sequencing. However, a smMIP-based approach will not likely scale as well to very large targets (*i.e.* thousands of genes), and may be less sensitive to large-scale genomic rearrangements. Compared to highly multiplexed droplet[143,144] or microfluidic[145] PCR, smMIP does not rely on sophisticated instrumentation, and, because of molecular tagging, is more sensitive and quantitative for low-frequency variation. However, smMIP may not be as compatible with very low sample inputs (i.e. less than ~10 ng) because the initial enrichment step is non-exponential.

There are a number of ways in which the smMIP assay can likely be improved. Coverage of poorly captured sites and capture uniformity from probe to probe will be improved by further probe rebalancing and supplementation of the probe set to more densely tile problematic regions. The detection of more complex variants such as the large *FLT3* insertions could also likely be achieved via denser tiling, or by adopting more sensitive analytical strategies using existing reagents. Further optimization of the capture protocol directed towards reducing formation of undesired low molecular weight artifacts should improve mapping rates, increasing sensitivity and reducing or eliminating the need for time- and labor-intensive size-selection steps during sample preparation. These improvements would be especially valuable in the context of the <72 hr rapid workflow. Improved algorithms and/or probe content may also facilitate the detection of loss-of-heterozygosity and copy number changes. Finally, the assay could be expanded to target additional genes for which clinical testing is becoming routine. Notably, the modular nature of the

smMIP capture reagent facilitates changes to the target definition on a rapid time-scale without the need to replace existing reagents.

In one colon cancer sample, we identified a *KRAS* p.G13D mutation and a low-frequency *NRAS* p.Q61R mutation. *NRAS* mutations are infrequent (~5% overall; ~3% p.Q61; ~1% p.Q61R) in *KRAS* wild-type colon cancers[167] but have similar implications, *i.e.* reduced response to the targeted anti-EGFR monoclonal antibodies cetuximab and panitumumab[168], and *KRAS* mutations have been detected in ~20% of cancers also harboring *NRAS* mutations[168]. Furthermore, colon cancers harboring *KRAS* p.G13D, unlike those harboring other *KRAS* mutations, may remain responsive to cetuximab/panitumumab[169], although this finding was not replicated in a subsequent study[168]. Based on our observation, one explanation for the disagreement between those studies is that some subset of tumors harboring *KRAS* p.G13D also harbor *NRAS* mutations at clonal or sub-clonal frequencies, and that *NRAS* mutation status is also influencing response to antibody therapy. Further study will be needed to better establish the prevalence of co-occurring *KRAS* p.G13D and clonal or sub-clonal *NRAS* mutations and the relationship between mutational status and response to therapy. Nevertheless, this observation highlights the utility of a single multiplex assay that is capable of detecting low-frequency variation in a large panel of cancer-related genes.

The ability to comprehensively, sensitively, and rapidly detect actionable mutations in both research and clinical settings has the potential to substantially transform cancer diagnosis and therapy. Individual cancers are shaped by the concomitant processes of mutation and clonal expansion, with extensive genetic heterogeneity emerging as a common attribute of human cancers[146-149]. Furthermore, the inevitable emergence of resistance to certain targeted therapies may be caused, at least in some cases, by the expansion of pre-existing low frequency sub-clones harboring resistance mutations[111,112]. In the long-term, the genomic characterization of individual cancers, whether for research or for clinical care, must include the reliable detection and quantitation of sub-clonal mutations. Such investigations will likely also benefit from the ability to characterize multiple independent samplings of neoplastic tissue from a given patient collected over time and space (*e.g.* independently sequencing multiple biopsies from the same tissue mass or multiple independent metastases). The speed, simplicity, parallelizability, and very

low substitution error rate (≤$3x10^{-5}$) of the smMIP assay are well suited to these challenges. These attributes may also render smMIP useful in other diagnostic scenarios where the sensitive and precise quantification of low-frequency variation is relevant, *e.g.* the detection of extremely low-frequency cancer-related mutations in circulating cell or cell-free DNA[145,170]; or as a complementary approach for non-invasively assaying fetal DNA from maternal plasma[171,172] at clinically relevant sites.

6.5 Notes

6.5.1 Acknowledgments

We thank C. Lee for assistance with sequencing, and members of the Shendure Lab for helpful discussions. We thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). We thank C. Smith, K. Koehler, M. Murillo, and H-S Yi for help genotyping clinical samples. Our work was supported by: a gift from the Washington Research Foundation; grant CA160080 from the National Cancer Institute (J.S.); grant AG039173 from the National Institute on Aging (J.B.H.); a fellowship from Achievement Rewards for College Scientists (J.B.H.); and the Department of Laboratory Medicine, University of Washington Medical Center.

Chapter 7 Overcoming read length and error rate limitations of massively parallel sequencing with molecular tagging: approaches and opportunities

This chapter is based on a manuscript being prepared for submission as a Perspective or Review article.

7.1 Introduction

The cost and throughput advances of massively parallel sequencing (MPS) platforms are offset by substantial drawbacks with respect to read length and base-calling accuracy[16]. For some applications, these limitations can be circumvented or ignored, but for many exciting potential uses of MPS, long and/or accurate reads are indispensable. These areas include *de novo* genome assembly[24,25], shotgun and amplicon metagenomics[26], structural variation in larger genomes[27], pathogen diversity[98], the detection of sub-clonal variation in targeted regions of larger genomes, and accurate ascertainment and phasing of variation in synthetic nucleic acid populations (for a more complete discussion of potential applications, see Chapter 3). To overcome these limitations, I and others have developed methods to enable unambiguous assignment of reads to individual progenitor molecules (*i.e.* "molecular tagging"), which in turn enables absolute quantification, error correction, and extension of effective read length. Here I review the emerging technological paradigm of molecular tagging. I first describe experimental strategies in general, followed by a detailed consideration of specific applications and the advantages compared to conventional MPS of each.

7.2 Molecular tagging strategies

All molecular tagging strategies share the same fundamental goal: to enable the assignment of sequencing reads to individual molecules present in the source nucleic acid from which a sequencing library is constructed. To achieve this goal, information that uniquely identifies a progenitor molecule must be encoded into the library in such a way that it can be subsequently decoded on the sequencing instrument. In practice, this has been achieved in one of two ways.

The first such method, introduced by Hiatt *et al*[128] and categorized as "endogenous" tagging by Kinde *et al*[156], relies on the fact that many nucleic acid fragmentation methods result in near-random breakpoints

such that any two molecules sampled from the same genomic locus are unlikely to have identical endpoints (how unlikely depends on the fragment length distribution, with molecules from more uniformly fragmented libraries more likely to serendipitously share endpoints). As the endpoints of the fragment are naturally accessed using most MPS platforms, endogenous tagging is a straightforward way to group reads or read-pairs based on shared molecular origin. Indeed, the earliest implementations of molecular tagging[128,156] (discussed in more detail below), employed endogenous tagging either entirely or at least in part. However, endogenous tagging has two important drawbacks: (a) as sampling depth of a particular locus increases, breakpoints will begin to "collide", resulting in inappropriate over-grouping of reads; (b) a random fragmentation step is not practical for many applications where long, highly accurate reads would be very useful (*e.g.* those involving amplicon sequencing). It should also be noted that many analytical workflows used in shotgun MPS have some awareness of progenitor molecular identity based on breakpoints; however, with at least one notable exception[53], these algorithms have typically discarded all but one read or read-pair corresponding to an inferred progenitor molecule (often termed "PCR duplicates") rather than performing any comparison of reads derived from the same progenitor. This strategy essentially achieves the digital quantitative precision[46-48] of molecular tagging without the same level of error correction.

The second and more powerful method of encoding uniquely identifying information, termed "exogenous" tagging, is the introduction of a highly complex sequence tag into the sequencing library early in the library construction process[20,153,154,156-158,173,174]. The sequence of this tag is then collected using the MPS instrument along with the sequence of the target nucleic acid, and is used to assign sequencing reads to progenitor molecules. Depending on the particular implementation, the molecular tag may be drawn from a complex but pre-determined[154,158] or degenerate[20,153,155-157] pool of sequences, although it should be noted that the use of pre-determined sequences necessarily limits sampling depth compared to the theoretical complexity attainable via even modestly degenerate sequences. Various application-dependent methods to incorporate the tag have been demonstrated, including ligation[20,153,154,173], priming or template-switching during cDNA preparation from mRNA[155,157,158], as part of a PCR primer that participates in an early (ideally initial) round of amplification[156], or as part of a Molecular Inversion

Probe[174]. Exogenous tagging enables discrimination of reads derived from two identical copies of the same nucleic acid sequence, and is thus the most generally applicable molecular tagging strategy.

7.3 De novo assembly of genomes and metagenomes from short reads

In 2010, we described a method called "Subassembly"[128], which was to our knowledge the first demonstration of molecular tagging in the context of MPS, and sought to improve the utility of MPS for *de novo* genome and metagenome assembly by increasing the effective MPS read length and decreasing the error rate (Chapter 4). We note that *de novo* assembly is a complex task affected by numerous factors[175]; however, holding all other factors constant, longer reads and higher accuracy base-calls are expected to generate more contiguous and accurate assemblies.

Subassembly works by creating nested sub-libraries from relatively long (~500 nt) DNA fragments so that short reads in sub-libraries can be appropriately grouped according to fragment of origin; these Tag-Defined Read Groups (TDRGs) are then individually assembled to accurately reconstruct the sequence of the longer fragments from which they were derived. We applied Subassembly to bacterial genomic DNA and an environmentally derived mixture of microorganism DNA (*i.e.* a metagenomic isolate). Beginning with 76 nt reads from the Illumina GAIIx platform, which had a substitution error rate of more than one in 100, we generated hundreds of thousands to millions of "subassembled reads" with median lengths ranging from 250 to 400 nt and less than one substitution error in 10,000 for the best 85% of base-calls. While this is indeed a selected subset of calls, we note that the quality estimation and subsequent data filtering is a routine component of most analysis strategies. We then explored the extent to which subassembled reads aided *de novo* genome assembly, finding that a combination of subassembled reads and a conventional paired-end short read library resulted in a more contiguous and more accurate assembly for both the *P. aeruginosa* and metagenomic samples compared to conventional paired-end data alone.

Subassembly as demonstrated in 2010 relied on endogenous tagging; however, as we noted then, the method is compatible with exogenous tagging and subsequent implementations of Subassembly indeed rely on that strategy as discussed below[20]. The method is not without its limitations, including the added complexity of library construction, and subassembled read length limits of ~1 kb corresponding to the

difficulty of clustering long molecules on the Illumina flow-cell. Additionally, since we first described Subassembly, existing "long read" MPS platforms have improved (*i.e.* the Roche/454 platform) and new platforms have emerged[15,43]. Indeed, *de novo* assembly methods using a hybrid of short and very long MPS reads have been recently established[24,25]; because of their simplicity, these methods are likely to become the standard *de novo* assembly approach for the near future. However, the more general conceptual framework introduced with Subassembly, *i.e.* a completely multiplex, single-tube, *in vitro* implementation of the hierarchical assembly strategy that underlies the quality and contiguity of the human reference sequence, may facilitate future breakthroughs in *de novo* assembly as well as haplotype phasing (discussed in more detail in the conclusion). Furthermore, currently available longer read MPS technologies are still beset by high per-base costs, high to extremely high error rates, and relatively limited availability compared to the most cost-effective platforms. These issues limit the application of the longer read platforms to many important applications, such as phasing variation in highly complex populations of synthetic constructs (discussed below).

7.4 Increasing quantitative precision of RNA-seq and genomic copy number estimation

MPS has also been applied broadly as a quantitative tool; two widely used applications are quantification of transcript abundance[18] and the determination of genome-wide genomic copy number[46-48]. For both of these applications, molecular tagging has been exploited to afford improved quantitative precision.

In the case of copy number determination, algorithms have generally employed molecular tagging, albeit implicitly, via an endogenous tagging strategy. The identity of unique progenitor fragments is inferred based on random fragmentation points and redundant "PCR duplicates" are discarded before quantification is performed. For large and complex sequence targets such as the non-repetitive portions of eukaryotic genomes, where fragmentation is sufficiently random and fold-coverage is sufficiently low, this is a powerful technique. However, at sufficiently high sampling depths, an endogenous tagging strategy is expected to fail in repetitive regions and for less complex targets (where truly independent source fragments may share fragmentation end-points by chance); exogenous tagging would overcome these limitations. Kivioja *et al* recently described an exogenous tagging-based strategy for "digital

karyotyping" based on estimates of the absolute number of molecules present in a sample; the absolute number of molecules is inferred based on the number of molecules observed and the distribution of associated tag counts with a given sampling depth[157]. To validate this method, they constructed a shotgun library from a complexity-bottlenecked sample of genomic DNA derived from a 1:1 mixture of DNA from a boy with Down syndrome and his mother, finding that tagging reduced noise in genome-wide copy number estimation. However, this comparison may be somewhat misleading, since it did not specifically investigate the added benefit of exogenous tagging relative to the endogenous tagging information also present in the library.

Depending on the library construction strategy employed, some RNA-seq workflows also implicitly use endogenous tagging to refine relative transcript abundance estimates[176]. Furthermore, Kivioja *et al* recently described a shotgun RNA-seq approach involving exogenous tagging and demonstrated it using total RNA isolated from *Drosophila melanogaster* S2 cells. They found that tagging improved the correlation between abundance estimates from libraries amplified for 15 and 25 PCR cycles compared to abundance estimates derived from read counts alone. However, this comparison also ignored the available endogenous tagging information present in the library. Finally, Shiroguchi *et al* recently described an alternate approach for RNA-seq with exogenous tagging which they validated using total RNA from *E. coli*[158]. Similarly to Kivioja *et al*, they demonstrated improved agreement of transcript abundance measurements between technical replicates using exogenous tagging compared to raw read counts; however, as in Kivioja *et al*, the comparison was made without considering endogenous tagging information, and as such is not a stringent test of the added value of exogenous tagging.

Nevertheless, as Shiroguchi *et al* point out, exogenous tagging may be especially useful in more specialized applications of genomic species abundance profiling such as nascent transcript quantification[65,66], translation rate measurements[67], small RNA sequencing, and protein-nucleic acid interaction studies[57]. Another such application is the use of molecular tagging for deep sampling of a highly identical but non-clonal population of nucleic acid molecules, such as that described by Jabara *et al*. In this study, the authors used exogenous tagging combined with Roche/454 pyrosequencing to deeply and accurately sequence the protease (*pro*) gene from a complex population of HIV-1 RNA

genomes[155]. In this application, exogenous tagging served two important purposes: first, to refine transcript abundance estimates, since no endogenous tagging information is available in this targeted scenario; and second, to create highly accurate consensus sequences for individual progenitor molecules. A more extensive discussion of the use of tagging for error correction follows.

7.5 Accurate detection of sub-clonal variation

The accurate and sensitive detection of intra-sample genetic heterogeneity (*i.e.* sub-clonal variation) is essential to numerous fields, including basic studies of evolution, massively parallel functional screens, and clinical diagnostics for infectious disease and cancer. However, the high error rate of MPS platforms has limited their application to these important problems. Molecular tagging has the potential to overcome this error rate limitation, which we showed using the Subassembly technique. However, the endogenous tagging strategy used in Subassembly is not amenable to some important applications. In 2011, Kinde *et al* demonstrated accurate and quantitative detection of sub-clonal variation using an exogenous tagging strategy, which they used to quantify the error rate of a proof-reading PCR polymerase[156]. They also demonstrate sensitive and quantitative detection of sub-clonal variation at 3% and 0.3% using a mixing study, and reduction of apparent mutation rate in human genomic and mitochondrial DNA of 24-fold and 15-fold, respectively, when comparing molecular tagging to raw data.

The PCR-based exogenous tagging strategy employed by Kinde *et al* is well-suited to detecting very low-frequency variation in a target that can be amplified and sequenced in one or a few PCR amplicons. However, this method is not expected to scale well to tens or hundreds of kilobases of target sequence. Multiplex detection of low-frequency variation in larger targets has important applications, including research and clinical cancer sequencing. To overcome these limitations, we combined the Molecular Inversion Probe (MIP) targeted capture technique[159,160] with an exogenous molecular tagging strategy[174] by adding a degenerate region to the common "backbone" sequence of the MIP to create "single molecule MIPs" (smMIPs, Chapter 6). We then validated the method using a panel of cell line genomic DNA mixtures and clinical tumor samples, many of which were derived from formalin-fixed, paraffin-embedded (FFPE) tissue. We demonstrated quantitative variant detection to 0.2% and reduction of substitution error rates in clinical samples from ~$1x10^{-4}$ to below $1x10^{-5}$. Finally, enabled by the deep

sampling of MIPs and accurate base-calling afforded by molecular tagging, we detected several low frequency variants at clinically relevant sites ranging in frequency from 0.2% to 4.7% that might have implications for patient care. The incorporation of molecular tagging into a rapid, simple, and clinically useful workflow is expected to substantially broaden the influence of this powerful paradigm. The modularity, sensitivity, and simplicity of the smMIP approach may render it especially well suited to the task of monitoring for disease recurrence or the emergence of specific drug resistance mutations based on relatively noninvasive sampling of patient tissue.

One limitation of some molecular tagging strategies is a susceptibility to polymerase errors (either stochastic or due to damaged source DNA) that occur very early in library construction. Strategies such as our tagged MIP approach that only sample one strand of a complementary duplex are especially susceptible to this type of error. To overcome this limitation, an exogenous tagging strategy was recently developed that independently labels complementary strands of a DNA duplex[173], enabling unambiguous reconstruction of independent consensus sequences for each strand of the progenitor double-stranded molecule and subsequent reconciliation of these two consensus sequences to correct first-cycle polymerase errors. Using this approach, Schmitt *et al* demonstrate single-stranded error rates of $3x10^{-5}$ and conclude that their double-stranded approach has a theoretical error rate of ~$4x10^{-10}$. The extreme sensitivity of this approach may be well suited to direct measurement of the mutation rate of DNA replication in normal and cancer cells. However, practical application of this method to the detection of ultra-low frequency variation at specific sites of interest will require integration with a targeted capture approach.

7.6 Phasing variation in complex populations of highly similar synthetic DNA constructs

Many groups have begun to explore the utility of MPS for applications beyond observational studies of genomic sequence and copy number, RNA abundance, and protein-DNA interactions. One particularly exciting area is the redevelopment of methods for saturation mutagenesis[101,177] in the context of massively parallel sequencing for the functional dissection of genomic elements regulating gene expression (*i.e.* promoters and enhancers)[20,59,178,179]. In general, these strategies work by generating a

large synthetic library of mutants of a set of regulatory elements of interest linked *in cis* to a reporter transcript harboring a sequence tag that uniquely identifies a given mutant, subjecting this library *en masse* to transcription, and using MPS to measure relative tag abundance as a proxy for the extent of transcriptional activation conferred by a given mutant regulatory element. The first such description of this approach by Patwardhan *et al*, as well as the subsequent efforts by Melnikov *et al* and Sharon *et al*, generated a tagged library of mutants by direct synthesis using programmable microarrays. The advantages of this approach are that specific single mutations and combinations of mutations can be virtually guaranteed, and that the relationship between mutant element genotype and tag sequence is specified *a priori*. However, this approach is limited in the length and number of mutant elements that can be synthesized using the microarray approach, which has in turn restricted the applicability of the method to promoters[59,179] and very small enhancers[178].

To overcome these limitations and enable saturation mutagenesis of longer elements in more complex libraries, we developed a method termed Massively Parallel Functional Dissection[20] (MPFD, Chapter 5). In MPFD, mutant regulatory element haplotypes are synthesized using polymerase cycling assembly (PCA) and cloned into a plasmid vector upstream of a minimal promoter and a reporter gene; a highly complex library of short degenerate sequence tags is separately cloned into the 3'-UTR of the reporter gene in the plasmid library. As in other implementations, these tags serve as reporters of relative transcriptional activation associated with a given mutant element; however, since the sequence of the mutant elements is no longer specified *a priori*, and the elements are themselves longer (259, 302, and 620 bp in our pilot study) than the longest Illumina read lengths (as of August 2012), the tags also serve to enable the highly accurate, full-length reconstruction of hundreds of thousands of mutant elements differing by only ~3% sequence divergence. We used an updated version of the Subassembly protocol to fully sequence the mutant elements. An exogenous tagging version of Subassembly therefore enables the full-length sequencing and subsequent functional read-out of libraries of hundreds of thousands of mutants of a given regulatory element that is itself much longer than the read-length of the most cost-effective and highest-throughput MPS platforms.

7.7 Alternatives to molecular tagging for long reads or highly accurate calls

Massively parallel sequencing is a heterogeneous grouping of related technologies. These platforms are similar in important ways but there exist many important differences with respect to, for example, library construction and sequencing workflow, instrument and run cost, per-day and per-run throughput, availability, as well as read length and error rate. As described earlier, no platform offers a combination of minimum cost, long reads, and highly accurate calls. However, there are platforms that perform impressively with respect to one of these two parameters.

Considering read length first, while the most widely available and cost-effective platform, the Illumina HiSeq, has relatively short read lengths (at most a total of 300 bp in August 2012), two other platforms offer the potential for substantially longer reads. The Roche/454 pyrosequencing-based platform[12], one of the first commercially available massively parallel sequencers, offers substantially longer read lengths (nearing one kilobase as of August 2012) than the Illumina platform. However, this platform suffers from important drawbacks, namely a labor-intensive and sensitive sequencing workflow, low throughput and high per-base cost, and high error rates, especially with respect to homopolymer sequences[16]. For these reasons, the Roche/454 platform has largely been used for relatively specialized applications, including metagenomic diversity profiling[26] and aiding *de novo* assembly[121]. Furthermore, because of the high error rates, the platform is still not generally well suited to sequencing complex but highly related populations of molecules; innovative approaches have been taken to improve sensitivity by sophisticated modeling of error processes[99,180-190] (some by explicitly considering linkage of putative variants on shared haplotypes), extending the utility of this application to deep viral gene sequencing. However, sensitivity remains limited to ~0.5%, substantially hampering application to synthetic functional genomics where haplotypes under investigation may be represented at a fraction of one in hundreds of thousands[20], and cost and throughput barriers remain substantial.

The recently released Pacific Biosciences (PacBio) instrument[43] generates even longer reads (up to ~6 kilobases in length as of August 2012). However, this instrument is also hampered by a number of important setbacks, including a high input DNA requirement, challenging workflow, extremely high error rate, and low throughput and high per-base cost. These have severely limited the applications of this

platform, although it has recently been applied successfully to aid in finishing *de novo* assemblies generated largely from Illumina sequence data[24,25]. Both the PacBio and Roche/454 platforms, therefore, have been applied successfully to some applications that lie outside the scope of the read length offered by the Illumina platform. However, throughput, cost, and error rate limitations, amongst others, continue to restrict these platforms to a subset of the long read application space.

Analytical strategies for the interpretation of raw sequencing data to detect biological variation have been a focus of substantial effort over the past several years. All of these strategies use sophisticated statistical models to accurately call variation in spite of the high error rate of massively parallel sequencing. However, they are suited to different applications, coarsely falling into one of three groups: (a) heterozygous or homozygous variant detection in entire genomes or large subsets of genomes[191-195]; (b) sub-clonal variant and haplotype detection and population reconstruction from pyrosequencing data[99,180-190]; and (c), sub-clonal variant detection from the most cost-effective and high-throughput short read sequencers[144,150-152,196-200]. Obviously, the latter two categories are most relevant to sub-clonal variant detection, although it should be noted that some data processing steps from the more traditional variant callers in the first category would likely be very useful for some sub-clonal variant detection applications. As described above, pyrosequencing-focused methods have been successfully applied to characterization of viral populations; however, these approaches remain limited by cost and throughput drawbacks of pyrosequencing.

For detection of sub-clonal variation, methods have been developed that seek to minimize the error rate and maximize the positive predictive value of variant detection from MPS data. While these methods differ in the underlying statistics, they generally work by constructing statistical models of the underlying error rate and searching for positions where variant calls above a certain quality are significantly over-represented relative to the predicted error rate. These models can be very powerful as illustrated by the very high sensitivity (to below 0.1% alternate allele frequency) and positive predictive value achieved across these various studies[144,150-152,196-200]. However, these methods are inherently limited by at least two key factors. First, the models are only as good as the training data. Models trained internally will necessarily lose sensitivity because truly variant positions are considered as errors during training. On the

other hand, a reliance on control data necessitates additional sequencing and, depending on the similarity between control and experimental samples and conditions, models trained on control data may not be applicable. Second, methods based on precise estimation of the underlying error rate remain fundamentally limited by that error rate, as they do not explicitly correct errors. As the contribution of sequencing error rates is reduced (due to improved chemistry, instrumentation, image analysis, and quality estimation, for example), PCR errors continue to limit sub-clonal variant detection. Explicit correction of PCR errors to maximize sensitivity for sub-clonal variation requires a molecular tagging approach.

7.8 Conclusions

The advent and maturation of MPS technologies has transformed many areas of biomedical research due to dramatic improvements in cost and throughput; however, many important applications in basic research and clinical medicine have remained refractory to these platforms because of drawbacks with respect to read length and error rate. To overcome these limitations and better exploit MPS for a number of exciting applications, several groups have developed methods that are individually unique but share the common paradigm of uniquely labeling individual input nucleic acid molecules such that sequencing reads derived from copies of the same progenitor molecule can be appropriately grouped (*i.e.* "molecular tagging"). Grouping of reads sharing a common molecular origin enables correction of polymerase and sequencing error, extension of effective read length, and improved quantitative precision.

Multiple factors influence the decision to employ a molecular tagging strategy for a given application. First, the introduction of molecular tagging may add complexity to library construction, sequencing, and analysis. However, many sequencing workflows already implicitly involve some form of endogenous tagging, and more sophisticated analytical approaches will more explicitly incorporate this information. Furthermore, even for applications that are not amenable to endogenous tagging, the introduction of exogenous tagging may be straightforward (*e.g.* into the molecular inversion probe targeted capture method[174]). Finally, continued development of bioinformatics tools for molecular tagging data analysis is expected to facilitate more widespread adoption.

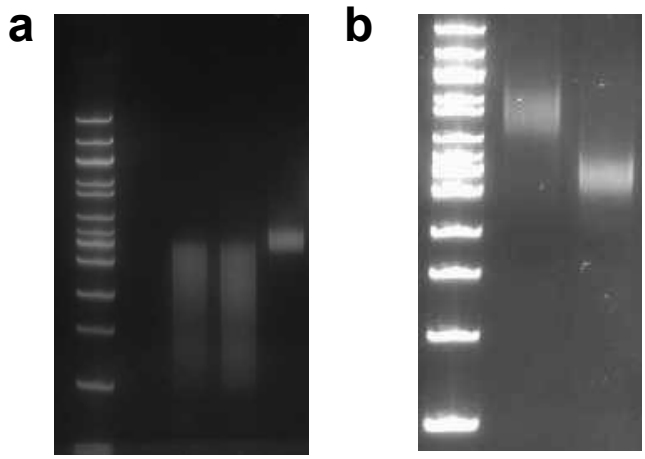
Another factor influencing adoption is the continuing development of commercial sequencing platforms, including continually improving read lengths and base-calling accuracy of the most cost-effective platforms, further development of and improved analytical strategies for the more specialized long read platforms (*i.e.* Roche/454, PacBio), and the emergence of even newer technologies that may bridge the gap between throughput and read length (but still seem to suffer from high systematic error rates[201]), namely the Ion Torrent[15] PGM and Proton. For some applications such as *de novo* genome assembly, a hybrid approach leveraging both short read and long read MPS platforms is emerging as the standard in the field, and currently practical molecular tagging methods may not be competitive. However, the massively parallel hierarchical assembly conceptual framework introduced by Subassembly may underpin a new generation of transformative methods as molecular biology and sequencing technologies continue to develop. For example, pioneering efforts in haplotype-resolved sequencing of human genomes using MPS have relied on labor- and cost-intensive physical compartmentalization of genomic subsets[171,202,203]; the development of methods to tag and amplify longer DNA fragments (*i.e.* tens or hundreds of kilobases) in an *in vitro* and multiplex fashion and sequence tagged fragments descended from these long progenitors, would dramatically increase the practicality, feasibility, and availability of haplotype-resolved genome sequencing. Such a method would also substantially aid *de novo* assembly efforts, especially for complex eukaryotic genomes with high repetitive sequence content.

In addition to *de novo* assembly and haplotyping, other applications such as synthetic functional genomics may continue to demand read lengths longer than those offered by the short read platforms and error rates lower than the long read platforms, representing a continuing opportunity for the application of molecular tagging to applications requiring long effective read lengths. As base-calling accuracy improves and if full use of the endogenous tagging information already available in most workflows becomes more widespread, exogenous molecular tagging will be most useful in situations where high quantitative precision is desired. If, however, a given application is not amenable to endogenous tagging but high base-calling accuracy and/or quantitative precision are desired, exogenous tagging remains an effective option.

Molecular tagging is therefore a general and broadly useful strategy to improve and broaden the utility of MPS. Despite improvements in commercial sequencing platforms as well as newly emerging platforms, there remain applications requiring some combination of long reads, highly accurate base-calls, and high quantitative precision that would benefit from a molecular tagging strategy in some form. Continued optimization of existing molecular tagging methods and analytical tools will lead to more widespread adoption of these useful approaches in the research and clinical communities, while the application of molecular tagging concepts in new contexts will facilitate further advances in biomedical research.

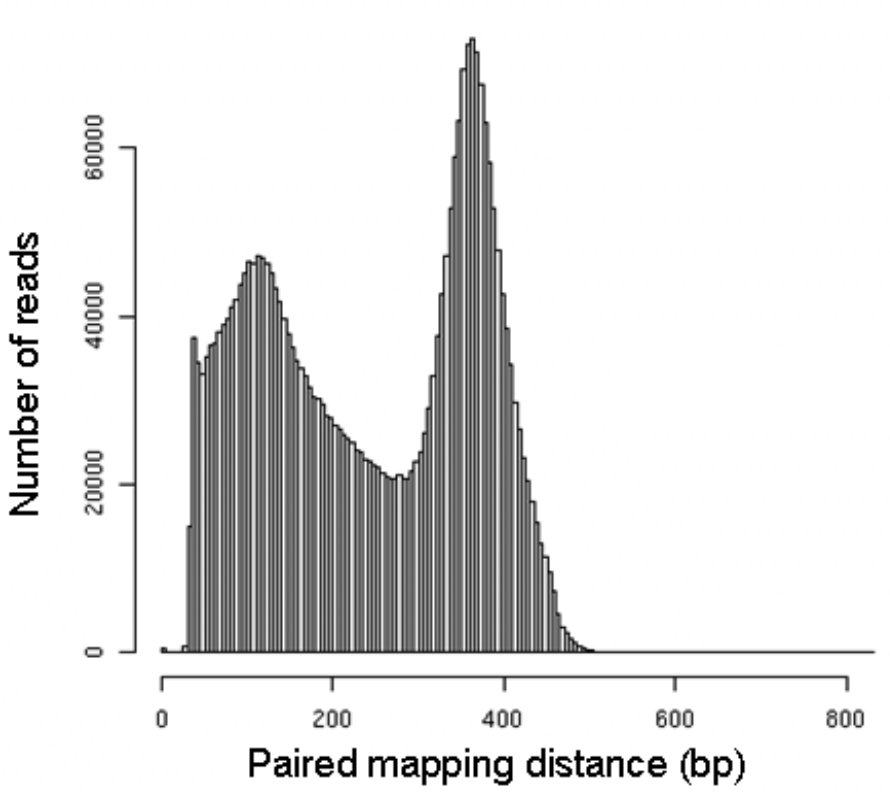
Appendix A Supplementary material for Chapter 4

Figure A.1. Length of library fragments by PAGE.



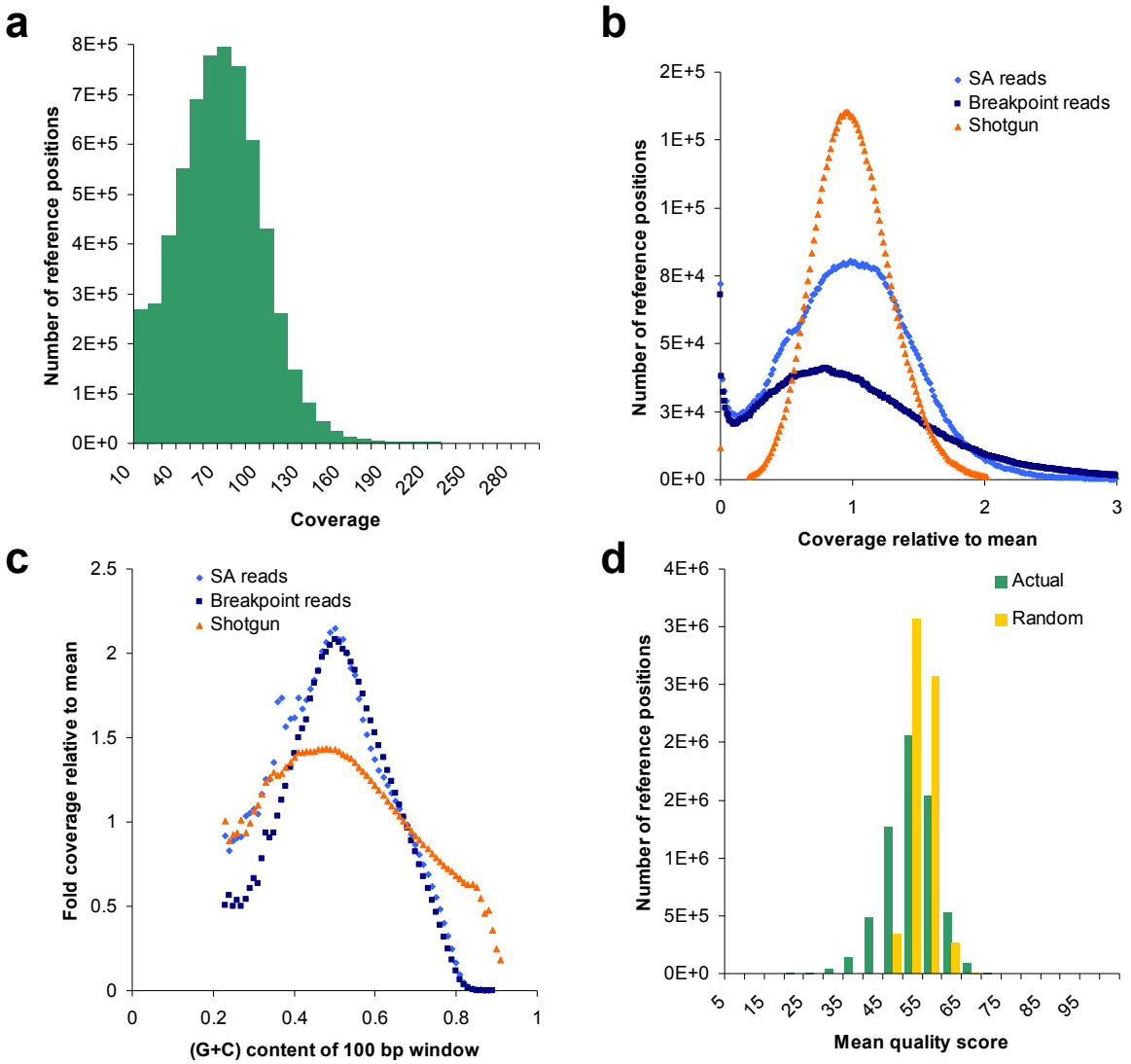
(a) PAGE of NEB 100 bp ladder and nebulized and size-selected ~550 bp *P. aeruginosa* fragments. (b) PAGE of NEB 100 bp ladder and Biorupted and size-selected Methylamine metagenomic fragments.

Figure A.2. Length distribution of subassembly fragments by paired-end sequencing



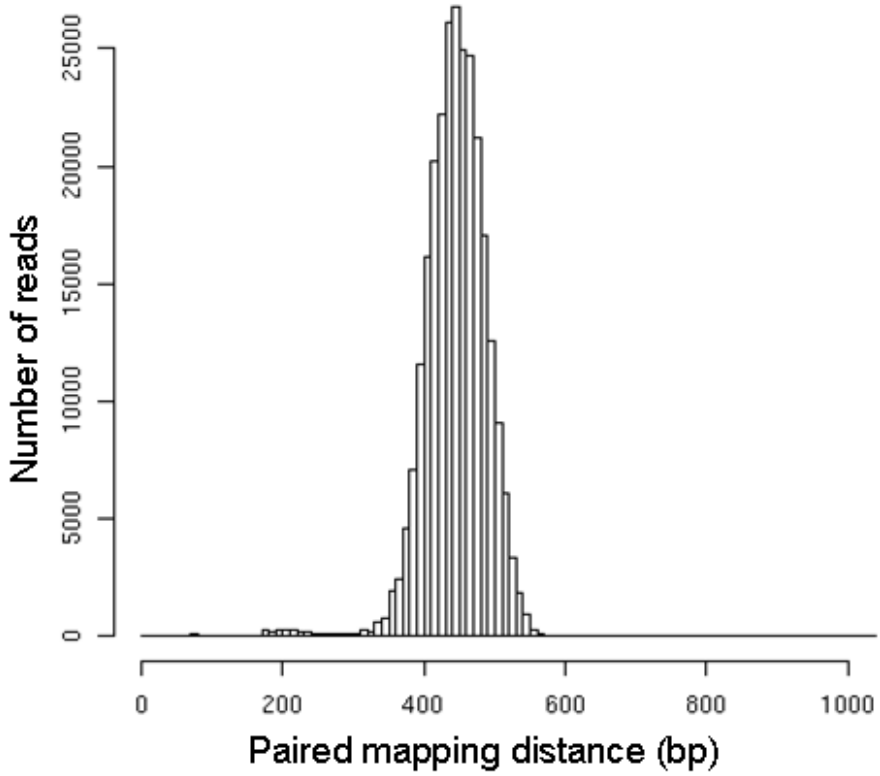
Histogram of mapping distance separating tag and breakpoint reads from the 450-600 bp size-selection performed at the end of the subassembly library construction protocol of a representative subset of the *Pseudomonas* data. Paired 20x76 bp reads were mapped to the PAO1 reference genome using *maq.* Shorter mapping distances are thought to arise from over-amplification during PCR, which causes shorter fragments to migrate with longer fragments during PAGE. Retained shorter fragments are then preferentially amplified and sequenced during the Illumina sequencing protocol. Careful PCR amplification is essential to prevent small fragments from completely dominating the sequencing reaction. The non-uniform nature of this distribution may contribute to the bimodal distribution of subassembled read length that we observed for this sample.

Figure A.3. Coverage of the PAO1 reference by SA reads



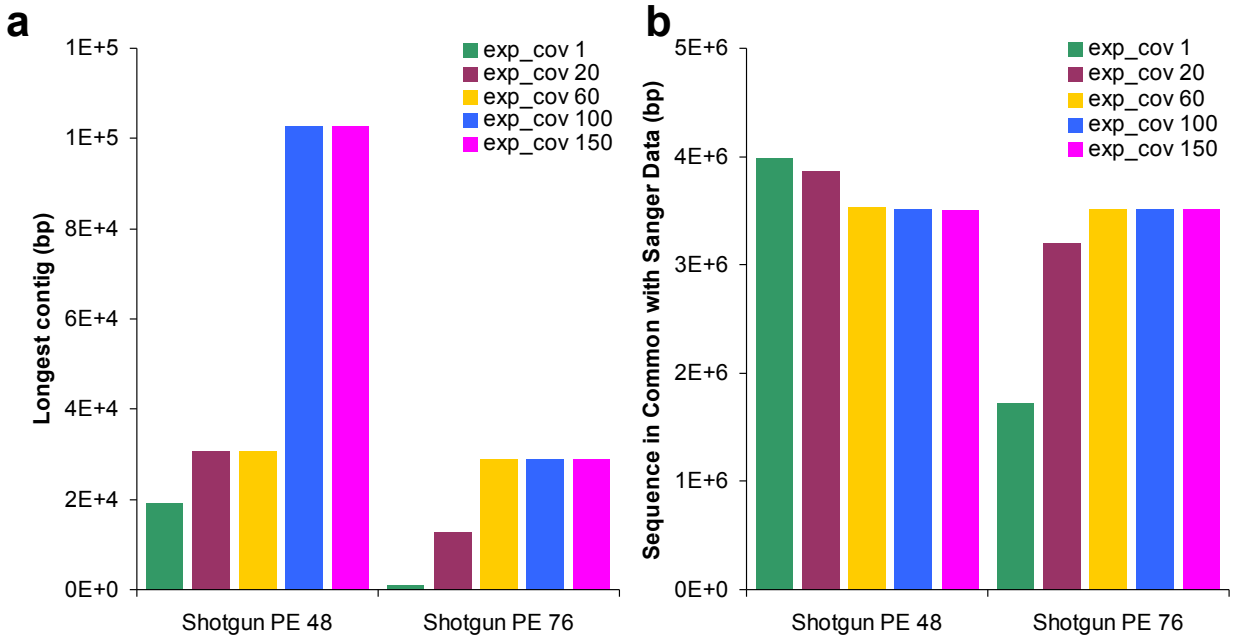
(a) Histogram of coverage of the PAO1 reference by SA reads as determined by BLAST alignment (bin size 10 bp). (b) Histogram of coverage of the PAO1 reference by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. (c) Mean (G+C) content in the 100 bp window around reference positions with a given coverage on the x-axis by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. A strong relationship between coverage and (G+C) content is observed. That is, reference bases in very high (G+C) content regions tend to have reduced coverage relative to the mean, and regions with intermediate (G+C) content are correspondingly overrepresented. This is likely due to (G+C) content biases present during the PCR steps of library construction, as a similar relationship is observed for the contributing 76 bp reads, and could likely be mitigated by PCR conditions designed to reduce (G+C) bias. (d) Distribution of mean quality score (and therefore predicted error rate) across the reference. The number of reference positions with a given mean quality score is plotted in green ("Actual"), while a simulated distribution was made by randomizing the full set of quality score assignments in SA reads and then recalculating mean quality scores for reference positions, and is plotted in yellow ("Random"). The standard deviation of the actual distribution was six compared to three for the random distribution, indicating a small systematic bias in quality score (and therefore error) distribution across the PAO1 genome.

Figure A.4. Length distribution of metagenomic fragments



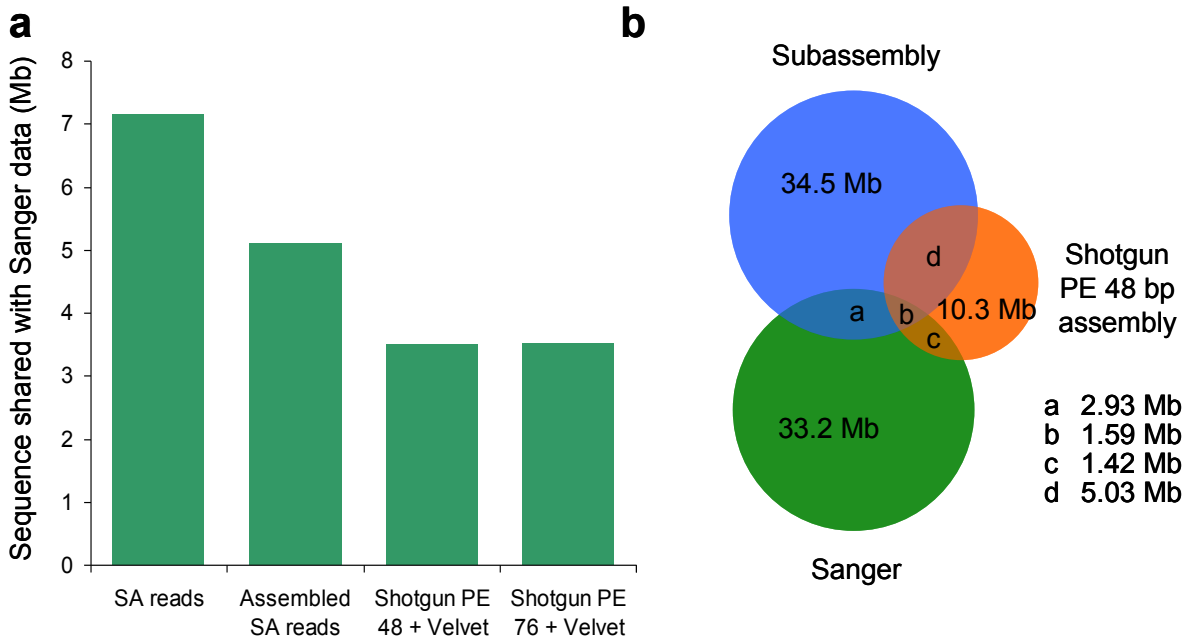
Histogram of the mapping distance separating paired tag reads from 36 bp paired-end sequencing data of the metagenomic library fragments (used to pair and merge TDRGs). Paired-end reads were mapped to the recently obtained Sanger data from the same sample using *maq.* Some selection for shorter molecules during the Illumina sequencing protocol may have taken place, shifting the peak of the distribution somewhat shorter than would be expected based on PAGE of the original fragments (Figure A.2).

Figure A.5. Optimization of Velvet parameters for shotgun metagenomic assembly



We optimized Velvet parameters for shotgun metagenomic assembly with respect to contig length and sequence shared with available Sanger data. (a) Maximum contig length as a function of changing Velvet parameters for assembly of shotgun paired-end 48 bp and paired-end 76 bp reads. Contig length was found to be very sensitive to the exp_cov parameter. (b) Sequence in common with the available Sanger data from the same sample as a function of changing Velvet parameters as in (a). Shared sequence was found to be somewhat sensitive to the exp_cov parameter in an unpredictable fashion, with shared sequence decreasing with increased exp_cov for the 48 bp reads and increasing with increased exp_cov for the 76 bp reads. To optimize length and coverage, we chose to perform subsequent analyses with the exp_cov = 100.

Figure A.6. Coverage overlap of metagenomic sample between sequencing methods



(a) Sequence shared with Sanger data for SA reads, assembled SA reads, and Velvet-assembled shotgun paired-end reads. Shared sequence was estimated by considering BLAST alignments with at least 98% identity across at least 100 bp. SA reads covered more than twice as much of the Sanger data as either shotgun assembly. (b) Venn diagram illustrating reciprocal coverage across data sets as determined by stringent BLAST analysis. Contigs produced by Celera assembly of SA reads, Velvet assembly of a 48 bp paired-end shotgun library with exp_cov=100, and the recently obtained Sanger sequencing data were compared to one another using BLAST. Coverage was defined as the best pair-wise match between bases as determined by the bit-score of the alignment as long as the alignment had at least 98% identity and was at least 100 bp long. The bases in common shown here are not in exact agreement with those presented in (a) because, for the purposes of constructing this diagram, each base was only allowed to align to one corresponding base in another dataset. Circles are drawn to scale; regions of overlap not to scale.

Note A.1. The importance of a complexity bottleneck

A complexity bottleneck is needed so that multiple overlapping, randomly-positioned breakpoint reads can be observed for each member of the long fragment library with a reasonable amount of sequencing. In other words, it is necessary to sample each nested sub-library in sufficient depth to reconstruct the sequence of the parent long molecule. For example, we obtained approximately 60 million read-pairs across six lanes of Illumina sequencing that enabled us to reconstruct part or all of the sequence of approximately one million long molecules. If we had used a library containing 100 million long molecules, we would only have observed, on average, less than one read-pair per long molecule, preventing any subassembly from taking place within most if not all sub-libraries. The only exception to this principle is in the case of a very small effective genome size and if the ends of molecules (and not degenerate synthetic adaptors) are used as tags. For example, in the case of a genome of only 500 kilobases, the maximum number of unique tag reads (assuming no repetitive sequence at the scale of the tag read) is one million, which we have shown to be a tractable library complexity. In such a situation, it might not be formally necessary to restrict library complexity.

Note A.2. Filtering of predicted misassemblies

Manual inspection of predicted misassemblies revealed four contigs that were incorrectly called misassemblies because of differences between the strain that we sequenced and the reference PAO1 strain. Three of these (2548, 2129 and 2115) exhibited extremely high sequence identity with a phage-like insertion in PAO1 that was recently added to GenBank (ID GQ141978.1) and that we have observed in independent shotgun sequencing data from our strain. Notably, the same phage-like insertion seems to have caused the lone misassembly in our scaffolds (Scaffold_LR7_3). The fourth contig (2622) spans a ~1 kb deletion in our strain that we have also observed in independent shotgun sequencing data (data not shown).

Note A.3. Comparison of *de novo* assembly to hybrid 454-Illumina approach

We compared the performance of our method to a recently published, high quality *de novo* assembly from a similar but significantly lower (G+C) content organism (66.6% versus 58.5%), which was generated by combining both long-read and long-range mate-paired 454 data with short distance paired-end Illumina data[121]. We find that our method compares very favorably to that approach with respect to N50 (445 kb versus 92 kb), longest scaffold (915 kb versus 389 kb), substitution error rate (~1/14,000 versus ~1/7,000), and number of rearrangements (one versus twenty). It should be noted that the authors of that study also performed sequencing and assembly of a related organism without a reference genome and achieved apparently better performance (N50 of 532 kb, longest contig of 794 kb), which they attempted to validate with limited Sanger sequencing. However, it is difficult to make a direct comparison with respect to accuracy in the absence of a reference genome. Our method also used significantly more raw data than that study, but only required a single sequencing platform, which may increase its general utility.

Note A.4. Estimated cost of subassembly protocol

Although it is difficult to draw firm conclusions in the face of rapidly changing costs associated with many second-generation sequencing platforms, it is clear that subassembly is significantly more expensive than standard shotgun Illumina sequencing if only the total amount of sequence produced is considered. However, as subassembly produces much longer reads at much higher per-base accuracy than the raw reads from the Illumina platform, such a comparison is not valid. Even the comparison to Roche/454 sequencing, which produced reads in the hundreds of base-pairs, is difficult because of the decreased accuracy of that method relative to the method we present here. Still, we estimate that our method is roughly cost-comparable to Roche/454 sequencing. For example, if a lane of sequencing is assumed to

cost ~$2,000, from six lanes of sequencing we generated 405 Mb of long SA reads for the *P. aeruginosa* sample, which corresponds to a cost of ~$30/Mb, or about half that of recently published estimates of the cost of Roche/454 Sequencing [16]. However, the reduced error rate is a critical differentiator, making the cost comparison tenuous. A major advantage of subassembly is that extremely low error rates and long effective read length is maintained independent of sample complexity. In the case of short read sequencing (Illumina, AB SOLiD, Helicos), read length and error limitations can be overcome through the use of very high coverage. The ability to achieve high coverage depends implicity on sample complexity and can be complicated by relatedness of sequences therein. With Roche/454 sequencing, read lengths are longer, but once again, errors can only be overcome with high coverage, which again may be impossible in the case of either very high sample complexity or the presence of highly related sequences. We therefore conclude that subassembly produces equivalently long sequences at below or equal to the cost of Roche/454 sequencing with length and error performance that remains independent of sample complexity and sequence relatedness, a feature of no other currently available second-generation sequencing method.

Table A.1. Phrap optimization

Min match	Min score	Force level	Index word size	# of TDRGs	Mean longest SA read	Median longest SA read	Fraction of non-BLASTing SA's	Fraction of SA's BLASTing <90% of length	Fraction of mismatches among BLASTing bases
12	12	1	10	2619	361.6	403	0.004964	0.02993	0.001513
10	12	1	10	2619	364.4	406	0.004964	0.0284	0.001543
10	12	1	8	2619	364.4	406	0.004964	0.0284	0.001543
10	10	1	8	2619	369.5	409	0.004964	0.04106	0.001551
8	10	1	8	2619	371.9	411	0.004964	0.04643	0.001579

A representative subset of 10,000 Pseudomonas TDRGs was randomly selected and subjected to phrap assembly using different parameters and the resulting lengths and qualities of the longest subassemblies from each TDRG were assessed. We determined that parameters of minmatch 10, minscore 12, force level 1, and index word size 8, achieved the optimal balance between assembly accuracy, measured as the fraction of subassembled reads BLASTing across at least 90% of their length in a single BLAST hit (and the fraction removed because of oppositely oriented reads, not shown), and subassembled read length.

Table A.2. Summary statistics for subassembled reads

Sample	Original fragment size	# of read-pairs	# of filtered TDRGs	Median length
P. aeruginosa	~550 bp	56.8M	1,031,537	338 bp
Metagenomic	~450 bp	21.8M	262,298	256 bp
Metagenomic (merged)	~450 bp	21.8M+1.8M	180,008 (90,004 pairs)	408 bp

For the two samples used and the two analyses performed of the methylamine-enriched metagenomic sample, listed is the approximate size of long fragments from which subassembly libraries were generated, the number of Illumina read-pairs that were used to generate subassembled (SA) reads (merged analysis also shows the number of reads used to pair tags), the number of TDRGs after filtering for successful assembly and properly oriented contributing reads, and the median length of the longest SA read from each filtered TDRG.

Table A.3. Summary statistics from assembly of metagenomic SA reads versus assembly of a standard shotgun library

Input	Assembly strategy	# of contigs	Median contig length	Sequence in contigs ≥ 200 bp	Longest contig
SA reads	Celera	86,418	390 bp	35.7 Mb	6,000 bp
Shotgun PE 48 bp	Velvet (exp_cov = 100)	17,618	332 bp	9.9 Mb	102,806 bp
Shotgun PE 76 bp	Velvet (exp_cov = 100)	33,374	315 bp	16.0 Mb	28,861 bp

Comparison of assembly of short reads from a standard Illumina shotgun library prepared from the metagenomic sample to Celera assembly of the full complement of SA reads from the same sample. Listed is the assembly input, the assembly strategy used, and, for contigs at least 200 bp long, the number of contigs produced, the median contig length, the total amount of sequence contained in such contigs, and the longest contig. 76 bp paired-end (PE) reads were collected from a standard shotgun library and were trimmed to 48 bp reads to match the amount of sequence collected per read-pair for subassembly (20+76). Velvet assembly was performed using both 48 bp and 76 bp paired-end reads, but the same total amount of raw sequence as collected for subassembly (2.2 Gb) was used in each shotgun assembly. Notably, while the shotgun assemblies achieve greater contiguity at the longest lengths, potentially due to deep sampling of abundant genomes or to misassemblies, subassembly produces at least twice as much sequence at the lengths necessary for phylogenetic analysis and gene prediction.

Table A.4. Oligo sequences

	Name	Sequence
Bottleneck adaptor oligos	Ad1	TCGCAATACAGAGTTTACCGCATT
	Ad1_rc	/5Phos/ATGCGGTAAACTCTGTATTGCGA
	Ad2	CTCTTCCGCATCTCACAACCTACT
	Ad2_rc	/5phos/GTAGGTTGTGAGATGCGGAAGAG
Illumina adaptor oligos	Illum_rev	CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
	Illum_rev_rc	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Bottleneck PCR primers	Ad1_amp	/5phos/TCGCAATACAGAGTTTACCGCATT
	Ad2_amp	/5phos/CTCTTCCGCATCTCACAACCTACT
TDRG merging PCR primer	Illum_amp_r_Ad 2	CAAGCAGAAGACGGCATACGAGATATCGAGAGCCTCTTCCGC ATCTCACAACCTACT
Sequencing PCR primers	Illum_amp_f_Ad 1	AATGATACGGCGACCACCGAGATCTACACCAATGGAGCTCGC AATACAGAGTTTACCGCATT
	Illum_amp_f_Ad 2	AATGATACGGCGACCACCGAGATCTACAC ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_amp_r	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGC TGAACCGCTCTTCCGATCT
Oligos used in sequencing	Ad1_seq	CAATGGAGCTCGCAATACAGAGTTTACCGCATT
	Ad2_seq	ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_seq_r	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Oligos were obtained from Integrated DNA Technologies. An asterisk indicates a phosphorothioate bond. /5Phos/ indicates a five-prime phosphate modification.

Methods

Subassembly library construction

Source DNA was fragmented by sonication, end-repaired and size-selected to ~550 bp (P. aeruginosa) or ~450 bp (metagenomic sample). Size-selected fragments were A-tailed and ligated to custom adaptors (Table A.4). Real-time PCR with phosphorylated primers was performed using serial dilutions of adaptor-ligated fragments to impose a complexity bottleneck and generate many copies of a limited number of long fragments. Complexity was estimated from the concentration of input material, the kinetics of PCR amplification and gel electrophoresis of the PCR product. After PCR, the product estimated to have resulted from ~105–107 long fragments was concatemerized to high molecular weight and then fragmented by sonication. Shearing products were end-repaired, A-tailed and ligated to the Illumina Read 2 adaptor. PCR amplification was then performed with one primer corresponding to the Read 2 adaptor and a second primer corresponding to one of the two original adaptors. Finally, the amplification products were size-selected to obtain a uniform distribution of shearing products across the original fragment (Figure A.2). For the metagenomic effort, an aliquot of the bottleneck PCR was subjected to an additional round of PCR to prepare the long fragments for paired-end sequencing and subsequently used for tag-pairing and TDRG merging.

Shotgun library construction

P. aeruginosa short insert (~200 bp) and long insert (~2.5 kb), and metagenomic short insert shotgun libraries were constructed according to manufacturer's specifications, except that standard oligonucleotides were obtained from IDT. For the metagenomic library, to conserve source material, size selection to the desired fragment length was performed before A-tailing and adaptor ligation rather than afterward so that the longer size range could be used for subassembly.

Illumina sequencing

For subassembly libraries, an Illumina GA-II instrument was used to collect paired-end reads according to manufacturer's specifications, except that custom sequencing primers (Table A.4) were used, and asymmetric read lengths were collected (20-bp first read and 76-bp second read). For the tag-pairing metagenomic library, paired-end 36-bp reads were collected according to manufacturer's specifications with custom sequencing primers. For shotgun libraries, paired-end reads were collected according to manufacturer's specifications.

Organizing breakpoint short reads into TDRGs

For all experiments, breakpoint reads paired with identical or nearly identical tag sequences were grouped into TDRGs. As millions of tag reads were involved, an all-against-all comparison to cluster similar tags was not feasible. Instead, a two-step strategy was used to group tag sequences in each experiment. First, perfectly identical tags were collapsed using a simple hash to define a nonredundant set of clusters. From this set, clusters with four or more identical tags were identified as 'core' clusters and, in descending order by size, were compared to all other tags. Tags matching a given core cluster with up to one mismatch were grouped with that core cluster (and removed from further consideration if they themselves defined a smaller core cluster). TDRGs with more than 1,000 members were excluded from downstream analysis to limit analysis of adaptors or other low-complexity sequence.

Subassembly of TDRGs

Each TDRG was assembled separately using phrap with the following parameters: "-vector_bound 0 -forcelevel 1 -minscore 12 -minmatch 10 -indexwordsize 8". Pre-grouping reads into TDRGs allowed us to use less stringent parameters than the defaults used in traditional assemblies. Parameters were optimized to balance SA read length and accuracy (Table A.1). A short-read assembler, Velvet, was also tested but did not produce substantial gains in SA read length relative to phrap (data not shown).

Trimming and filtering of SA reads and assignment of consensus quality scores

SA reads were masked using the cross_match program provided as part of the phrap suite, using the following parameters: "-minmatch 5 -minscore 14 –screen". Determination of consensus quality scores and further trimming was performed as follows. Because it permits multiple alignments per read, the Bowtie short-read alignment tool [204] was used to map contributing 76-bp breakpoint reads to the SA reads to generate consensus quality scores for SA read base calls. Only alignments within TDRGs were allowed (that is, alignments of breakpoint reads to SA reads from another TDRG were ignored). Bowtie was also used to map the 20-bp tag reads back to the SA reads to facilitate end trimming where the SA read had extended into adaptor sequence. Next, SA reads were trimmed using both tag read mapping and adaptor masking information. SA reads were first trimmed from the 3′ end using the mapping location of the tag read; if bases remained that had been masked by cross_match because of the presence of adaptor, the masked bases were removed and the longest remaining continuous sequence was retained. Finally, any sequence containing a base call with quality below 10 within 5% of the 3′ end of the SA read was discarded.

In all subsequent analyses, only SA reads that were at least 77 bp long and were assembled from identically oriented short reads were considered. The read orientation filter was only applicable to SA reads from individual, unmerged TDRGs. In addition, for length and quality analyses, only the longest SA read from each TDRG was analyzed.

Quality assessment

The longest SA read (after trimming as described above) from each TDRG containing at least 10 member reads was aligned to the P. aeruginosa PAO1 reference genome using BLAST with the following parameters: "-p blastn -e 1e-6 -m 8 -F F -a 4".

Error rate as a function of quality score and position in the SA read was then determined as follows. BLAST alignments containing at least 95% of the length of the SA read query and without any gap openings were used to define the position in the reference of the SA read in question (the BLAST coordinates were extended to encompass the entire length of the SA read). Every base in an SA read whose alignment meets the above criteria was compared to the corresponding reference base. If less than 100% of the SA read aligned, the comparison was forced to extend to the ends of the SA read. From the base-by-base comparison, the error rate as a function of base call quality or position in the SA read was calculated.

We did not perform a base-by-base comparison for cases in which BLAST used a gap opening in making an alignment, which could potentially suppress our error rates if such SA reads were substantially more error-laden. Accuracy of such SA reads within aligned regions was slightly lower (99.56% accurate compared to 99.86% in SA reads without gaps), and such sequences only comprised less than 1% of the sequence being analyzed. We therefore concluded that errors in these sequences that fall outside of aligning regions are unlikely to substantially alter our estimates of error rate as a function of base quality. We performed a similar analysis for SA reads containing larger gaps with respect to the reference (those with a BLAST alignment less than 95% of their length), as we did not perform a base-by-base comparison

for such SA reads either. Once again, the accuracy with aligned regions was somewhat lower (99.4% versus 99.86% in those with complete or nearly complete alignments). Such errors probably reflect larger-scale misassemblies owing to repetitive sequence in the true reference sequences. Notably, aggressive trimming substantially reduced the relative abundance of such sequences; only 1.5% of the total number of bases analyzed was contained in such sequences, and only 2.3% of BLAST alignments fell into this category. Once again, forcing the alignment to the very edges of such SA reads was not likely to substantially alter the relationship between error rate and base call quality score.

To analyze quality as a function of raw read base quality, maq was used to align contributing 76-bp breakpoint reads to the reference, Illumina base calls were compared to the reference and, for a randomly chosen subset of 1 million bases, the error rate as a function of Illumina base call quality was determined.

To analyze quality as a function of raw read position, a representative lane of contributing 76-bp breakpoint reads used for the subassembly process was aligned to the reference genome using maq, and the error rate at each position was determined by comparing read base calls to reference bases for each read.

Assembly of SA reads using the Celera assembler (CABOG)

For P. aeruginosa and metagenomic samples, all trimmed, orientation- and length-filtered SA reads (not only the longest per TDRG) were subjected to assembly using the Celera assembler. Assembly was guided by consensus quality scores generated as described above. The Celera assembler (CABOG) was run with default parameters and “unitigger=bog”.

Assessment of assembled SA read quality

Contigs produced by the Celera Assembler from SA reads were aligned to the reference using BLAST with the following parameters: “-p blastn -e 1e-6 -m 8 -F F”. Substitution error rate was measured as the number of mismatches within the best BLAST alignment for each contig. To account for a potentially higher error rate in misassembled contigs, if a contig aligned across less than 95% of its length, other BLAST alignments were also considered as long as they comprised at least 10% of the contig length.

Scaffolding of contigs for P. aeruginosa

For de novo assembly of the P. aeruginosa genome, we used independently produced shotgun sequencing libraries to scaffold the contigs produced from SA reads as follows. The resulting contigs were scaffolded using a custom script that used 36-bp shotgun paired-end Illumina reads from one lane each of short-insert (~200 bp) and long-insert (~2.5 kb) libraries. The gap between each pair of adjacent contigs in a scaffold was dynamically estimated based on the distance of the read pairs connecting the two contigs from the ends of the contigs and the expected insert size of the library from which they were derived. Scaffolds were then constructed by separating the contigs by a string of unknown nucleotides (Ns) as long as the estimated gap size. For cases where the expected gap size was close to zero or negative (indicating a possible overlap), the adjacent ends of the two contigs were subjected to a Smith-Waterman alignment and merged accordingly if a match was detected.

TDRG merging algorithm

Paired 36-bp reads were obtained from a sequencing library prepared from bottlenecked, adaptor-ligated metagenomic fragments, then trimmed computationally to 20 bp to correspond to the length of the tag reads that were obtained during sequencing of the subassembly libraries.

To prevent sequencing errors at the ends of the reads from creating spurious tags and tag pairs, we trimmed the reads further to the first 15 bp. If multiple TDRGs (defined by 20-bp tags) could correspond to a single 15-bp tag from a merging read pair, the TDRG with the most members was chosen. In descending order of tag-pair abundance, we defined TDRG pairs, removing tags that had been assigned to TDRG pairs as we proceeded.

Velvet assembly of shotgun metagenomic library

Paired-end shotgun reads constructed according to standard Illumina protocols were assembled using Velvet with the following parameters: "-cov_cutoff 2 -exp_cov [variable] -ins_length 250 -unused_reads yes".

If exp_cov was set to 1, cov_cutoff was set to 0. As Velvet (along with all other short-read assemblers) is not designed for assembly of metagenomic sequences, considerable effort was made to optimize its performance with respect to length of sequences produced and agreement with the available Sanger sequencing data to make the fairest comparison possible. We found that contig length was sensitive to the exp_cov parameter (Figure A.5). However, we observed unpredictable performance with respect to agreement with the Sanger sequencing data when altering this parameter, as agreement improved for the paired-end 76-bp reads but degraded for the paired-end 48-bp reads. We therefore chose an exp_cov value of 100 as the best compromise of sequence length and coverage for the comparator datasets.

Resulting scaffolds were then split into contigs that did not contain Ns, as we reasoned that key goals of metagenomic sequencing such as gene discovery and phylogenetic classification would depend solely on the length of contiguous regions of defined bases.

Comparison to Sanger sequencing data with BLAST

Contigs produced from SA reads with CABOG and contigs produced from shotgun short reads with Velvet were aligned to one another and to the recently collected Sanger sequencing data from the same sample (JGI IMG/M Taxon Object ID 2006207002, NCBI accession number ABSR01000000) using BLAST with the following parameters: "-p blastn -e 1e-6 -m 8 -F F". Two bases were considered to be a shared position between two datasets if they were contained in a BLAST alignment at least 100 bp long and with at least 98% identity. For the Venn diagram (Figure A.6), an additional restriction was added so that mappings between the three datasets were not ambiguous: the two bases were required to be in the BLAST alignment with the highest bit score of all the BLAST alignments between the two datasets involving either base.

Appendix B Supplementary material for Chapter 5

Note B.1. Multiple linear regression on entire haplotypes

While linear models constructed on a position-by-position basis best represent the effect size of individual mutations, they may not perform optimally as predictors of the transcriptional activity of entire haplotypes, which contain many such mutations. To assess the ability of models constructed from our data to predict overall haplotype activity, we built two multiple linear regression models for each enhancer. The first model was composed of *n* binary variables (where *n* is the length of the enhancer) for whether or not a position was wild-type in an enhancer haplotype, and the second model was composed of 3*n* binary variables for whether a position was a particular mutant nucleotide in an enhancer haplotype (Table B.1). While all the models were significant as measured by comparison of mean squared error calculated from actual versus data versus data with the outcome vector permuted ($p<0.01$), the explanatory power of these models (R^2) ranged from 0.03 to 0.3, suggesting that complexity bottlenecking has limited the ability of our models to explain large fractions of the observed variation for entire haplotypes. Specifically, the relatively few numbers of tags with which individual haplotypes are associated, and the relatively few aliquots in which individual tags are observed, adds considerable stochastic noise to the system.

Table B.1. Predictive power and significance of multiple linear regression models

	n term model		3*n* term model	
Library	**R^2**	**p**	**R^2**	**p**
ALDOB	0.03	< 0.005	0.05	< 0.01
ECR11	0.12	< 0.005	0.19	< 0.01
LTV1 rep. 1	0.21	< 0.005	0.29	< 0.01
LTV1 rep. 2	0.22	< 0.005	0.30	< 0.01

Multiple linear regression models taking into account all positions (*n* term model, where *n* is the length of the enhancer) or all mutations at all positions (3*n* term model), were constructed for each of the enhancers. Listed here are R^2 values and p-values (computed by constructing models for 200 or 100 random permutations of the outcome vector and comparing mean squared errors from the permuted data models to the actual data model).

Table B.2. Predicted transcription factor binding sites.

Enhancer	Start pos.	End pos.	Strand	Factor	Core Match	Matrix Match	Enhancer	Start pos.	End pos.	Strand	Factor	Core Match	Matrix Match
aldob	16	29	-	C/EBPbeta	0.833	0.799	ecr11	212	226	-	HNF-3beta	1	0.842
aldob	17	27	+	AP-1	1	0.975	ecr11	217	230	-	C/EBPbeta	0.829	0.786
aldob	41	52	+	Oct-1	1	0.91	ecr11	219	228	+	GATA-3	0.977	0.886
aldob	45	58	-	C/EBPbeta	0.833	0.776	ecr11	221	236	+	GR	1	0.848
aldob	92	105	-	HNF-4	0.825	0.811	ecr11	247	259	-	CHOP - C/EBPalpha	0.778	0.819
aldob	112	125	-	C/EBPbeta	0.821	0.762	ecr11	249	262	+	C/EBPbeta	0.833	0.863
aldob	120	130	-	AP-1	0.935	0.868	ecr11	249	262	-	C/EBPbeta	0.816	0.861
aldob	135	151	+	HNF-1	1	0.795	ecr11	264	273	+	USF	0.918	0.868
aldob	140	149	+	TATA	1	0.888	ecr11	272	281	+	USF	0.918	0.873
aldob	141	155	-	HNF-3beta	1	0.835	ecr11	361	371	+	AP-1	0.935	0.854
aldob	148	158	-	AP-1	0.935	0.866	ecr11	382	401	-	YY1	1	0.855
aldob	149	166	-	NF-1	1	0.983	ecr11	390	407	+	NF-1	0.911	0.878
aldob	150	163	-	HNF-4	0.796	0.802	ecr11	394	403	+	USF	0.905	0.88
aldob	153	162	+	USF	0.945	0.932	ecr11	407	420	-	C/EBPbeta	0.816	0.787
aldob	157	170	-	HNF-4	0.988	0.892	ecr11	429	438	+	USF	0.905	0.854
aldob	170	179	+	USF	0.931	0.852	ecr11	438	447	+	TATA	1	0.882
aldob	201	217	+	HNF-1	0.942	0.706	ecr11	460	474	-	HNF-3beta	1	0.894
aldob	205	215	+	AP-1	0.935	0.922	ecr11	465	478	-	C/EBPbeta	0.829	0.786
aldob	209	225	-	HNF-1	0.771	0.632	ecr11	465	481	-	HNF-1	0.829	0.712
aldob	217	230	+	C/EBPbeta	0.859	0.784	ecr11	467	476	+	GATA-3	0.977	0.886
aldob	247	256	+	GATA-3	1	0.935	ecr11	474	487	+	C/EBPbeta	0.883	0.815
aldob	247	256	-	GATA-3	1	0.946	ecr11	489	498	+	GATA-3	0.945	0.91
ecr11	18	27	+	GATA-3	0.968	0.912	ecr11	520	533	+	C/EBPbeta	0.865	0.78
ecr11	25	41	+	HNF-1	0.92	0.657	ecr11	521	537	-	HNF-1	0.874	0.71
ecr11	31	47	-	HNF-1	1	0.647	ecr11	545	561	-	HNF-1	0.835	0.732
ecr11	44	53	+	USF	0.918	0.865	ecr11	549	562	-	C/EBPbeta	0.883	0.829
ecr11	51	64	+	C/EBPbeta	0.854	0.811	ecr11	567	580	+	C/EBPbeta	0.833	0.788
ecr11	62	71	+	TATA	1	0.952	ecr11	571	585	+	HNF-3beta	1	0.907
ecr11	64	78	-	HNF-3beta	1	0.85	ecr11	573	585	-	CHOP - C/EBPalpha	0.882	0.81
ecr11	71	87	-	HNF-1	0.771	0.706	ecr11	603	619	+	HNF-1	0.796	0.657
ecr11	74	88	+	HNF-3beta	1	0.832	ltv1	6	16	+	AP-1	0.935	0.892
ecr11	87	100	-	C/EBPbeta	0.816	0.84	ltv1	8	21	+	HNF-4	1	0.873
ecr11	87	101	-	HNF-3beta	1	0.836	ltv1	13	23	+	AP-1	0.935	0.847
ecr11	87	104	-	NF-1	1	0.961	ltv1	18	35	+	NF-1	0.921	0.88
ecr11	87	96	-	TATA	1	0.9	ltv1	36	49	+	C/EBPbeta	0.848	0.789
ecr11	98	107	-	USF	0.945	0.931	ltv1	43	58	+	GR	0.978	0.85
ecr11	102	118	+	HNF-1	0.829	0.818	ltv1	86	103	+	NF-1	1	0.952
ecr11	102	112	+	AP-1	0.811	0.825	ltv1	91	100	-	USF	0.931	0.911
ecr11	106	116	+	AP-1	1	0.969	ltv1	101	115	-	HNF-3beta	1	0.879
ecr11	108	124	-	HNF-1	1	0.78	ltv1	126	139	-	C/EBPbeta	0.828	0.804
ecr11	116	129	-	C/EBPbeta	0.842	0.859	ltv1	127	137	-	AP-1	0.935	0.889

ecr11	120	134	-	HNF-3beta	0.93	0.868	ltv1	159	176	-	NF-1	0.921	0.88
ecr11	144	153	+	GATA-3	0.981	0.908	ltv1	163	176	+	C/EBPbeta	0.888	0.837
ecr11	146	155	-	GATA-3	0.981	0.908	ltv1	199	209	+	AP-1	0.935	0.887
ecr11	147	163	-	HNF-1	0.795	0.682	ltv1	202	219	-	NF-1	0.905	0.864
ecr11	167	182	-	GR	1	0.886	ltv1	213	222	+	USF	0.987	0.857
ecr11	175	189	-	HNF-3beta	1	0.926	ltv1	223	232	-	USF	0.926	0.883
ecr11	187	204	+	NF-1	0.921	0.875	ltv1	243	254	-	CREB	1	0.975
ecr11	191	207	-	HNF-1	0.794	0.672	ltv1	244	254	-	AP-1	0.935	0.868
ecr11	196	205	-	GATA-3	0.896	0.881	ltv1	248	265	-	NF-1	0.921	0.898
ecr11	198	211	-	C/EBPbeta	0.996	0.885	ltv1	285	298	-	C/EBPbeta	0.854	0.781

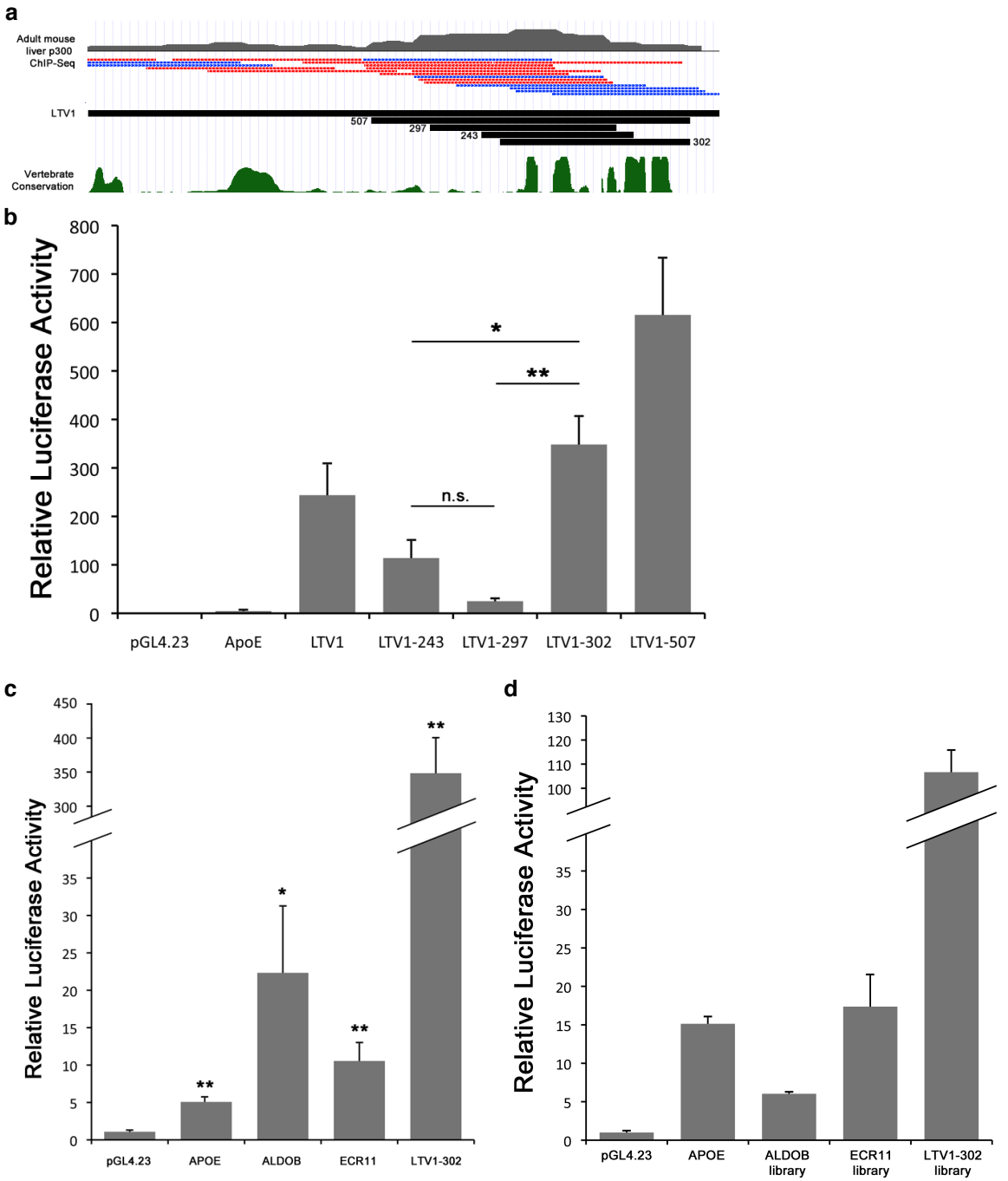
We used the MATCH web server[131] to predict transcription factor binding sites (TFBS) in the three enhancers under study using liver-specific profiles and cutoff selection set to minimize false negatives.

Table B.3. Characteristics of interacting positions from pairwise multiple regression models.

		ALDOB		ECR11		LTV1	
		<=10 nt	>10 nt	<=10 nt	>10 nt	<=10 nt	>10 nt
Not significant		2509	30798	5706	178301	2787	41143
Significant		22	60	17	182	28	17
p-value		< 1e-4		< 1e-3		< 1e-4	
Univar. model coeff. signs	**Interaction term sign**	**ALDOB**		**ECR11**		**LTV1**	
		Not significant	**Significant**	**Not significant**	**Significant**	**Not significant**	**Significant**
-/-	-	7378	0	4195	0	9690	0
-/-	+		2		4		18
+/-	-	4471	36	5387	20	8315	4
+/-	+		6		1		4
+/+	-	654	0	1682	1	1760	5
+/+	+		12		23		3

For pairs of positions that were mutated together in at least 20 haplotypes, we built multiple linear regression models with three binary variables to predict the number of RNA aliquots in which a haplotype was observed. Two variables encoded whether each position was mutant or wild-type in a given haplotype and the third encoded whether both were mutant together. We then compared whether pairs of positions with significant interaction terms (FDR<0.05) were enriched for nearby pairs (separated by ≤10 nt) compared to those with non-significant interaction terms (p-value obtained by comparing the number of nearby pairs with significant interaction terms to the null distribution of this quantity, obtained by randomly permuting the position vector 10,000 times and each time computing the number of nearby pairs with significant interaction terms). We also classified models on the basis of the sign (positive or negative) of the coefficient from the univariate position-by-position models ("Univar. model coeff. signs"), the interaction term sign, and whether or not the interaction term was significant in the pairwise model (note that non-significant interactions terms cannot be distinguished from zero and therefore do not have a sign).

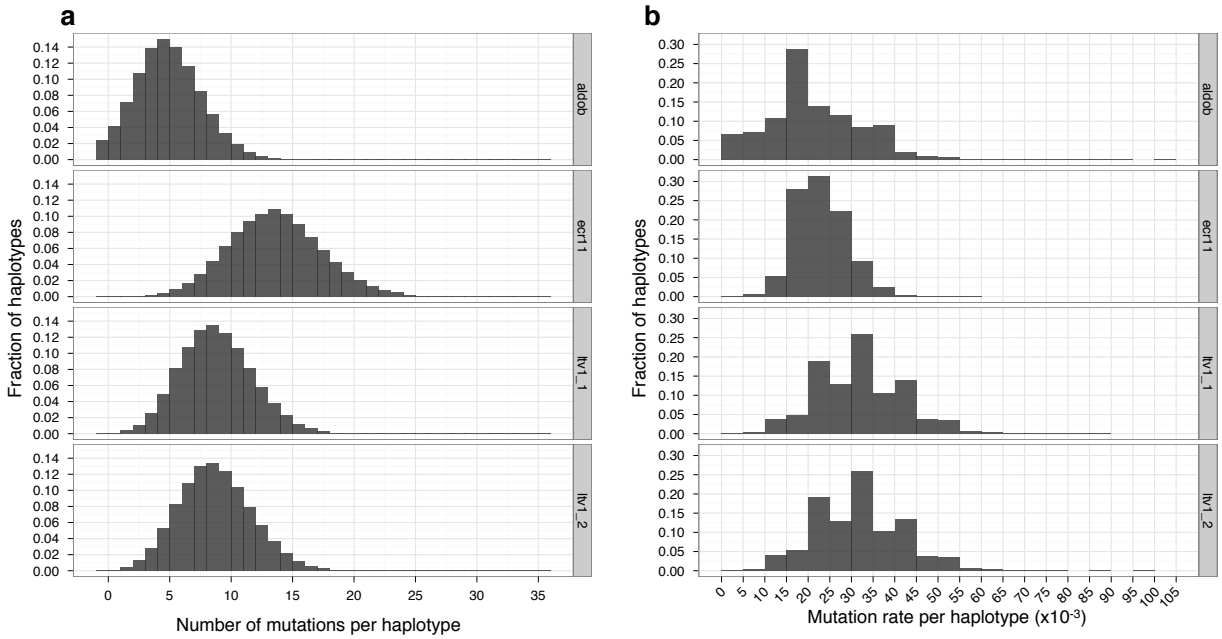
Figure B.1. Activity of wild-type enhancers and variant pools by tail vein luciferase assay.



(a) Identification of LTV1 based on p300 ChIP-Seq from early adult mouse liver and position of deletion fragments constructed to refine enhancer position. The labels indicate the name and size (bp) of each fragment (b) Relative luciferase activity driven by the various LTV1 fragments compared to the APOE liver enhancer and minimal promoter only (pGL4.23). *:p<0.05, **:p<0.01, One way analysis of variance (ANOVA) with Tukey post-hoc test to compare groups. (c) Relative luciferase activity driven by the three wild-type enhancers used in this study compared to the APOE liver enhancer and minimal promoter only (pGL4.23). *:p<0.05, **:p<0.01, Student's unpaired two tailed t-test (d) Aggregate relative luciferase

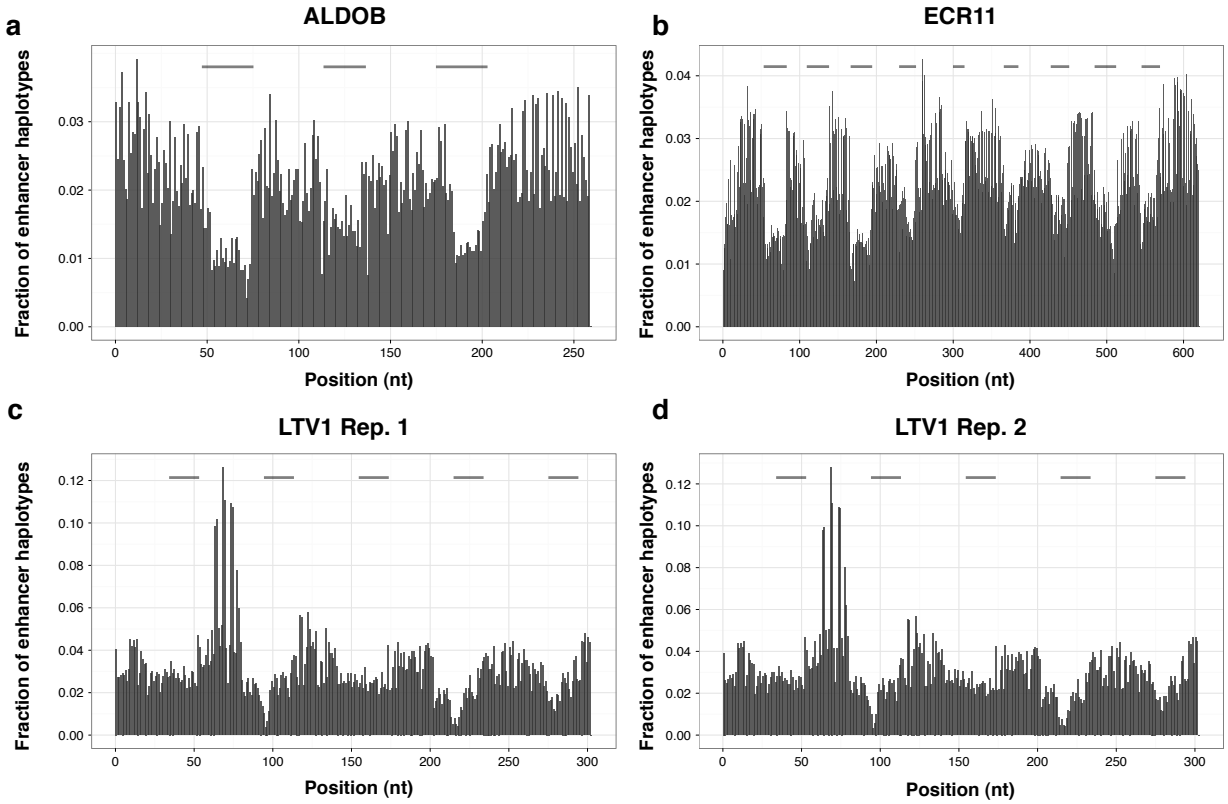
activity driven by a pool of all the enhancer haplotypes for each of the three enhancers under study and compared to the ApoE liver enhancer and minimal promoter only (pGL4.23).

Figure B.2. Distribution of mutations per enhancer haplotype.



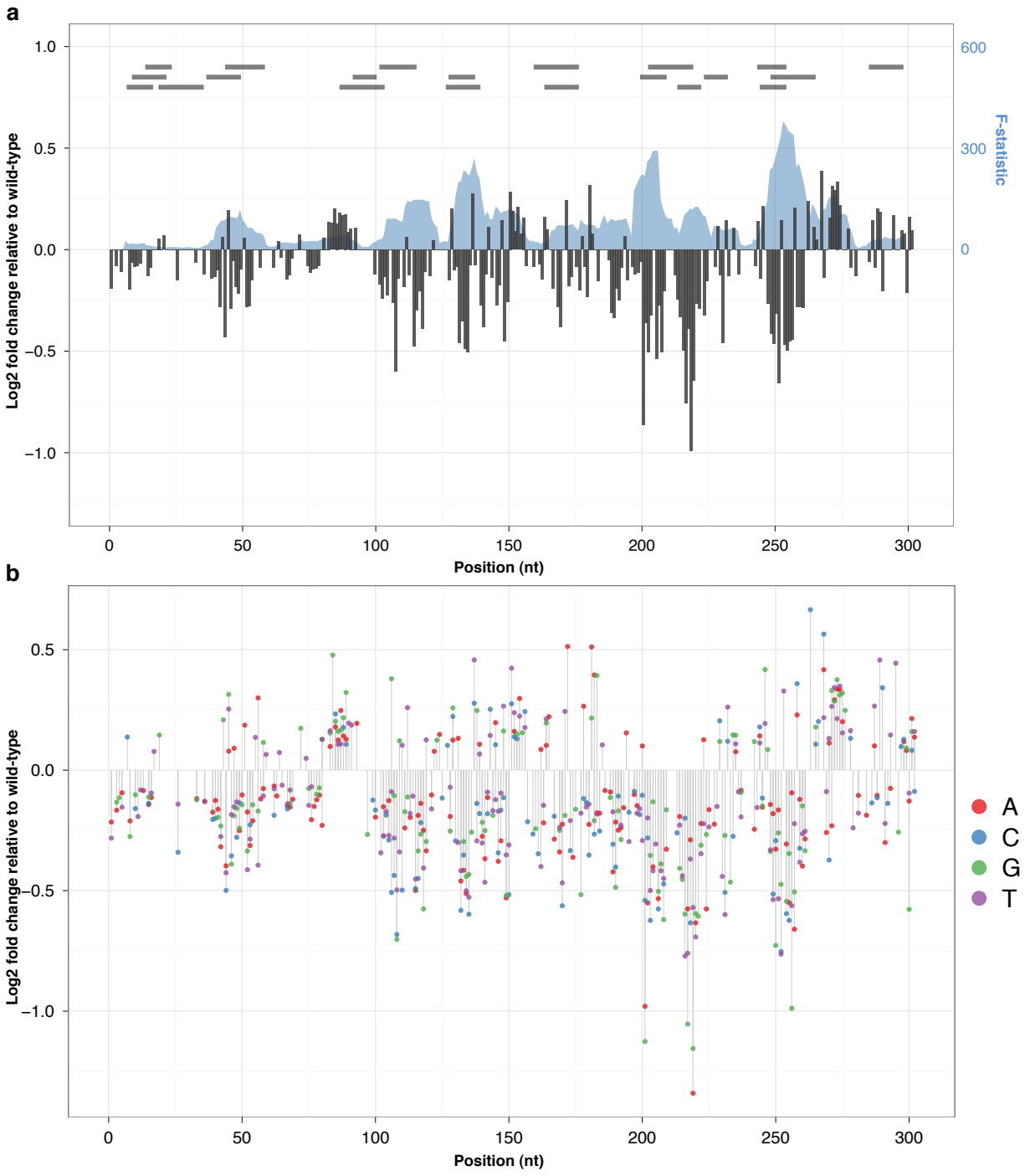
The fraction of enhancer haplotypes containing a given number of mutations (a) and the fraction of enhancer haplotypes with a given per-base mutation rate (b).

Figure B.3. Distribution of mutations by position in enhancer haplotypes



Per-base mutation rate as a function of position in the enhancer for ALDOB (a), ECR11 (b), and the two replicates of LTV1 (c, d). As would be expected, dips in mutation rate correspond to overlap regions during the PCA process (horizontal gray bars). Nonetheless, all possible substitution mutations were observed in at least 42 distinct enhancer haplotypes and all pairs of positions were disrupted together in at least one haplotype with the exception of a single pair of positions in LTV1.

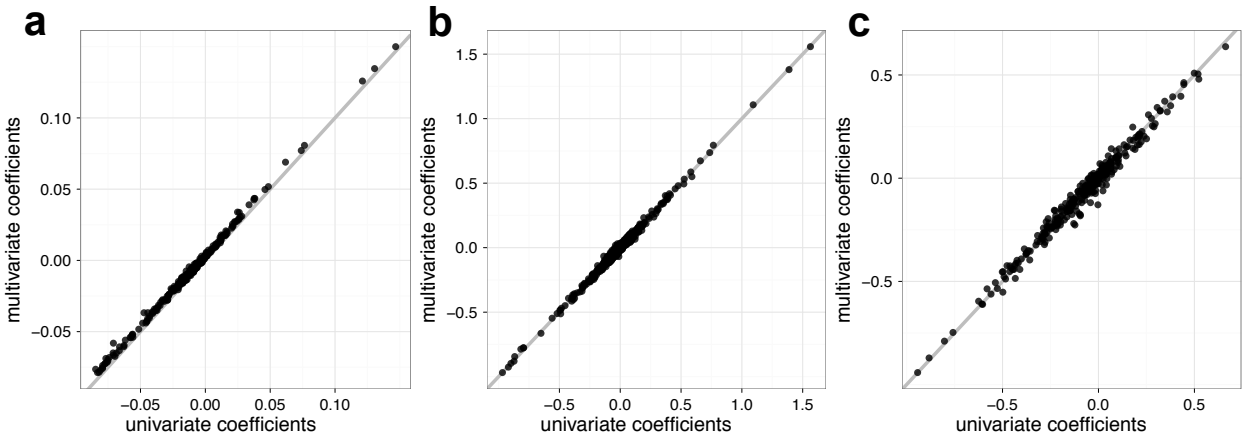
Figure B.4. Mutation effect size in the second replicate of LTV1.



Position-specific mutation effect sizes based on coefficients from univariate (grey columns, left axis) (a) and trivariate models (A:red, C:blue, G:green, T:purple) (b) are plotted here. Effect sizes were calculated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide. Effect sizes are only shown for positions where model coefficients had associated p-values less than or equal to 0.01. Multiple linear regression was used to predict the number of aliquots in which a given enhancer haplotype was observed, using sets of 10

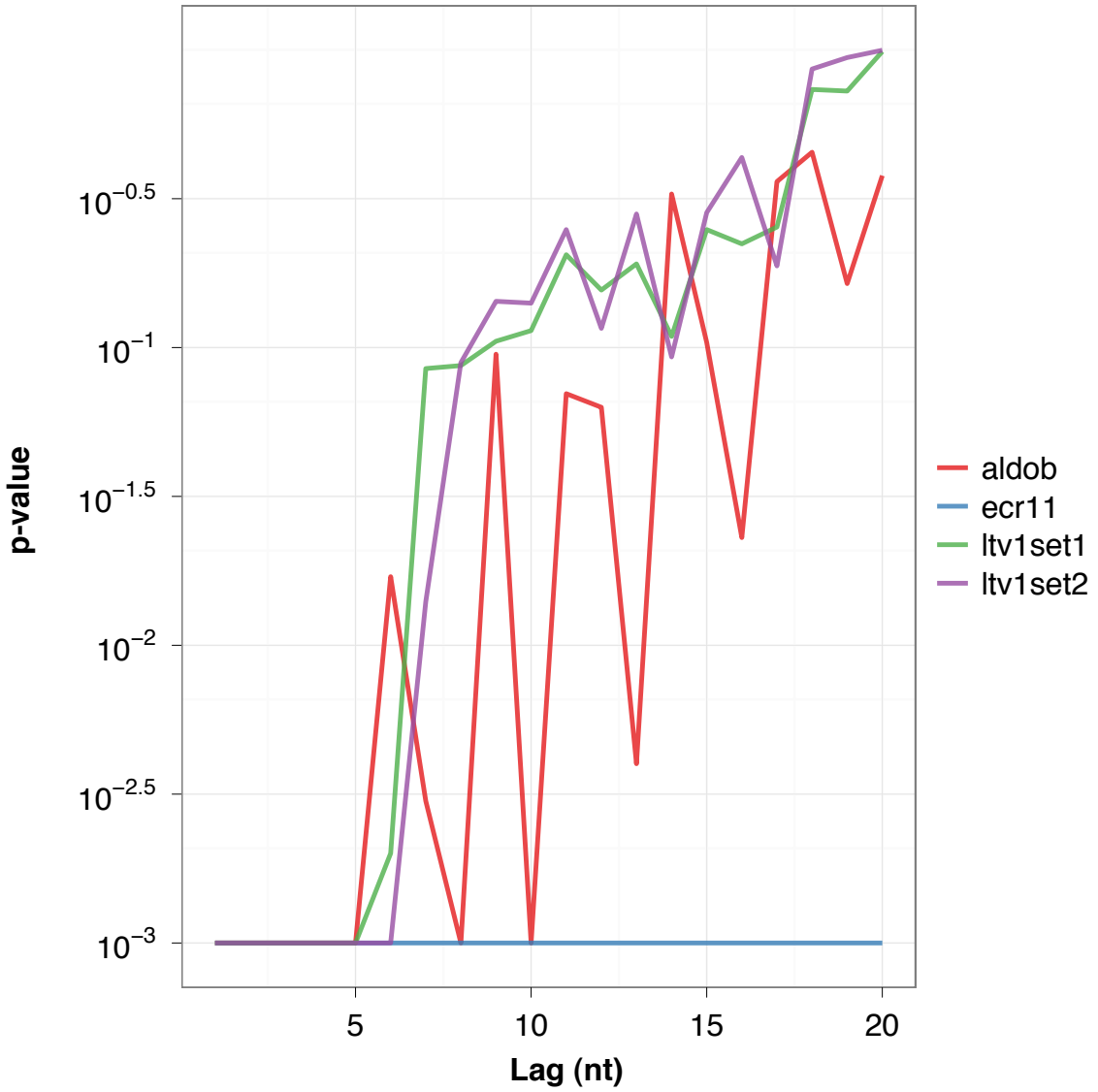
adjacent positions (coded as binary vectors based on whether a mutation was present in each enhancer haplotype) as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (blue shadow, right axis) (a). The locations of TFBS predictions using the MATCH web server are shown as horizontal grey bars at the top of the plot.

Figure B.5. Comparison between univariate and multivariate linear regression coefficients.



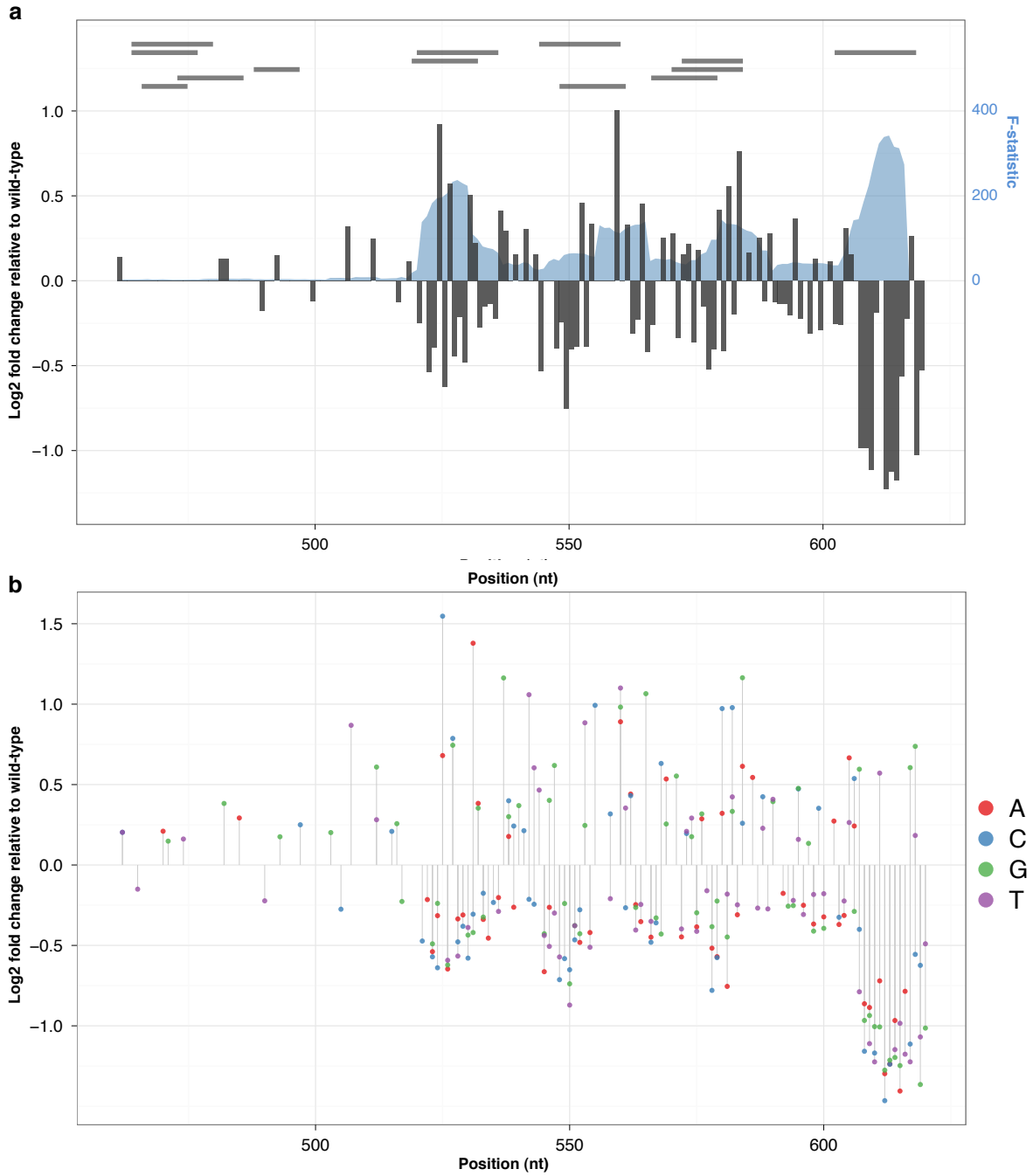
Coefficients calculated via univariate linear regression (i.e. only considering mutational status at a single position) are plotted against coefficients calculated via multivariate linear regression (simultaneously considering mutational status at all sites in the enhancer) for ALODB (a), ECR11 (b), and LTV1 (c). The line y=x is shown in gray in all three plots.

Figure B.6. P-value for the similarity of effect sizes of nearby positions.



To assess the similarity of the effect sizes of mutations at nearby positions in each enhancer, we summed the absolute difference between effect sizes at all positions separated by a fixed "lag" distance. We then recalculated this quantity 1000 times after randomly permuting the effect sizes. We obtained a p-value by calculating the fraction of times that the quantity computed on the permuted effect sizes was at least as small as the quantity computed on the real data. This was repeated for a range of values of the lag distance. The p-value is plotted here as a function of the lag distance. Positions separated by ~5 nucleotides or fewer show substantially similar effect sizes ($p<0.01$) across all three enhancers assayed.

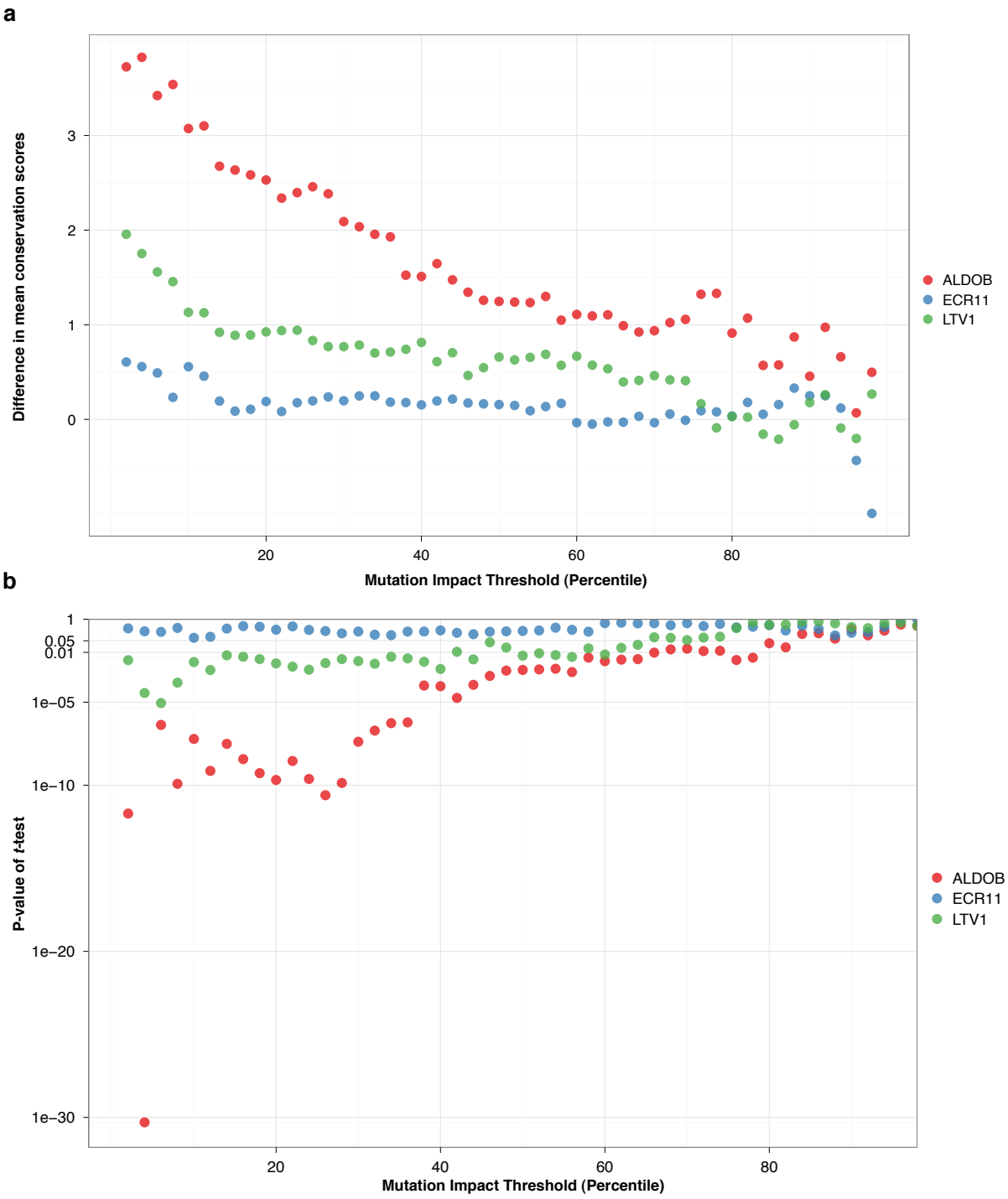
Figure B.7. Mutation effect size in the distal 160 nt of ECR11.



Position-specific mutation effect sizes based on coefficients from univariate (grey columns, left axis) (a) and trivariate models (A:red, C:blue, G:green, T:purple) (b) are plotted here. Effect sizes were calculated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide. Effect sizes are only shown for positions where model coefficients had associated p-values less than or equal to 0.01. Multiple linear regression was used to predict the number of aliquots in which a given enhancer haplotype was observed, using sets of 10 adjacent positions (coded as binary vectors based on whether a mutation was present in each enhancer haplotype) as predictors. The F-statistic of these models, representing the extent to which the model is

predictive of the outcome, is plotted (blue shadow, right axis) (a). The locations of TFBS predictions using the MATCH web server are shown as horizontal grey bars at the top of the plot.

Figure B.8. Single nucleotide relationships between evolutionary and functional constraint.



Positions were rank-ordered based on the absolute vale of their effect size (from position-based, i.e. univariate, linear models) and the difference in the mean conservation score for the top *x* percent of positions versus the mean conservation score for the bottom 100-*x* percent of positions is shown (a). For example, a cutoff at the tenth percentile separates the highest impact ten percent of positions from the lowest impact ninety percent. A *t*-test was then performed to compare the means of the two distributions of conservation scores for a given impact threshold cutoff and the p-values associated with each test are

shown in (b). The highest impact mutations tend to be significantly more conserved than the remainder of positions for ALDOB and ECR11.

Methods

Oligonucleotides used for PCA	
ALDOB_PCA_OLIGO1	AGGACCGGATCAACTTCTTCA
ALDOB_PCA_OLIGO2	TCCCTGTAAACAGTATTAGTTTGAATTATCATTTTCACTGTTATTCTGGTTGAGTCAGCATACCCAGA TTGAAGAAGTTGATCCGGTCCT
ALDOB_PCA_OLIGO3	ATAATTCAAACTAATACTGTTTACAGGGAGTTAAACTTCTACAGTGGGATTAAAGGTCTGTACCACG TTAGCACAAATGTCACCTCTCTG
ALDOB_PCA_OLIGO4	CCATCCCAGGTTGTCCTCCTGTCTCCTTGTGGTGAACATTGGCCTGTGACCCTGTTTTATGATTAA CAGAGAGGTGACATTTGTGCTAAC
ALDOB_PCA_OLIGO5	GGAGGACAACCTGGGATGGGTAATGACAAAGAACGATTTCCGTACTCCTAAGCCTCTGCTCTCTC AGATCTCAAGCCATTGCGTGAACCG
ALDOB_PCA_OLIGO6	TCGGTTCACGCAATGGCTTG
ECR11_PCA_OLIGO1	AGGACCGGATCAACTCTCTGAAGCTCAAAAGCAATG
ECR11_PCA_OLIGO2	AAACATTTAGTATTTTTAAAGGTGTTGGAATTCAAGTGTTAAAAATCGAAGCCTTATCAAATCATTGC TTTTGAGCTTCAGAGAGT
ECR11_PCA_OLIGO3	ATTCCAACACCTTTAAAAATACTAAATGTTTCCCATTTTAAACAAGCCAAGTGAATGACTGAATTCTT AACCAAAAATAAATGTGA
ECR11_PCA_OLIGO4	GGCCAGAGAATATTTATATAATGTTCTGTATGGACAAAGAGTGATATCAATCTACTTCACATTTATTT TTGGTTAAGAATTCAGTC
ECR11_PCA_OLIGO5	ACAGAACATTATATAAATATTCTCTGGCCTTACTATCTAGCAAGGCAGGAAAAATAGATCAATTTGT TCTCACTCATAGGTGGGAA
ECR11_PCA_OLIGO6	CCCCACAACAGGCCCCGATGTGTGATGTTCCCCTTCCTGTGTCCATGTGTTCTCATTGTTCAATTC CCACCTATGAGTGAGAACAA
ECR11_PCA_OLIGO7	GGGGCCTGTTGTGGGGTGGGGGGAGGGGGGAGGGATAGCATTAGGAGATATATCTAACGTTAAA TGACGTGTTAATGGGAGCAGCA
ECR11_PCA_OLIGO8	TAAGTTTTAGGGTACATGTGCACAACATGCAGTTTGTTACATATGTATACATGTGCCATGTTGGTGT GCTGCTCCCATTAACACGT
ECR11_PCA_OLIGO9	TTGTGCACATGTACCCTAAAACTTAAAGTATAATAAGAAAAATAGATCAATTTACTCTACATCTGAGA TTAAAAAGCAGAAAGACT
ECR11_PCA_OLIGO10	TTCTCGCTGTTACTCTATTTCTGGTTCTGAATGTCAAATACTGAAACTCTGTGAGTGAGTCTTTCTG CTTTTTAATCTCAGATGTA
ECR11_PCA_OLIGO11	ACCAGAAATAGAGTAACAGCGAGAACTTGAACTATTTCAGTTTAGCCTCCCACCCTCTCTGCTATC ACTTCCCAAAACATTGCGTG
ECR11_PCA_OLIGO12	TCGGTTCACGCAATGTTTTGGGAAGTG
LTV1_PCA1_OLIGO1	ATCACAAGTTTGTACAAAAAAGCAGGCTCCGCGGCCGCCCCCTTCACCTTTGGGTGACCCCTGAC CCTGGCCGCCTGGGCTC
LTV1_PCA1_OLIGO2	ACAGGGCCAAGGAAGGAGGGCGGGGTGGGGCGGGGCGGCGAGGACGGAATGTGCGGGAAGGC GAGCCCAGGCGGCCAGGGTC
LTV1_PCA1_OLIGO3	CCCTCCTTCCTTGGCCCTGTGGGGACGGAAACATCCCGTTCCTGCCCAAGCTGGGTCAAGAGCC GGAGGGACAGGACCAGAG
LTV1_PCA1_OLIGO4	AGGCGTGGCGAGATGAGGTCACCCAGTAGGAACAAGGAGAGCTAGTTCTGGCGTAAGGGGTGCT CTGGTCCTGTCCCTCCGG
LTV1_PCA1_OLIGO5	GACCTCATCTCGCCACGCCTCCTCAGGTGAACACCCGGGCTGGTAACGTCACTTCCTGCCAGGTA AGCGCCCCCAGGCAGCA
LTV1_PCA1_OLIGO6	ATCACCACTTTGTACAAGAAAGCTGGGTCGGCGCGCCCACCCTTCAGACCTTTCCGTGAGCAGTG CTGCCTGGGGGCGCTTAC
LTV1_PCA2_OLIGO1	AGCAGGCTCCGCGGCCGCCCCCTTCACCTTTGGGTGACCCCTGACCCTGGCCGCCTGGGCTCGC CTTCCCGCACATTCCG
LTV1_PCA2_OLIGO2	GGATGTTTCCGTCCCCACAGGGCCAAGGAAGGAGGGCGGGGTGGGGCGGGGCGGCGAGGACG GAATGTGCGGGAAGGCGA
LTV1_PCA2_OLIGO3	CTGTGGGGACGGAAACATCCCGTTCCTGCCCAAGCTGGGTCAAGAGCCGGAGGGACAGGACCAG AGCACCCCTTACGCCA
LTV1_PCA2_OLIGO4	GTTCACCTGAGGAGGCGTGGCGAGATGAGGTCACCCAGTAGGAACAAGGAGAGCTAGTTCTGGC GTAAGGGGTGCTCTGG
LTV1_PCA2_OLIGO5	CCACGCCTCCTCAGGTGAACACCCGGGCTGGTAACGTCACTTCCTGCCAGGTAAGCGCCCCCAG

	GCAGCACTGCTCACGG
LTV1_PCA2_OLIGO6	ATCACCACTTTGTACAAGAAAGCTGGGTCGGCGCGCCCACCCTTCAGACCTTTCCGTGAGCAGTG CTGCCTGG
LTV1_PCA1_P1	GCTAGCCTCGAGGATATCACAAGTTTGTACAAAAAAGCAGGCTCCG
LTV1_PCA1_P2	ACGGGCCAAGGAAGGAGGGC
LTV1_PCA1_P3	CCCTCCTTCCTTGGCCCTGTGG
LTV1_PCA1_P4	AGGCGTGGCGTGATGAGGTCAC
LTV1_PCA1_P5	GACCTCATCTCGCCACGCCTCC
LTV1_PCA1_P6	AGGCCAGATCTTGATATCACCACTTTGTACAAGAAAGCTGGGTCG
LTV1_PCA2_P1	GCTAGCCTCGAGGATATCACAAGTTTGTACAAAAAAGCAGGCTCCGCGGCC
LTV1_PCA2_P2	GGATGTTTCCGTCCCCACAGGG
LTV1_PCA2_P3	CTGTGGGGACGGAAACATCCCG
LTV1_PCA2_P4	GTTCACCTGAGGAGGCGTGGCG
LTV1_PCA2_P5	CCACGCCTCCTCAGGTGAACAC
LTV1_PCA2_P6	AGGCCAGATCTTGATATCACCACTTTGTACAAGAAAGCTGGGTC
LTV1_OUTER_F	GCTAGCCTCGAGGAT
LTV1_OUTER_R	AGGCCAGATCTTGAT
Oligonucleotides used for cloning, tagging, and sequencing	
VH_F	GCTAGCCTCGAGGATTGCCTAGGACCGGATCAACT
VH_R	AGGCCAGATCTTGATGAGCTTCGGTTCACGCAATG
ENHANCER_FWD	AATGATACGGCGACCACCGAGATCTACACTGCCTAGGACCGGATCAACT
LTV1_F	GGGAGGTATTGGACAGGCCGC
Nextera Adapter1	AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAG
Nextera BP1	AATGATACGGCGACCACCGA
BARCODE_PE_F	AATGATACGGCGACCACCGAGATCTACACAGTCGCCTATACGGTGATGG
BARCODE_PE_R	CAAGCAGAAGACGGCATACGAGATATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX1	CAAGCAGAAGACGGCATACGAGATCGTGATTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX2	CAAGCAGAAGACGGCATACGAGATGCCTAATCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX3	CAAGCAGAAGACGGCATACGAGATCACTGTTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX4	CAAGCAGAAGACGGCATACGAGATATTGGCTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX5	CAAGCAGAAGACGGCATACGAGATTCAAGTTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX6	CAAGCAGAAGACGGCATACGAGATCTGATCTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX7	CAAGCAGAAGACGGCATACGAGATAAGCTATCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX8	CAAGCAGAAGACGGCATACGAGATGTAGCCTCGACTCTAGATGGGATTAAACGGGGAGAC
BARCODE_SEQ_F	GACCACCGAGATCTACACAGTCGCCTATACGGTGATGG
TAG_OLIGO	GTGTAATAATTCTAGAAGCTTAGTCGCCTATACGGTGATGGNNNNNNNNNNNNNNNNNNNN GTCTCCCCGTTTAATCCCATCTAGAGTCGGGGCGG
TAG_EXTEND	CCGCCCCGACTCTAGATG

Details of the reactions for the assembly of LTV1 haplotypes using PCA

The two LTV1 designs were assembled separately. For each design, pairs of oligos i.e. oligos 1 and 2, oligos 3 and 4, and oligos 5 and 6 were each assembled in parallel and the products of the three reactions were then assembled together into the final product in a single reaction.

	Templates	**Primers**
Step 1, Reaction 1	LTV1_PCA[1/2]_OLIGO1,	LTV1_PCA[1/2]_P1,

	LTV1_PCA[1/2]_OLIGO2	LTV1_PCA[1/2]_P2
Step 1, Reaction 2	LTV1_PCA[1/2]_OLIGO3, LTV1_PCA[1/2]_OLIGO4	LTV1_PCA[1/2]_P3, LTV1_PCA[1/2]_P4
Step 1, Reaction 3	LTV1_PCA[1/2]_OLIGO5, LTV1_PCA[1/2]_OLIGO6	LTV1_PCA[1/2]_P5, LTV1_PCA[1/2]_P6
Step 2	Products of reactions 1, 2, and 3	LTV1_OUTER_F, LTV1_OUTER_R

Construction of enhancer haplotypes from short, doped oligonucleotides using PCA

Sets of overlapping oligonucleotides for each enhancer were designed either by manual inspection (LTV1) or using the program DNAWorks (ALDOB and ECR11). Common flanking sequences were included on either side to allow for amplification of the full-length enhancer haplotypes during PCA. For LTV1, two versions of overlapping oligonucleotides were designed, such that the overlap region in each was different. Oligonucleotides were synthesized by Integrated DNA Technologies (IDT). All positions corresponding to the enhancer region were synthesized using a hand-mix doped at a ratio of 97:1:1:1 (that is, designated base at a frequency of 97%, and every other base at a frequency of 1%). Sequences of all oligonucleotides are listed in Tables B.2 and B.3.

For ALDOB as well as ECR11, the full-length haplotypes were assembled in a single step. We used 50 fmol of each oligonucleotide (ALDOB_PCA_OLIGO[1...6] or ECR11_PCA_OLIGO[1...12]) in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), and 0.5× SYBR Green II, with the following cycling conditions: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Four such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product representing a complex pool of enhancer haplotypes was purified using QIAquick columns (Qiagen). The assembled enhancer haplotypes were then subjected to an additional around of PCR to add 15 bp of vector homology on either side to render them competent for cloning using InFusion (Clontech). We used 20 ng of template in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix, 0.5× SYBR Green II, and each primer (VH_F and VH_R) at 0.3 µM final concentration. Thermal cycling was done with the following program: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Sixteen such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product was purified using QIAquick columns (Qiagen).

The two LTV1 designs were assembled separately. For each design, pairs of oligonucleotides, that is, oligonucleotides 1 and 2, oligonucleotides 3 and 4, and oligonucleotides 5 and 6, were each assembled in parallel and the products of the three reactions were then assembled together into the final product in a single reaction. The combinations of primers and oligonucleotides used in each reaction are listed below. Each 50 µl PCR reaction was prepared on ice with 1× iProof Ready Mix (Bio-Rad), 0.5× SYBR Green II, forward and reverse primers each at 0.5 µM final concentration and 50 fmol of each template oligo. Thermal cycling was done in a MiniOpticon Real-time PCR system (Bio-Rad) with the following program: 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 62 °C for 30 s and 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. PCR products were purified on a QIAquick column (Qiagen). The haplotypes obtained from each of the two LTV1 designs were pooled after the PCA step. Two aliquots were drawn from this pool, and then carried through subsequent steps as two independent samples and were associated with entirely different sets of tags.

Cloning of enhancer haplotypes and the degenerate tag into pGL4.23 plasmid

For ALDOB and ECR11, we first cloned in the degenerate tag to create a complex library of tagged pGL4.23 plasmids. We then cloned in the enhancer haplotypes into these tagged pGL4.23 plasmids. For LTV1, we first cloned in the enhancer haplotypes and then cloned in the degenerate tag. Details of each cloning step remained the same, irrespective of the order in which they were carried out, and are described below.

Cloning of degenerate tag into pGL4.23 plasmid

The tag oligonucleotide (TAG_OLIGO) was made double-stranded using primer extension in a 50 µl reaction volume with 1× iProof Master Mix, 0.5 µg single-stranded tag oligo, 0.5 µg reverse primer (TAG_EXTEND). The reaction was incubated at 95 °C for 3 min, 61 °C for 10 min and then 72 °C for 5 min. The product was purified using a QIAquick column and eluted in 50 µl EB. It was further subjected to ExoI treatment in 40 µl reaction volume for 1 h at 37 °C to degrade any remaining single-stranded DNA, and purified again using QIAquick columns. The resulting double-stranded tag oligo was then cloned into pGL4.23 at the XbaI site (at 1,799 bp) using standard InFusion (Clontech) protocol. The InFusion reaction was diluted to 100 µl using TE8. We used 1.5 µl of this diluted cloning reaction to transform 50 µl of chemically competent FusionBlue cells (Clontech) using the standard protocol. When the tag was being cloned in first, 16 such transformation reactions were pooled and grown overnight in four 50-ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Cloning enhancer haplotypes into pGL4.23 vector

The enhancer haplotypes were cloned into the EcoRV site (at 42 bp) of the pGL4.23 plasmid, using standard InFusion protocol. We used 1.5 µl of the cloning reaction to transform 50 µl of chemically competent FusionBlue cells using standard protocol. Five transformations reactions were pooled and grown overnight in 50 ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Tail vein injections

Enhancers were injected using methods as previously described [127]. Briefly, each library was injected into mice using the TransIT EE Hydrodynamic Gene Delivery System (Mirus Bio) following the manufacturer's protocol. We injected 10 µg of each library, alongside 2 µg of pGL4.74[hRluc/TK] vector to correct for injection efficiency, into the tail vein of CD1 mice (Charles River). After 24 h, mice were euthanized and livers were harvested.

Measurement of luciferase activity

Firefly and renilla luciferase activity were measured on a Synergy 2 Microplate Reader (BioTek Instruments) for each liver using the Dual Luciferase Reporter Assay System (Promega). The firefly luciferase to renilla luciferase ratios were determined and expressed as relative luciferase activity. All mouse work was approved by the UCSF Institutional Animal Care and Use Committee.

Isolation of RNA from mouse livers

Fresh liver tissue was immediately stabilized in RNAlater solution (Ambion). Samples were homogenized in TRIzol reagent (Invitrogen) and RNA was isolated from the samples according to the manufacturer's instructions.

DNase treatment of RNA

To remove any DNA contamination in the RNA extracted from mouse livers, it was subjected to DNaseI treatment using DNA-free (Ambion). Each reaction was prepared with 1× DNA-free buffer, 1 µl of rDNaseI enzyme, 10 µg of RNA and RNase-free water to 50 µl. The reactions were incubated at 37 °C for 1 h, with an additional 1 µl of enzyme added mid-way through the incubation. The reaction was stopped by adding 7 µl of the inactivation reagent and incubating for 2 min at 25 °C with frequent shaking. The reaction was centrifuged in a microcentrifuge at 10,000g for 1.5 min, and the supernatant containing RNA was carefully transferred to a fresh tube.

RT-PCR

Aliquots of RNA obtained after DNase treatment were reverse transcribed to cDNA and amplified by PCR using the Qiagen One-Step Kit. The PCR sought to amplify the 20-bp degenerate tag encoded at the 3′ end of the luciferase transcript. The reactions were assembled on ice in a 25 µl total volume with the following reagents: 1× Qiagen One-Step RT-PCR buffer, 400 µM of each dNTP, 0.6 µM of forward primer (BARCODE_PE_F), 0.6 µM of relevant reverse primer (BARCODE_PE_R_ILMN_INDEX[1-8]), 0.5× SYBR Green II and 5 µl (~1 µg) of RNA template. Thermal cycling was done on a Bio-Rad MiniOpticon Real-Time PCR system with the following program: 50 °C for 30 min (reverse transcription), 95 °C for 15 min (inactivation of reverse transcriptase and heat-activation of the DNA polymerase), then 30 cycles of 94 °C for 30 s, 65 °C for 30 s and 72 °C for 30 s. Each reaction was monitored and extracted from the PCR machine when the fluorescence began to plateau. The cDNA products were purified using the QIAquick PCR Purification Kit (Qiagen) and eluted in 35 µl EB. The primers used for the RT-PCR contained the necessary sequences for compatibility with the Illumina flow-cell. Thus, the cDNA library obtained at the end of this step was ready for sequencing, eliminating the need for a separate sequencing-library construction step. The reverse primer additionally included 6 bp barcodes allowing for several RT-PCR reactions to be pooled into a single lane for sequencing.

Sequencing of RNA-derived tags

The pooled RT-PCR reaction products were sequenced on an Illumina GAIIx using a sequencing primer (BARCODE_SEQ_F) designed to read into the tag sequence. Each run was 36 cycles with an additional 6 cycles to read the indexing barcode using the index sequencing primer (BARCODE_SEQ_INDEX).

For each aliquot, reads were filtered based on the quality scores for the first 20 bases, which correspond to the degenerate tag. The numbers of occurrences of each tag were counted and tags that were supported by at least ten reads were classified as being 'present' in that aliquot.

Associating tags with enhancer haplotypes

The enhancer haplotypes and tags were situated more than 1,000 bp away from each other on the pGL4.23 plasmid. To bring them adjacent and facilitate the subassembly method, we digested the pGL4.23 plasmids using HindIII, which had two cut sites, one just 3′ of the enhancer, and one just 5′ of the tag, thus resulting in excision of the intervening region. Cut site 1 was already a part of the pGL4.23 backbone. Cut site 2 was engineered in as a part of the tag oligo. The digest was carried out in a 50 µl volume with 1× NEB Buffer 2, 1 µg of plasmid and 1 µl of HindIII Enzyme (New England BioLabs) and incubated at 37 °C for 3 h. The digested plasmid was purified using a QIAquick column.

The digested plasmids were then recircularized using intramolecular ligation, resulting in the tag becoming adjacent to the 3′ end of the enhancer. Ligation was performed using T4 DNA ligase (New England BioLabs) in a 20 µl reaction with 15 ng of template per reaction. The reaction was incubated for 15 min at 25 °C, followed 20 min at 65 °C to inactivate the ligase.

The enhancer and tag region were amplified from recircularized plasmids using PCR with the forward primer targeting the region immediately 5′ of the enhancer (ENHANCER_F for ALDOB and ECR11, and LTV1_F for LTV1) and the reverse primer targeting the region immediately 3′ of the tag (BARCODE_PE_R). The reaction was carried out in a 25 µl volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, and each primer at 0.3 µM final concentration. Thermal cycling was done using Bio-Rad MiniOpticon Real-Time PCR system using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The reactions were then pooled and purified using QIAquick columns.

The amplicons were then subjected to the subassembly protocol as conceptually described in [128] with some modifications as follows. The random fragmentation step was carried out using the Nextera Tn5 transposase (EpiCentre) instead of mechanical shearing. The Nextera reaction was purified using MinElute column (Qiagen) and size-selected by PAGE (LTV1: 100+; ECR11:100-300,300+; ALDOB: no size-selection performed). The size-selected fragments were subjected to PCR in a 25 µl reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, Nextera Adaptor 1 at 10 nM final concentration, and primers Nextera BP1 and BARCODE_PE_R at 0.3 µM final concentration each. Thermal cycling was carried out using BioRad Mini Opticon System using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The PCR products were purified using a QIAquick column and then sequenced on either an Illumina GAIIx or a Hi-Seq 2000. Read1 collected 76 bp/101 bp of the enhancer sequence staring at random breakpoints along the enhancer. Read 2 collected the 20-bp tag sequence.

The reads were then grouped by tag. Reads belonging to each group were then aligned to the wild-type enhancer sequence to identify the mutations on the haplotype associated with that tag using a custom analysis framework.

Estimation of effect size of mutation at each position along the enhancer (univariate model)

All linear regression analyses were done using the lm() or lsfit() functions available in the R Statistical Package. To quantity the effect of mutation at any given position on the number of aliquots in which an enhancer haplotype was observed, we built a separate linear regression model at every position along the enhancer, with a single predictor representing whether the given position was wild type or mutant. The predictor was thus a binary variable representing presence (1) or absence (0) of a mutation at that position.

$$y_i = \beta_{0j} + \beta_{1j} X_{ij}$$

where,

y_i = number of aliquots in which the *i*th haplotype was observed (referred to as aliquot counts), and

X_{ij} = 1 if position *j* was mutant and 0 if position *j* was wild type in the *i*th haplotype.

To facilitate comparison between positions and between enhancers, we calculated the effect size of mutation at a position *j* as

$$\log_2\left(\frac{\beta_{0j}+\beta_{1j}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{1j} was used to judge whether the effect size was significant.

For LTV1, as a single haplotype was typically associated with multiple tags, we normalized the aliquot counts for a given haplotype by dividing by the number of tags associated with that haplotype. In the case of ALDOB and ECR11, as the enhancer haplotypes were cloned in second, almost all haplotypes were associated with single tags, and thus the aliquot counts for tags were used directly as the aliquot counts of their linked haplotypes.

Estimation of effect size of each specific nucleotide change at each position along the enhancer (trivariate model)

To explore whether the estimated effect sizes for each position were being driven by specific nucleotide substitutions, we modified the model just described to include three predictors, each representing one of the three possible nucleotide substitutions at that position. The factors were set up as binary variables representing the presence (1) or absence (0) of the particular change at that position.

$$y_i = \beta_{0j} + \beta_{1j}X_{ij_1} + \beta_{2j}X_{ij_2} + \beta_{3j}X_{ij_3}$$

Effect sizes were then calculated from the coefficients produced by the models as follows (for *k* = 1,2,3):

$$\log_2\left(\frac{\beta_{0j}+\beta_{kj}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{kj} was used to judge whether the effect of a given nucleotide substitution at a given position was significant.

Spatial structure

To quantify whether nearby positions tend to have similar effect sizes, we calculated the sum of the absolute values of the differences in effect sizes between positions located at a given distance (lag) from each other. In other words, we calculated

$S(k) = \sum_{j=k+1}^{N} |r_j - r_{j-k}|$,

where k = 1, 2, ..., 20 denotes the lag, N denotes the length of the enhancer, and r_i is the effect size of position *i*.

For each value of the lag k, we also calculated $S_{1^*}(k),...,S_{1000^*}(k)$, each of which measures the sum of the absolute values of the differences in effect sizes between positions at a distance k from each other, after permuting the effect sizes $(r_1,...,r_N)$. We then calculated a P-value associated with each value of the lag k as the fraction of the $S_{1^*}(k),...,S_{1000^*}(k)$ that was as small or smaller than S(k).

Models to estimate combined predictive power of blocks of adjacent positions

To further characterize the nature of the spatial structure of the effect sizes and to explore whether certain regions along the enhancer were enriched for positions with larger effect sizes, we focused on blocks of

adjacent positions in a 10-bp sliding window along the length of the enhancer. For each window, we built a multiple linear regression model with one predictor for each position within the window. Each predictor was set up as a binary variable denoting the presence (1) or absence (0) of mutation at that position. The response variable y was the number of aliquots in which a given haplotype was seen.

$$y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{i(j+1)} + \ldots + \beta_{10} X_{i(j+9)}$$

The F-statistic from each model was used as a measure of the collective predictive power of positions within each window.

Multiple linear regression models based on the entire haplotype

The multiple linear regression model included one predictor for each position along the enhancer, encoded as a 1 or 0 to indicate presence or absence of a mutation at that position on a given haplotype, and the response variable y represented the number of aliquots in which the haplotype was observed. Here N is the number of positions within a given enhancer.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_N X_{iN}$$

A P-value for the model was calculated by comparing the mean squared error (MSE) of the model to MSEs of 200 models built using randomly shuffled versions of the response variable. A P-value for the model was estimated by calculating the fraction of times that the MSE for models built using a shuffled response vector was at least as small as the MSE computed using real data.

We then expanded the model, such that each position was represented by three predictors to indicate which of the three possible nucleotide substitutions was observed at that position.

$$y_i = \beta_0 + \beta_{1j} X_{i1_1} + \beta_{2j} X_{i1_2} + \beta_{3j} X_{i1_3} + \cdots + \beta_{1j} X_{iN_1} + \beta_{2j} X_{iN_2} + \beta_{3j} X_{iN_3}$$

A P-value for the model was calculated by repeatedly permuting the outcome vector as described immediately above; however, only 100 permutations were used, due to the high computational burden of constructing this model.

Identification of epistatic interactions (that is, nonadditive effects) among pairs of mutations

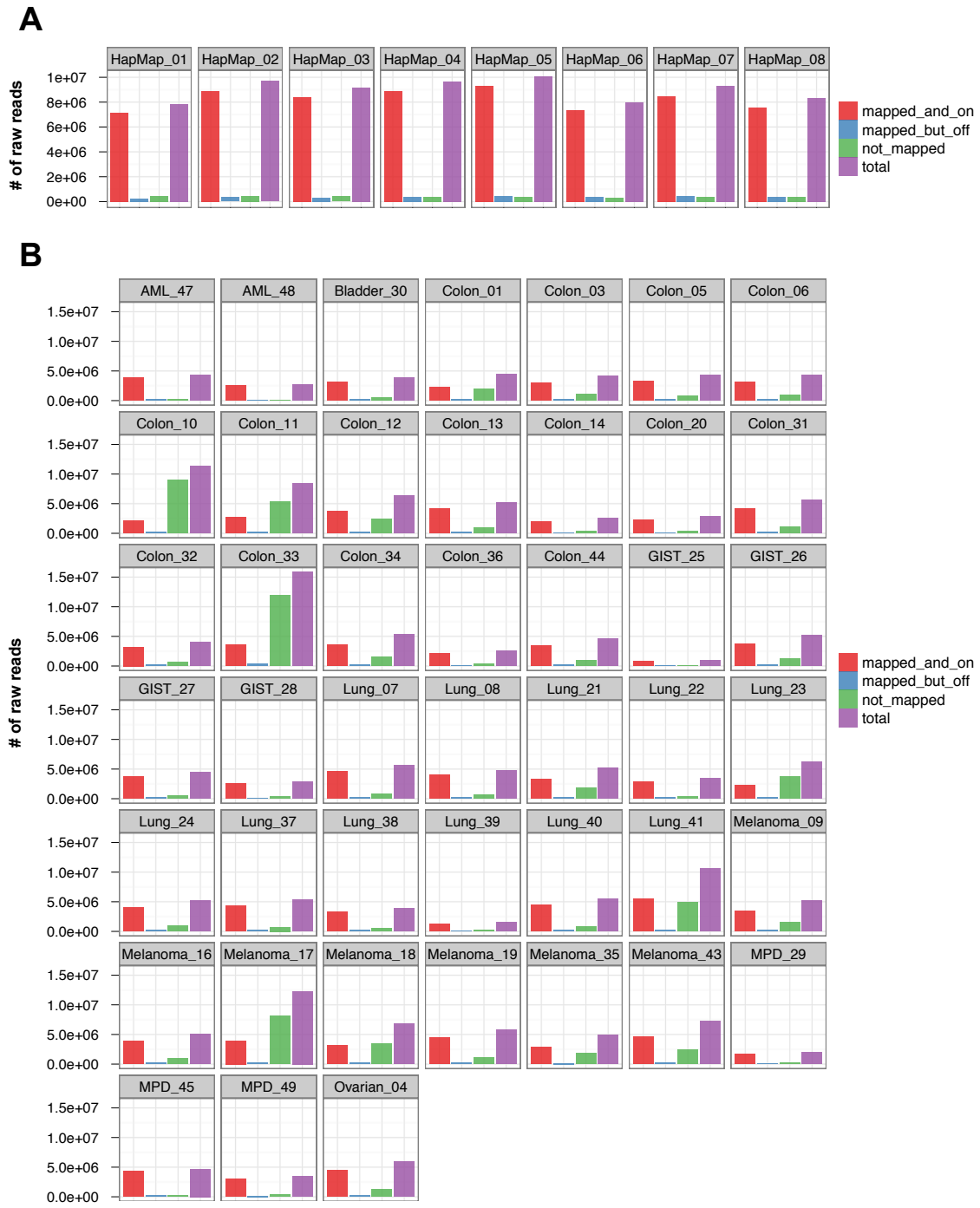
For each pair of positions, we built a linear multiple regression model with three predictors: one predictor each to indicate the presence (1) or absence (0) of a mutation at each of the two positions and a third (referred to as the “interaction term”) whose value was set to 1 if both positions were mutant on the given haplotype and 0 otherwise. Only pairs of positions that were both mutant on at least twenty haplotypes were considered.

$$y_i = \beta_{0jk} + \beta_{1jk} X_{ij} + \beta_{2jk} X_{ik} + \beta_{3jk} X_{ij} X_{ik}$$

We used the P-values for the interaction terms for the resulting models to calculate a FDR for each interaction term (using the p.adjust() function in R, with method = “BH”). Interaction terms with FDR < 0.05 were considered significant and used for downstream analyses of epistatic interactions.

Appendix C Supplementary materials for Chapter 6

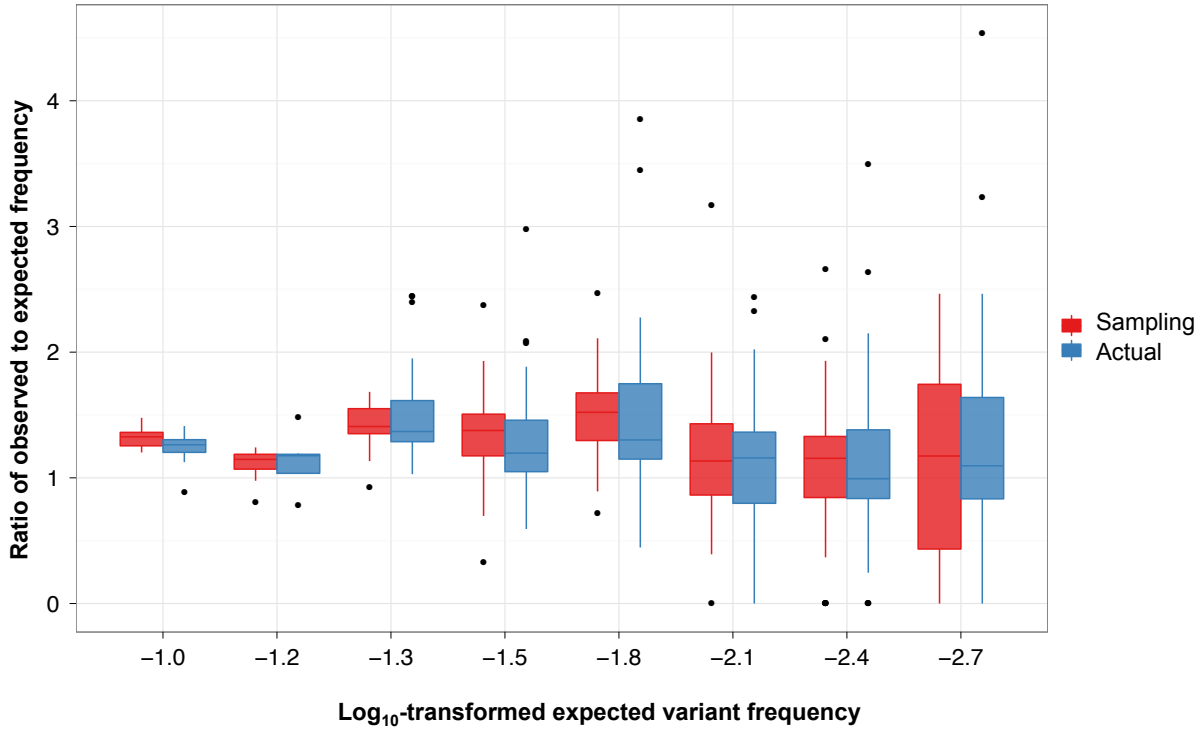
Figure C.1. Mapping and on-target rates of HiSeq data.



The total number of raw reads, mapping and on-target raw reads ("mapped_and_on") as determined by expected read start position and orientation with respect to the reference, mapping but off-target raw

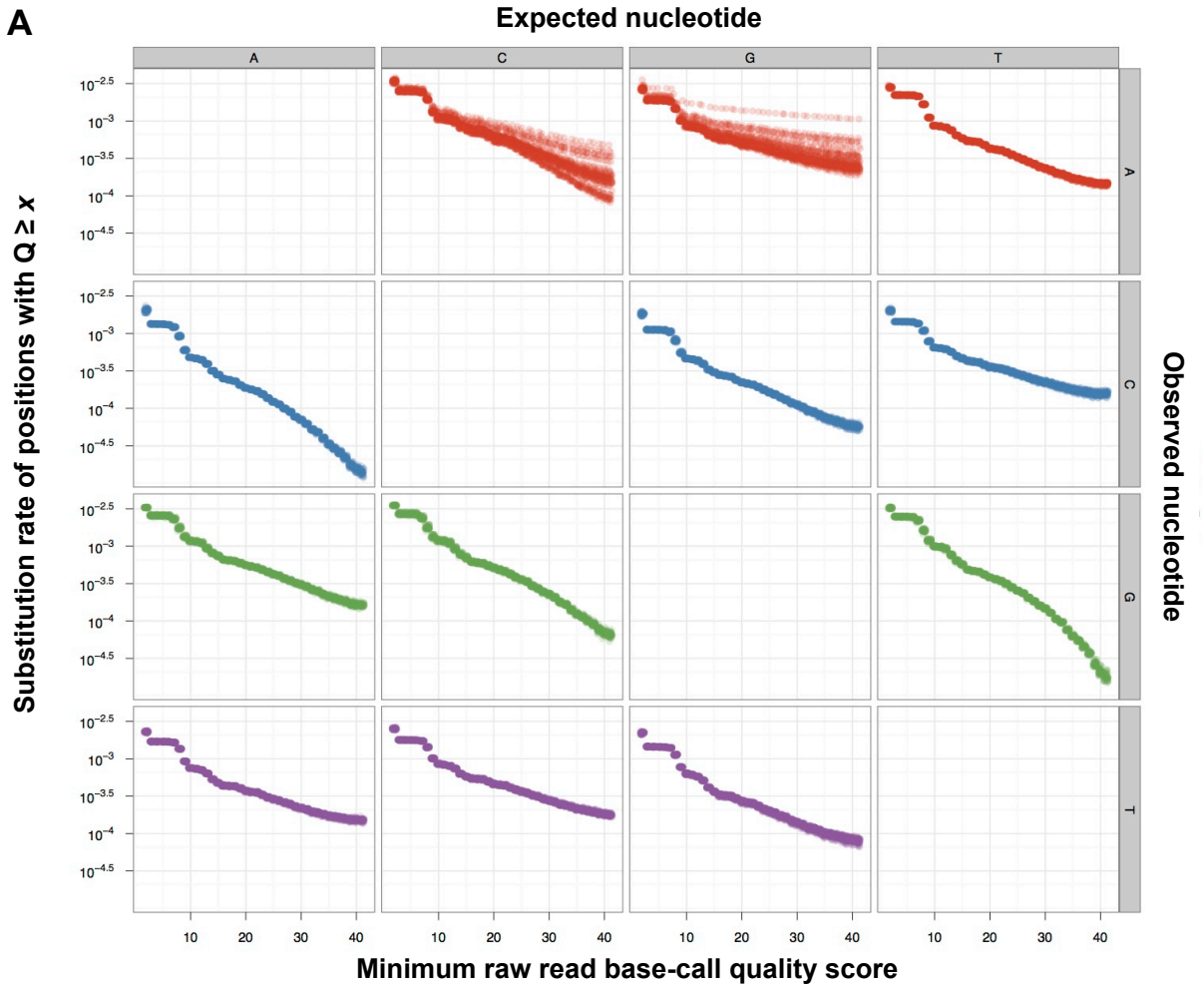
reads (“mapped_but_off”), and unmapped raw reads (“not_mapped”) are shown for the eight HapMap cell line samples (a) and the 45 successfully captured clinical samples (b).

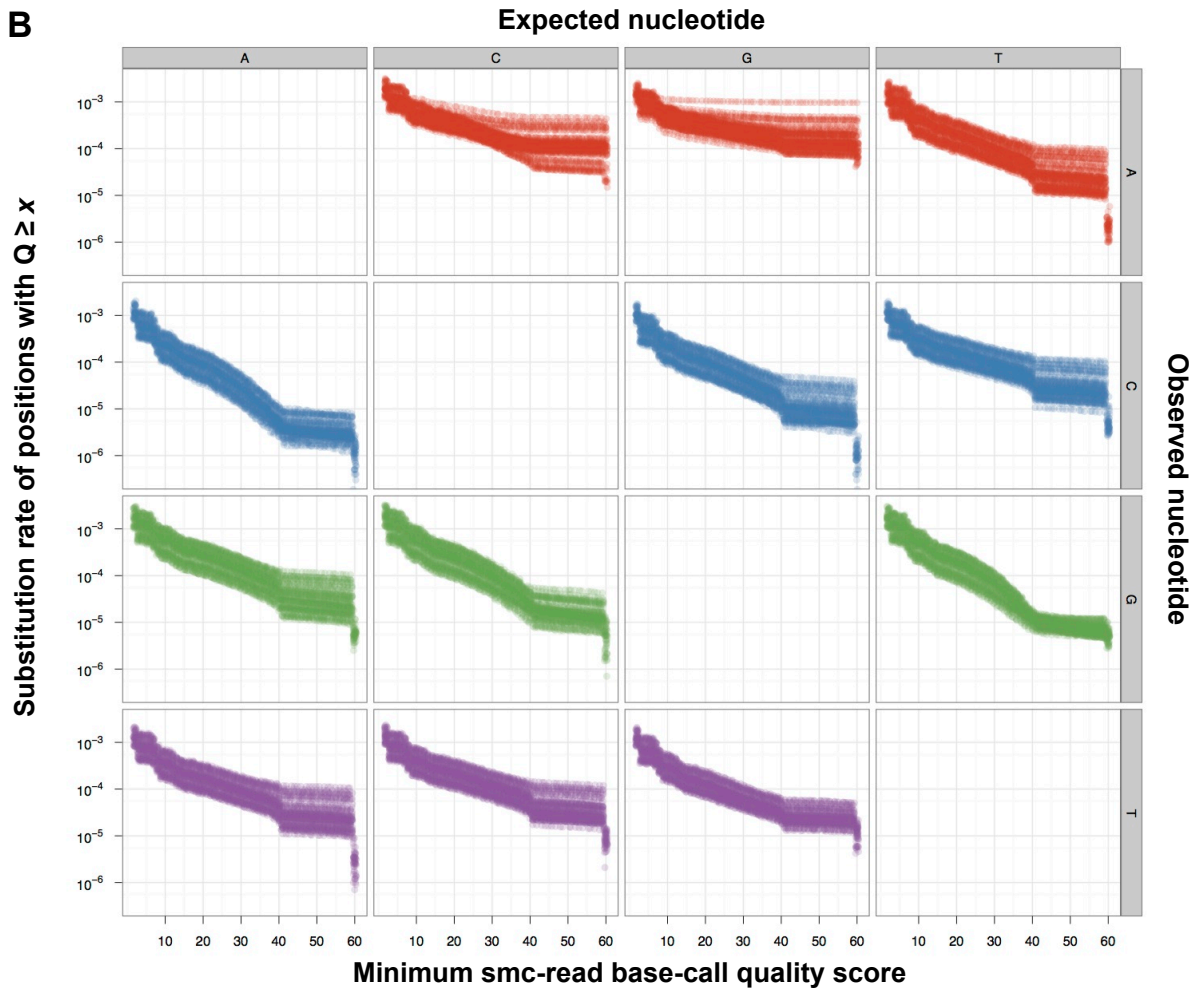
Figure C.2. Comparison of observed deviation from expected variant frequencies to Poisson sampling.



Distributions of the ratio of observed to expected variant frequency across log-transformed expected variant frequency bins for actual smc-read base-calls compared to Poisson sampling using the number of observations and the observed frequency. Restricted to positions with coverage at least 100x. The comparable distributions show that deviation from expectation, especially at low frequencies, is largely explained by sampling statistics.

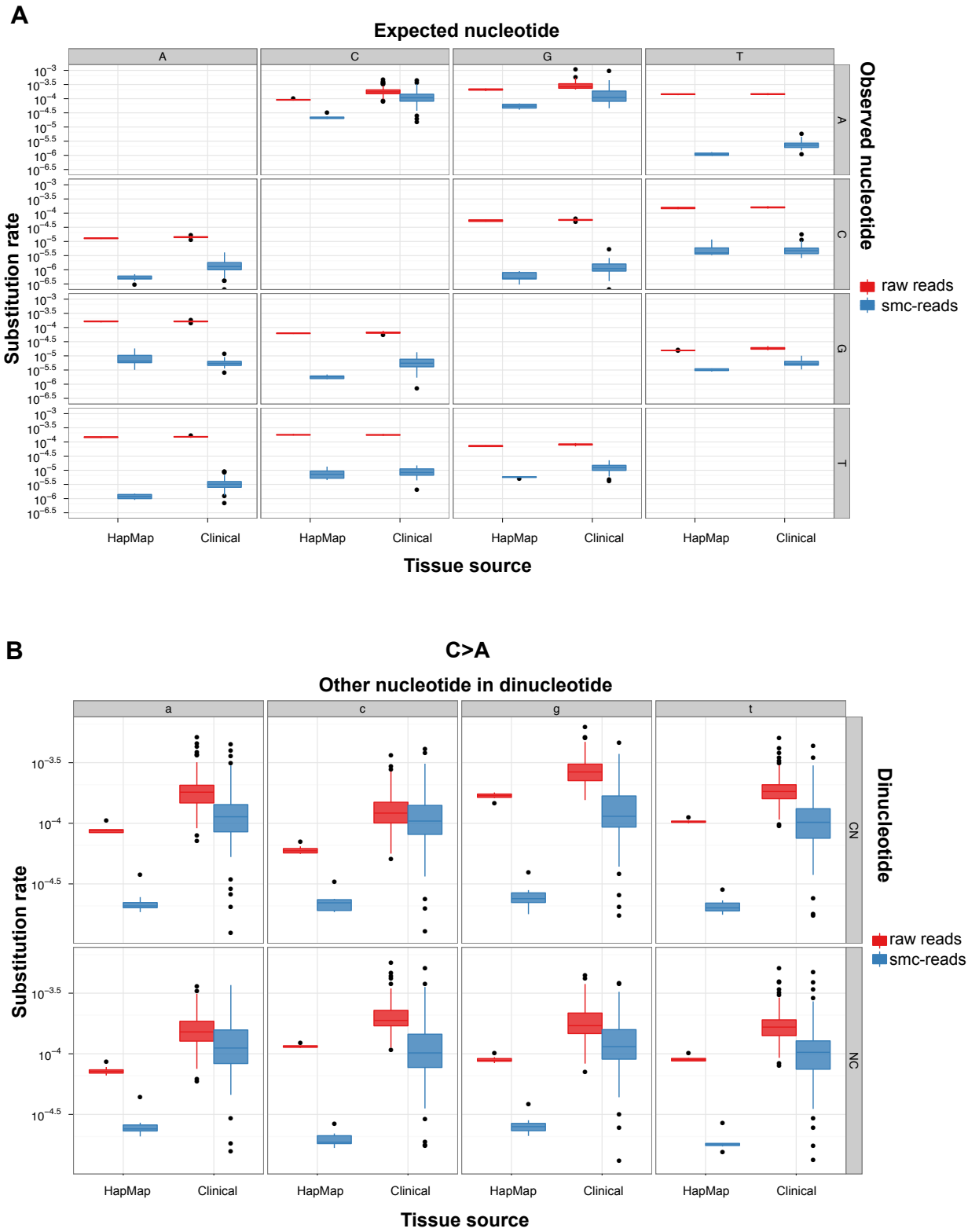
Figure C.3. Substitution rate as a function of minimum base-call quality score and expected and observed gap-fill nucleotides.

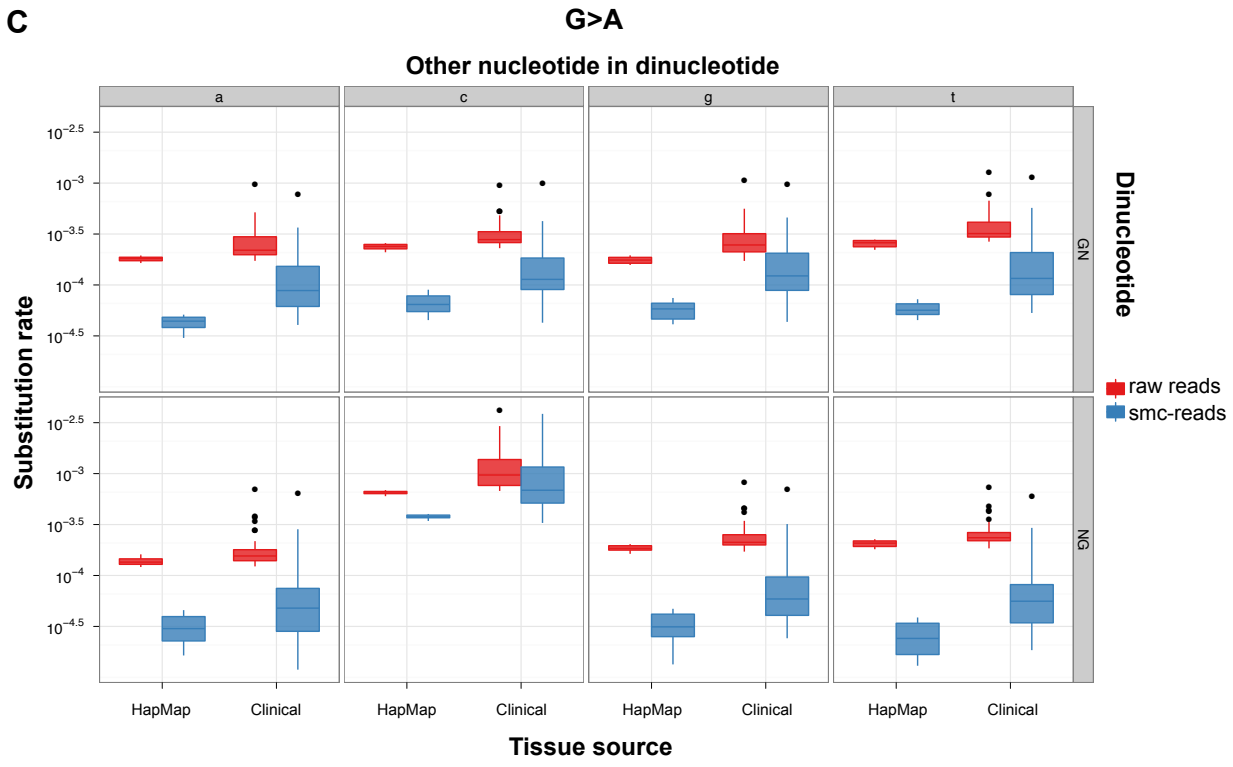


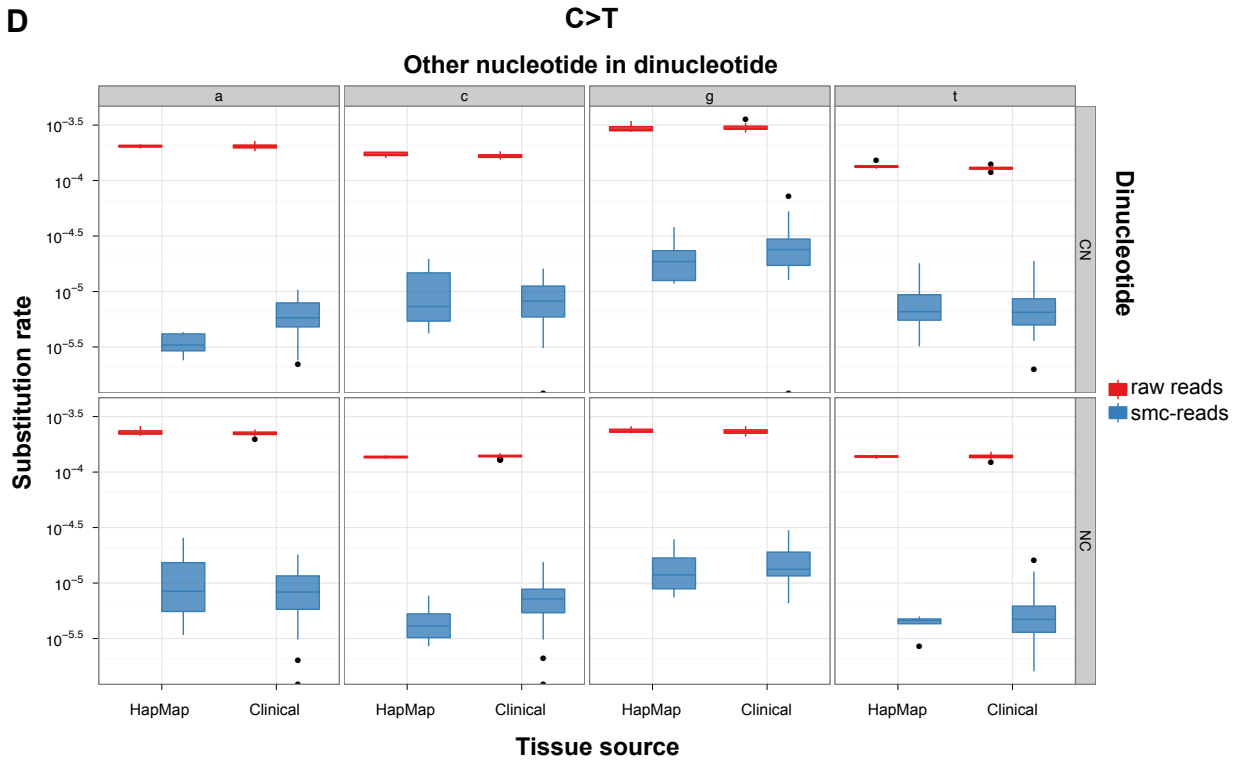


For putative homozygous reference positions, plotted is the substitution error rate as a function of observed and expected nucleotide in the gap-fill for all base-calls in raw reads (a) and smc-reads (b) with at least a given quality score. We note that an smc-read call must have multiple independent raw reads contributing to achieve a quality of 60, as a raw call with Q60 (due to the merging process) while be rescaled to 59 according to the likelihood ratio quality estimation process and subsequent integer casting.

Figure C.4. Substitution rate distributions as a function of expected and observed gap-fill nucleotide.



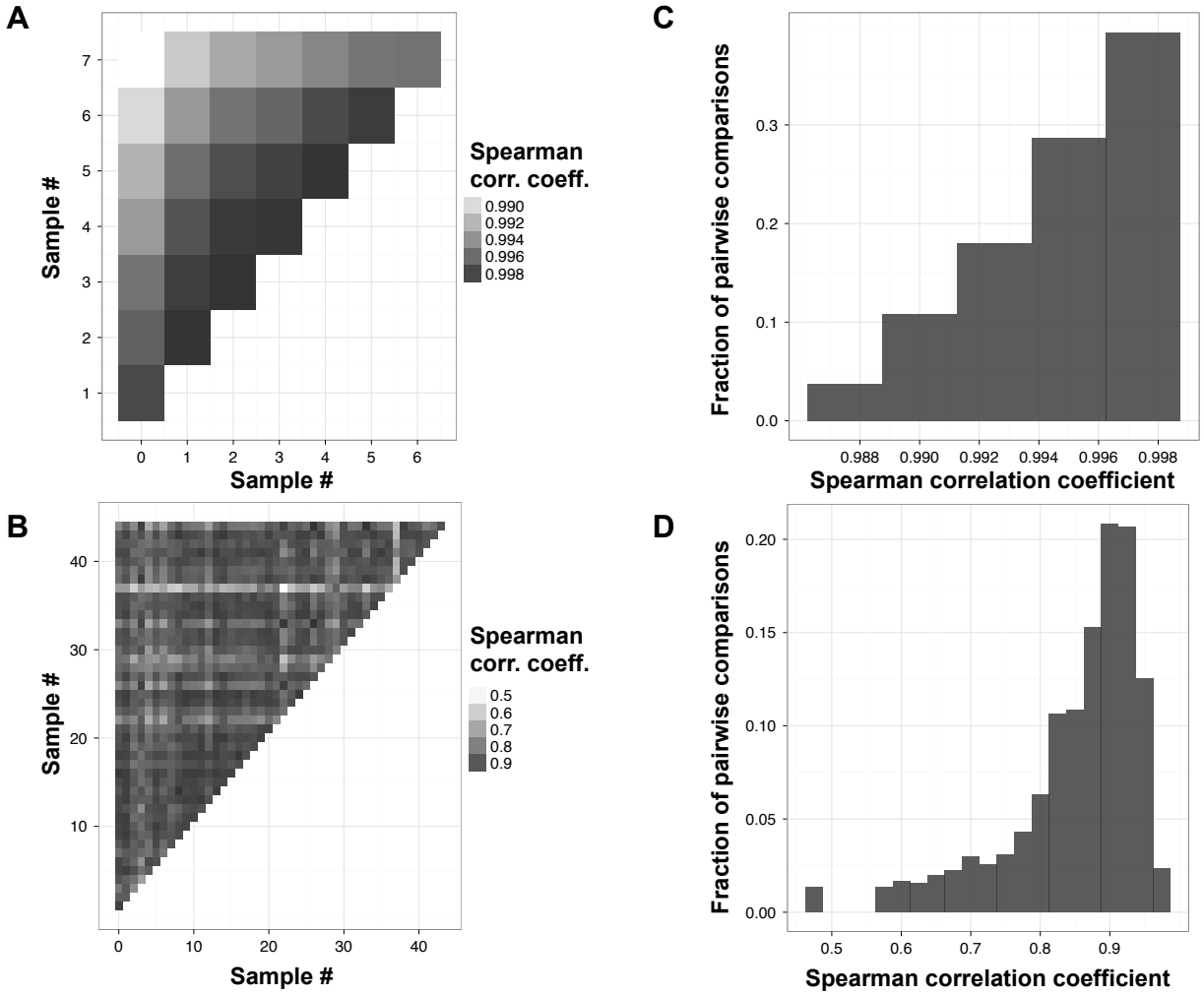




All rates are shown for a given expected gap-fill mono- or dinucleotide, that is the complementary nucleotide(s) to the nucleotide(s) present in the target genomic DNA, considering only ≥Q41 raw read

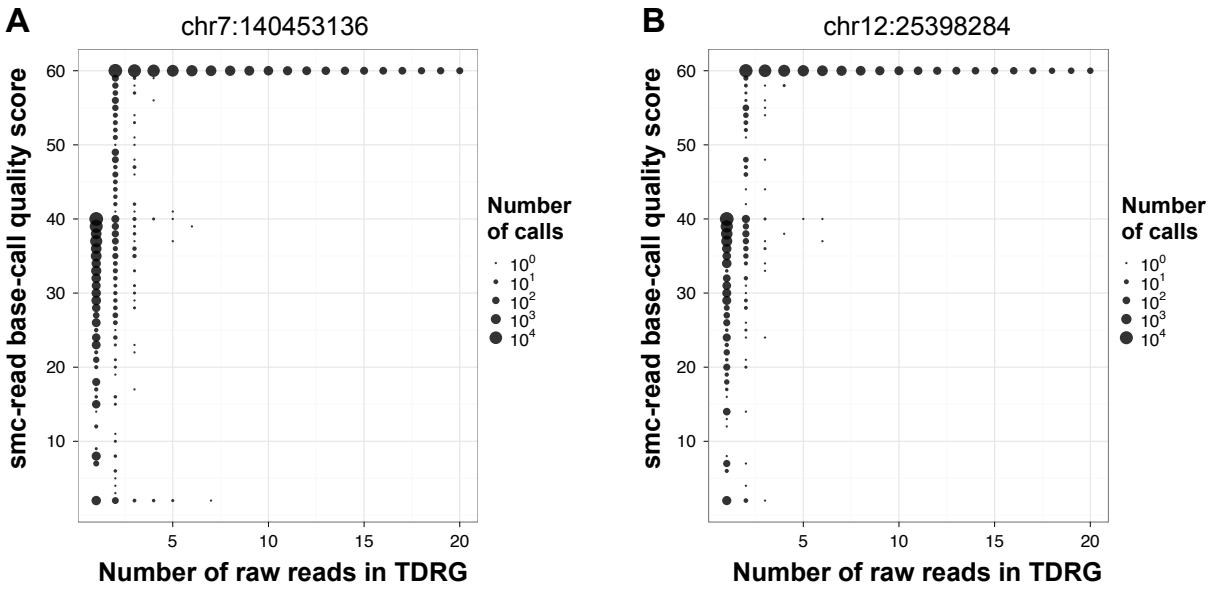
base-calls and Q60 smc-read base-calls at putative homozygous positions based on GATK calls. (a) Mononucleotide substitution rates for raw and smc-reads, and HapMap and clinical samples, for all twelve possible substitutions. (b) Substitution rate of cytosine to adenosine for all possible gap-fill dinucleotides involving cytosine as one of the two nucleotides, for raw and smc-reads and for HapMap and clinical samples. (c) Substitution rate of guanine to adenosine for all possible gap-fill dinucleotides involving guanine as one of two nucleotides, for raw and smc-reads and for HapMap and clinical samples. (d) Substitution rate of cytosine to thymine for all possible gap-fill dinucleotides involving cytosine as one of two nucleotides, for raw and smc-reads and for HapMap and clinical samples.

Figure C.5. Reproducibility of relative probe capture efficiency.



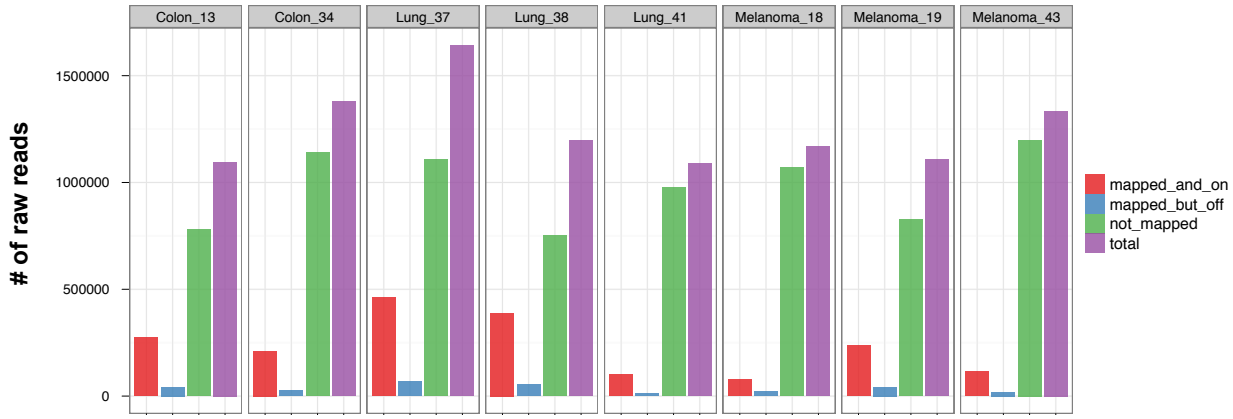
Spearman correlation coefficients of probe capture efficiency (measured as the number of TDRGs associated with each probe) between capture reactions are shown for the 8 HapMap cell line samples (a) and 45 clinical samples (b). Also shown are histograms summarizing correlation coefficients for HapMap samples (c) and clinical samples (d).

Figure C.6. smc-read base-call quality score as a function of TDRG size.



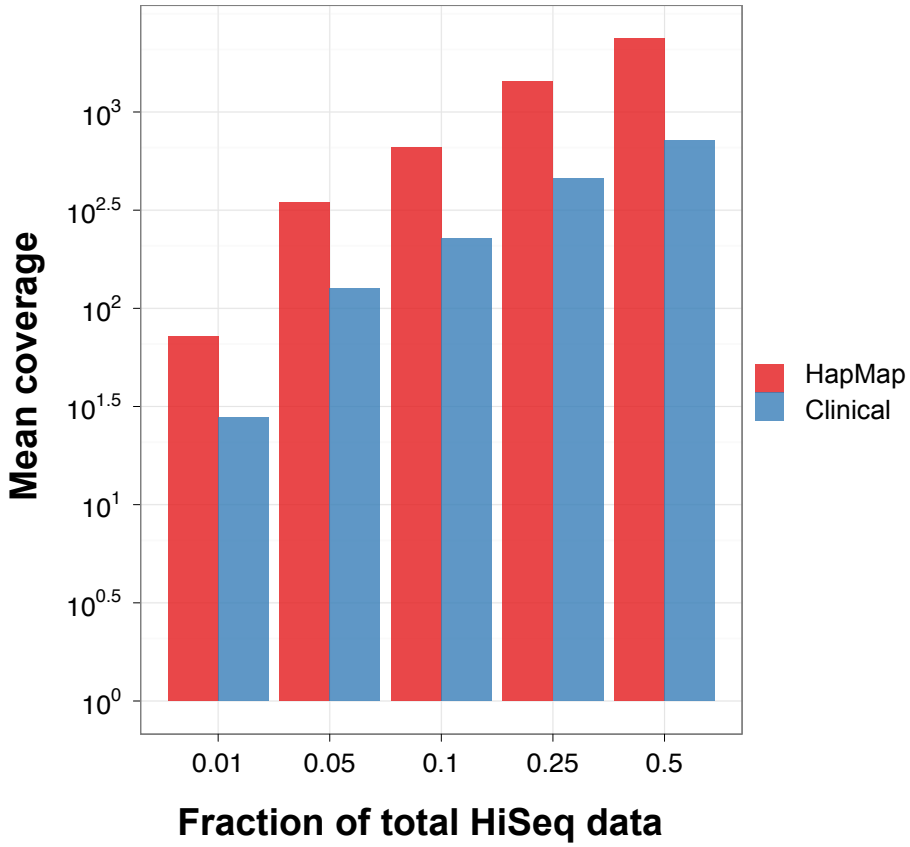
Plotted is the number of smc-read base-calls, across all samples and for smc-reads aligning without insertion or deletion edits, as a function of the number of raw reads in the TDRG associated with that smc-read and the estimated quality score of the base-call at chr7:140453136 (*BRAF* p.V600) (a) and chr12:25398284 (*KRAS* p.G12) (b). smc-read base-call quality score is a measure of TDRG clonality, as positions with discordant calls will be assigned lower quality scores.

Figure C.7. Mapping and on-target rates from MiSeq rapid workflow assay.



The total number of raw reads, mapping and on-target raw reads ("mapped_and_on") as determined by expected read start position and orientation with respect to the reference, mapping but off-target raw reads ("mapped_but_off"), and unmapped raw reads ("not_mapped") are shown for the eight clinical samples subjected to the "low-coverage/rapid workflow" smMIP implementation.

Figure C.8. Downsampling effect on mean coverage.



Mean fold coverage across 8 HapMap (red) and 45 clinical (blue) samples for various fractions of overall raw HiSeq data (3.2×10^8 paired-end reads or ~64.8Gb of sequence) ranging from 1% to 50%.

Table C.1. Targeted genes.

Gene	# of coding bases	# of discrete intervals	# of targeted coding bases	# of smMIPs
ABL1	3529	12	3482	47
AKT1	1443	13	1443	29
AKT2	1446	13	1446	28
APC	8697	16	8697	105
BRAF	2301	18	2301	41
CDK4	912	7	912	15
CDKN2A	912	5	912	16
CSF1R	2919	21	2919	47
CTNNB1	2346	14	2346	35
EGFR	3889	30	3889	67
ERBB2	3768	27	3768	63
FGFR1	2635	19	2635	45
FGFR2	2710	20	2640	48
FGFR3	2572	18	2572	43
FLT3	2982	24	2982	50
HRAS	633	5	633	12
JAK2	3399	23	3399	53
JAK3	3375	23	3375	56
KIT	2931	21	2931	47
KRAS	687	5	687	13
MET	4227	20	4227	61
MLH1	2271	19	1872	44
MYC	1365	3	1365	17
NRAS	570	4	570	8
PDGFRA	3270	22	3270	49
PIK3CA	3207	20	3207	51
PTEN	1212	9	1212	22
RB1	2787	27	2787	58
RET	3377	20	3377	52
SRC	1611	11	1611	27
STK11	1302	9	1302	23
TP53	1263	12	1215	31
VHL	642	3	400	9
Total	81190	513	80384	1312

Gene name, size, number of discrete genomic intervals, number of coding bases for which probes were successfully designed, and number of probes for targeted capture of the coding sequence of 33 cancer-related genes harboring clinically actionable mutations.

Table C.2. Description of clinical samples.

Sample ID	Estimated % tumor nuclei	Diagnosis	Site	Tumor category	Collection date	DNA isolation date
1	10	Rectal adenocarcinoma	Liver	Colon	7/08	6/09
2*	1	Colorectal adenocarcinoma	Colon	Colon	NA	12/13/11
3	30	Colorectal adenocarcinoma	Lymph node	Colon	6/10	1/11
4	60	Ovarian adenocarcinoma	Ovary	Ovarian	5/09	NA
5	50	Rectal adenocarcinoma	Rectum	Colon	12/10	1/11
6	30	Colorectal adenocarcinoma	Colon	Colon	3/10	1/11
7	40	Metastatic papillary lung adenocarcinoma	Lymph node	Lung	9/10	10/10
8	100	Lung adenocarcinoma	Brain	Lung	4/11	4/11
9	80	Melanoma	Lymph node	Melanoma	5/10	2/11
10	50	Colorectal adenocarcinoma	Cecum	Colon	3/11	6/11
11	50	Metastatic colorectal adenocarcinoma	Liver	Colon	2/10	12/11
12	60	Colorectal adenocarcinoma	Colon	Colon	8/09	1/11
13	50	Colorectal adenocarcinoma	Terminal ileum	Colon	9/10	11/10
14	40	Colorectal adenocarcinoma	Colon	Colon	6/10	9/10
15*	60	Colorectal adenocarcinoma	Sigmoid rectum	Colon	4/10	12/11
16	NA	Melanoma	Thigh	Melanoma	3/10	4/10
17	70	Melanoma	Axillary mass	Melanoma	6/11	11/11
18	90	Metastatic melanoma	Lymph node	Melanoma	10/08	3/11
19	45	Melanoma	Brain	Melanoma	4/08	11/11
20	50	Colorectal adenocarcinoma	Colon	Colon	10/09	12/11
21	50	Non-small cell lung cancer	Lung	Lung	12/09	7/10
22	NA	Non-small cell lung cancer	Pleural effusion	Lung	4/10	5/10
23	50	Lung adenocarcinoma	Lung	Lung	2/07	11/10
24	80	Lung adenocarcinoma	Lung	Lung	NA	9/10
25	80	Gastrointestinal stromal tumor	NA	GIST	NA	NA
26	75	Gastrointestinal stromal tumor	Stomach	GIST	2/11	11/11
27	50	Gastrointestinal stromal tumor	Stomach	GIST	5/11	5/11
28	NA	Gastrointestinal stromal tumor	Ileum	GIST	5/11	5/11
29**	NA	Polycythemia vera	Peripheral blood	MPD	11/05	11/05
30	NA	Metastatic urothelial carcinoma	Lung	Bladder	9/04	5/10
31	NA	Colorectal adenocarcinoma	Sigmoid colon	Colon	3/11	4/11
32	NA	Colorectal adenocarcinoma	Sigmoid rectum	Colon	8/09	4/11
33	70	Rectal adenocarcinoma	Liver	Colon	2/11	7/11
34	20	Colorectal adenocarcinoma	Lung	Colon	1/11	2/11
35	50	Melanoma	Nasal cavity	Melanoma	NA	8/10
36	70	Metastatic colorectal adenocarcinoma	Lung	Colon	9/09	7/10
37	75	Metastatic lung	Small intestine	Lung	8/10	8/10

		adenocarcinoma				
38	50	Metastatic lung adenocarcinoma	Cerebellum	Lung	10/11	11/11
39	30	Metastatic non-small cell lung cancer	Lymph node	Lung	10/10	10/10
40	50	Lung adenocarcinoma	Lung	Lung	11/10	11/10
41	50	Lung adenocarcinoma	Lymph node	Lung	3/10	8/10
43	40	Melanoma	Skin	Melanoma	7/10	10/11
44	10	Metastatic colorectal adenocarcinoma	Liver	Colon	7/08	12/11
45**	NA	Polycythemia vera	Peripheral blood	MPD	11/05	11/05
47**	NA	Acute myeloid leukemia	Bone marrow	AML	4/10	4/10
48**	NA	Acute myeloid leukemia	Peripheral blood	AML	11/08	12/08
49**	NA	Polycythemia vera	Peripheral blood	MPD	12/05	12/05

Numerical sample ID, histologically estimated percent tumor nuclei, clinical diagnosis, biopsy/collection site, tumor category (MPD: myeloproliferative disorder; GIST: gastrointestinal stromal tumor; AML: acute myeloid leukemia), tissue collection date (month/year), and DNA isolation date are shown for samples collected during routine clinical practice. Specimens were FFPE tissue unless otherwise indicated. NA: not applicable or information not available. *Failed capture, excluded from subsequent analysis. **Not FFPE tissue.

Table C.3. Substitution rates for guanine to adenine.

		Raw reads		smc-reads		Fold-reduction in sub. rate
		Calls	Sub. rate	Calls	Sub. rate	
HapMap cell lines	No CG>CA	1.39E+09	1.93E-04	6.19E+08	4.24E-05	4.55
	Only CG>CA	4.90E+07	6.50E-04	2.38E+07	3.77E-04	1.72
Clinical samples	No CG>CA	3.17E+09	2.76E-04	1.14E+09	1.26E-04	2.18
	Only CG>CA	1.37E+08	1.24E-03	5.47E+07	9.11E-04	1.36

Total number of calls, substitution rates, and the fold-reduction in substitution rate comparing smc-reads to raw reads for very high-confidence (≥Q41) base-calls from raw reads and smc-reads (Q60), shown for the incorporation of dA into the probe when dG was expected, as a function of dinucleotide context. Although the substitution rate for CG>CA in smc-reads is higher by ~8-fold compared to all other G>A substitutions, CG is an infrequent dinucleotide, and the substitution rate of G>A in the absence of CG>CA remains elevated, suggesting that deamination of 5mC is not the only factor causing high rates of G>A substitution, and that deamination of C to U may also be playing a substantial role.

Table C.4. Detection of previously ascertained substitution mutations in clinical samples.

Sample #	Cancer type	Gene	Chr.	Position	Ref. allele	Sample genotype	# of ref. allele obs.	# of alt. allele obs.	Alt. allele fraction	Mut. in protein
9	Melanoma	BRAF	7	140453136	A	A/T	2118	594	0.22	p.V600E
16	Melanoma	BRAF	7	140453136	A	A/T	1842	1250	0.40	p.V600E
17	Melanoma	BRAF	7	140453136	A	A/T	358	698	0.66	p.V600E
18	Melanoma	BRAF	7	140453136	A	A/T	906	1546	0.63	p.V600K*
18	Melanoma	BRAF	7	140453137	C	C/T	905	1548	0.63	p.V600K*
19	Melanoma	BRAF	7	140453136	A	A/T	2874	1843	0.39	p.V600K*
19	Melanoma	BRAF	7	140453137	C	C/T	2875	1844	0.39	p.V600K*
20	Colon	BRAF	7	140453136	A	A/T	2502	1596	0.39	p.V600E
21	Lung	EGFR	7	55259515	T	G/T	386	319	0.45	p.L858R
22	Lung	EGFR	7	55259515	T	G/T	1659	793	0.32	p.L858R
29	MPD	JAK2	9	5073770	G	G/T	1443	865	0.37	p.V617F
45	MPD	JAK2	9	5073770	G	G/T	643	6234	0.91	p.V617F
49	MPD	JAK2	9	5073770	G	G/T	3462	597	0.15	p.V617F
1	Colon	KRAS	12	25398285	C	C/A	468	81	0.15	p.G12C
3	Colon	KRAS	12	25398284	C	C/T	663	103	0.13	p.G12D
8	Lung	KRAS	12	25398285	C	C/A	478	614	0.56	p.G12C
11	Colon	KRAS	12	25398284	C	C/T	141	100	0.41	p.G12D
12	Colon	KRAS	12	25398281	C	C/T	602	285	0.32	p.G13D
13	Colon	KRAS	12	25398284	C	C/A	593	319	0.35	p.G12V
14	Colon	KRAS	12	25398281	C	C/T	801	435	0.35	p.G13D
37	Lung	KRAS	12	25398284	C	C/A	624	1684	0.73	p.G12V
44	Colon	KRAS	12	25398285	C	C/A	480	109	0.19	p.G12C
43	Melanoma	NRAS	1	115256529	T	T/C	225	65	0.22	p.Q61R
26	GIST	PDGFRA	4	55152093	A	A/T	859	334	0.28	p.D842V
27	GIST	PDGFRA	4	55152093	A	A/T	1451	475	0.25	p.D842V

Numerical sample ID, cancer type, gene name, chromosome, position, reference allele, sample genotype, number of smc-read reference allele calls, number of smc-read alternate allele calls, relative fraction of alternate allele calls, and resulting protein change are shown for mutations at clinically relevant sites for which these samples were previously genotyped. Two *KRAS* mutations in Lung samples that were not previously genotyped at that site but were subsequently confirmed are also shown (*KRAS* mutations in samples 8 and 37). *p.V600K is caused by adjacent substitutions at positions chr7:140453136 and chr7:140453137.

Table C.5. Detection of previously ascertained insertion and deletion mutations in clinical samples.

Sample #	23	24	28	25
Cancer type	Lung	Lung	GIST	GIST
Gene	EGFR	EGFR	KIT	KIT
Chromosome	7	7	4	4
Position	55242467	55242469	55593600	55593646
Exon	19	19	11	11
Reference allele	AATTAAGAGAAGCAAC	TTAAGAGAAGCAACATCTC	CAGTGGAAGGTTGTTG	T
Sample genotype	AATTAAGAGAAGCAAC / A	TTAAGAGAAGCAACATCTC / T	CAGTGGAAGGTTGTTG / C	T / TAGACCC
Reference allele counts	358	817	1143	1033
Alternate allele counts	155	363	1186	766
Alternate allele fraction	0.30	0.31	0.51	0.43
Reference amino acids	LREAT	LREATSP	WKVVE	-
Reference amino acid positions	747-751	747-753	557-561	after 571
Alternate amino acids	-	S	-	DP

Numerical sample ID, cancer type, gene name, chromosome, position, exon number, reference allele, sample genotype, number of smc-read reference allele calls, number of smc-read alternate allele calls, relative fraction of alternate allele calls, reference amino acids, positions of reference amino acids, and amino acids in place of reference amino acids caused by mutation. and resulting protein change are shown for mutations at clinically relevant sites for which these samples were previously genotyped. Two larger insertions in *FLT3* (67 and 104 bp, but of unknown exact position) were not detected in the smMIP data using the analysis strategy currently employed.

Table C.6. Effect of down-sampling on substitution error rate.

	Fraction of data	Raw sequence data	Raw reads		smc-reads		Fold-reduction in sub. rate
			Calls	Sub. rate	Calls	Sub. rate	
HapMap cell lines	0.01	1.30e+09	1.02E+08	1.08E-04	1.84E+06	1.03E-05	10.46
	0.05	6.48e+09	5.13E+08	1.09E-04	4.25E+07	8.31E-06	13.06
	0.1	1.30e+10	1.03E+09	1.08E-04	1.55E+08	8.50E-06	12.74
	0.25	3.24e+10	2.57E+09	1.08E-04	7.39E+08	8.36E-06	12.89
	0.5	6.48e+10	5.15E+09	1.08E-04	2.04E+09	8.45E-06	12.80
Clinical samples	0.01	1.30e+09	2.16E+08	1.26E-04	1.17E+07	3.06E-05	4.12
	0.05	6.48e+09	1.09E+09	1.27E-04	1.97E+08	2.90E-05	4.37
	0.1	1.30e+10	2.18E+09	1.27E-04	5.67E+08	2.90E-05	4.36
	0.25	3.24e+10	5.45E+09	1.27E-04	1.93E+09	2.89E-05	4.39
	0.5	6.48e+10	1.09E+10	1.27E-04	4.17E+09	2.80E-05	4.53

Fraction of raw reads (prior to raw read formation) used, number of base-calls, substitution rate and fold-reduction in substitution rate comparing smMIP Q60 calls to raw read ≥Q41 calls when using subsets of the $3.32x10^8$ total read-pairs collected for the study for both the HapMap cell line and tumor-derived samples.

Table C.7. Oligo sequences.

Name	Sequence
MIP backbone	CTTCAGCTTCCCGATCCGACGGTAGTGTNNNNNNNNNNNN
Forward PCR primer	AATGATACGGCGACCACCGAGATCTACACATACGAGATCCGTAATCGGGAAGCTGAAG
Reverse PCR primer*	CAAGCAGAAGACGGCATACGAGATXXXXXXXXACACGCACGATCCGACGGTAGTGT
Index 1	CTCTAGCA
Index 2	AGCTCTCA
Index 3	TGAGTGAC
Index 4	ACGCTTAT
Index 5	CAGATAGT
Index 6	GTCACCAT
Index 7	TGGTCGAA
Index 8	GCAATATA
Index 9	CACATGCA
Index 10	TCCTTCGA
Index 11	CTGATGTA
Index 12	CACTGCAA
Index 13	TCGGAGAA
Index 14	CTGAGCTT
Index 15	CTTGGTAC
Index 16	CTGACAAT
Index 17	TGGTACAG
Index 18	GGTCTCAA
Index 19	TGGCTAAT
Index 20	AGAGGATC
Index 21	CGAATACA
Index 22	AGCGTTAC
Index 23	TGACCTCA
Index 24	TAGTTGCC
Index 25	GTTGCAGT
Index 26	ATAGAGGC
Index 27	CTTGACTG
Index 28	AACCTCGA
Index 29	TCAACCGA
Index 30	GCTATGGA
Index 31	TCTTGACC
Index 32	ACATGGAT
Index 33	ATGACAGC
Index 34	AACTCCTG
Index 35	ACTTAAGG
Index 36	TCTTGCAT
Index 37	GACTGTTC
Index 38	ACTGACCT
Index 39	TGTGTCCA
Index 40	CCTGTCAT
Index 41	ATGTACTT
Index 42	GAATAATC
Index 43	CCTTAGAA
Index 44	GAATTCGC
Index 45	CTAGTCCT
Index 46	TCAGAGGT
Index 47	GACATTCT
Index 48	CTAACACG
Index 49	GTGTGATC
Index 50	CTGTTCAC
Index 51	GGTCAGTT
Index 52	AACCGATC
Index 53	CAGCTAAG
Read 1 sequencing primer	CATACGAGATCCGTAATCGGGAAGCTGAAG
Index sequencing primer	ACACTACCGTCGGATCGTGCGTGT
Read 2 sequencing primer	ACACGCACGATCCGACGGTAGTGT

Sequences of oligos used in MIP design, library construction, and sequencing. Probe sequences available upon request. All oligos, including probes, were procured from IDT using standard purification, with no modifications. *“XXXXXXXX” is replaced with one of 53 index sequences.

Note C.1. Two layers of indexing in the smMIP strategy.

As is now commonplace, sample index sequences are incorporated at the PCR step so that the capture products of several samples can be sequenced as a single pool. However, we also employed molecular tagging, wherein we modified the common region of each MIP to contain a twelve nt degenerate sequence, such that each capture event is associated with a unique tag sequence. This tag sequence is used to distinguish molecularly distinct capture events at the same locus within the same sample. Raw reads sharing the same sample index, aligning to the same locus, and sharing the same molecular tag sequence (i.e. tag-defined read groups, TDRGs) are compared to one another to yield a highly accurate single molecule consensus sequence that is expected to be devoid of polymerase and sequencing errors (Figure 6.1), with the important exception of polymerase errors that occur during the initial gap-fill event. These single molecule consensus reads (smc-reads) are also expected to yield more precise estimates of variant frequency by correcting for non-uniform amplification and ascertainment of different sequences.

Note C.2. Variation in fraction of aligning reads.

We observed variation in the fraction of reads that aligned, with half of all samples having greater than 80% of raw reads align to the target and ~80% (44/54) having greater than 60% aligning to target (Figure C.1). Across all samples, a very low fraction (<5%) of raw reads aligned to the reference but not to the target, and the remaining raw reads were unmapped. This phenomenon was much more prevalent in the clinical samples and corresponded to the presence of a low-molecular weight artifact after PCR, suggesting that low quality and imprecisely quantified input DNA may contribute to artifact formation and reduced mapping rates. Ongoing efforts are directed at reducing the formation of this artifact by altering capture conditions and removing the artifact prior to sequencing via SPRI bead-based size enrichment.

Note C.3. Coverage distributions.

We further explored the distribution of coverage across the target and the samples sequenced, finding that for 78% (35 of 45) of the clinical samples and all HapMap cell line samples, at least 85% of all targeted coding bases had at least 100x smc-read coverage. Furthermore, for all of the HapMap samples, at least 60% of all targeted coding bases had at least 1,000x smc-read coverage, and 42% (19 of 45) of the clinical samples had at least 50% of targeted coding bases with at least 1,000x smc-read coverage (Figure 6.2**a**). Finally, all HapMap samples and 80% (36 of 45) of the clinical samples had at least 97% of targeted coding bases with at least 10x smc-read coverage, while the worst performing sample had 94.5% of targeted coding bases covered with at least 10x smc-read coverage (Figure 6.2b).

Note C.4. Error correction by smc-reads and estimation of smc-read base quality.

In contrast with raw reads, subsets of which are amplification products of the same progenitor molecule, each smc-read represents the consensus of an independent progenitor molecule. We therefore expect that given a sufficient sampling of raw reads derived from the same progenitor molecule, sequencing and PCR errors should be corrected by our approach, with the important exception of errors introduced by the polymerase during the original gap-fill event that copies the target sequence into the molecular inversion probe. However, to accommodate TDRGs with few reads and discordant positions within TDRGs, we developed a likelihood ratio-based framework to estimate confidence in base-calls within smc-reads. Then, for the purposes of estimating and comparing error rates, we considered only maximum quality base-calls in raw reads and smc-reads (Figure C.3).

Note C.5. Variation in substitution rates as a function of expected and observed nucleotides.

Median substitution rates in raw reads ranged from ~$2x10^{-4}$ for the G>A substitution in tumor samples to ~$2x10^{-5}$ for the T>G substitution in HapMap cell lines, and for smc-reads, ranged from ~$1x10^{-4}$ for the same G>A substitution in tumor samples to ~$1x10^{-6}$ in HapMap samples for the A>T substitution (Figure

C.4a). However, for almost all substitutions and for both types of samples, smc-read base-call median substitution rates were at or below $1x10^{-5}$. Two notable exceptions to this pattern are the cases where an A is incorporated into the MIP when a G or a C was expected. We asked whether the nucleotide 5' or 3' to the nucleotide of interest affected the observed substitution rate. While the 5' or 3' nucleotide does not appear to influence the rate of C>A substitutions (Figure C.4b), G>A substitutions appear much more common in the CpG context (Figure C.4c, Table C.3). However, the CpG dinucleotide is rare and G>A mutation rates remain substantially elevated in the absence of the CpG dinucleotide.

Many factors may contribute to these patterns, including spontaneous polymerase errors during the gap-fill step, some types of DNA damage that predispose to such an error, and genetic heterogeneity in the input DNA. The C>A substitution in particular may be most simply explained by the presence, in the strand to which the MIP hybridizes, of the oxidative damage product 8-oxo-G, which is preferentially paired with A relative to C[205]. The C>A substitution is also notably elevated in the tumor samples relative to the HapMap cell lines, which could reflect oxidative damage induced by FFPE treatment and subsequent DNA isolation as well as higher rates of oxidative damage in tumors. One explanation for the G>A substitution, on the other hand, is spontaneous deamination of cytosine to uracil or 5-methyl-cytosine (5^mC) to thymine. Given that the G>A substitution is especially elevated in the CpG context and the polymerase used in the gap-fill step has a propensity to stall at Uracil residues[206], this pattern may, at least in part, reflect *in vivo* deamination of 5^mC prior to DNA isolation. If this indeed reflects *in vivo* deamination in a replicating population of cells, one would also expect to observe the corresponding C>T mutation in the CpG context at elevated rates. Indeed, this substitution, while at low rates overall, appears modestly elevated in the CpG context (Figure C.4d), but it should also be noted that some fraction of deamination events will be repaired prior to DNA replication. Finally, after removing the G>A substitutions occurring in the CpG context, the G>A substitution rate remains elevated (Table C.3). These substitutions may reflect an elevated error rate on the part of the gap-fill polymerase in the absence of any DNA modifications; alternatively, despite the tendency for the polymerase to stall at Uracil, it also possesses the ability to bypass these lesions at some rate[207], which could explain the elevated G>A substitution rate in the absence of the CpG context.

Despite the existence of these patterns, conservatively assuming that all observed substitutions are indeed false positives, smMIP reduces the error rate of targeted sequencing to ~$3x10^{-5}$ in DNA derived from clinical samples (most of which were FFPE tissues), and ~$8x10^{-6}$ in cell line-derived DNA, and should therefore be a highly accurate method for detecting low-frequency variation in specific genomic regions of interest. Furthermore, an awareness of these patterns will enable the development of more precise models to estimate confidence that a given mutation is a true sub-clonal variant compared to a library construction artifact.

Methods

Preparation of smMIP capture reagent

MIPs were designed as described elsewhere (*B.J.O. et al, manuscript under review*) against the coding exons of 33 cancer-related genes (Table C.1) with 50 nt of "splash" on either side of each exon. We designed probes with targeting arms summing to 40 nt in length, with ligation arms ranging in length from 16-20 nt and extension arms ranging in length from 20-24 nt. The gap-fill length was fixed to 112 nt. Targeting arms were joined by a constant 40-mer "backbone" sequence (common oligo sequences can be found in Table C.7) containing a stretch of twelve random nucleotides, such that each probe could exist in $\sim 4^{12}=1.67 \times 10^{6}$ unique sequences. After adding probes to accommodate sites of common variation in the genome that fell in targeting arms, we had a set of 1,312 probes targeting ~88 kb of coding sequence and ~125 kb overall. These 80mer probes were procured individually as column-synthesized oligos at 25 nanomole scale in 96-well plate format without any modifications or purification at a cost of $7.20 per probe. While a non-trivial up-front cost, this represents an effectively infinite supply, as each capture reaction consumes less than one ten-millionth of the supply of a given probe. Aliquots of each probe were pooled at equimolar ratios and 85 microliters of this pool was 5'-phosphorylated using 50 units of T4 Polynucleotide Kinase (NEB) and 1X T4 DNA ligase buffer in a total volume of 100 microliters for 45 minutes at 37'C followed by 20 minutes at 80'C to inactivate the kinase. Test captures using cell line genomic DNA were then carried out as described below, using the equimolar probe pool at a one thousand-fold probe-to-target molar excess. Based on sequencing results from these test capture reactions, probes were ranked with respect to capture efficiency and the worst-performing ~30% of probes were spiked into the main probe pool at a one hundred-fold relative molar excess.

Capture and library construction

Genomic DNA for HapMap cell-line samples (NA12892 and NA19239) was purchased (Coriell). Clinical specimens consisted of DNA prepared from formalin-fixed paraffin-embedded (FFPE) tissue, peripheral blood, or bone marrow aspirates from patients with sporadic colorectal cancer (n=18), melanoma (7), non-small cell lung cancer (11), bladder cancer (1), ovarian cancer (1), gastrointestinal stromal tumor (4), acute myeloid leukemia (2), and myeloproliferative disorders (3). De-identified residual clinical specimens were obtained from the University of Washington molecular diagnostics laboratory in accordance with the declaration of Helsinki and ethics guidelines of the local institutional review board. Hematoxylin and Eosin-stained slides were used as a guide to manually dissect areas of tumor tissue from unstained slide sections for FFPE tissue samples. Genomic DNA was prepared with the Gentra Puregene DNA Isolation Kit (Qiagen). A 3-hour to overnight proteinase K digestion step was included for FFPE samples. Genomic DNA from the HapMap sample NA12892 was serially diluted two-fold and added to 500 ng of genomic DNA from HapMap samples NA19239 at six relative ratios ranging from 1:8 to 1:256. Captures of the six cell line mixtures, two pure cell line samples, and 47 clinical samples were then performed using ~500 ng of each genomic DNA.

Captures were performed as previously described [160] with some modifications. 500 ng of genomic DNA, 330 femtomoles of probe mixture (ignoring the contribution of the spiked-in poor performers), and 1 microliter of 10X Ampligase DNA ligase buffer (Epicentre) were added to molecular biology-grade water for a total of 10 microliters. For the probe hybridization phase, these mixtures were incubated in a thermocycler (Bio-Rad) with heated lid at 98'C for 3 minutes, 85'C for 30 minutes, 60'C for 60 minutes, and 56'C for 120 minutes. For the gap-fill and ligation phase, we added 300 picomoles each dNTPs (NEB), 7.5 micromoles betaine (Sigma), 20 nanomoles NAD^{+} (NEB), one microliter of 10X Ampligase buffer, 5 units of Ampligase DNA ligase (Epicentre), 3.2 units of Phusion DNA polymerase (NEB), and

molecular biology grade water to 10 microliters for a total reaction volume of 20 microliters. The gap-fill and ligation phase was carried out at 56'C for 60 minutes and 72'C for 20 minutes.

Following the gap-fill and ligation phase, the reactions were cooled to 37'C, and to each reaction we added 20 units of Exonuclease I (NEB) and 100 units of Exonuclease III (NEB) to degrade un-circularized probe and genomic DNA. The digestion was incubated at 37'C for 45 minutes, heated to 80'C, and incubated for 20 minutes to inactivate the exonucleases.

After exonuclease treatment and heat-inactivation, the samples were cooled on ice, and, optionally, stored at -20'C. For each capture reaction, two PCR reactions were prepared, each with Phusion HF buffer to 1X (Fermentas), forward primer and indexed reverse PCR primers to 500 nM, SYBR green (Invitrogen) to 0.5X, dNTPs to 200 micromolar each (NEB), 2 units of Phusion Hot-Start II polymerase, 10 microliters of capture reaction, and nuclease-free water to fifty microliters. PCR cycling conditions were an initial denaturation step for 2 mins at 95'C, followed by 26 cycles of: 15 seconds at 98'C, 15 seconds at 65'C, and 45 seconds at 72'C. A subset of samples was run on a real-time PCR instrument (Bio-Rad MiniOpticon) to estimate the required number of cycles; the remaining samples were run without real-time monitoring (Bio-Rad DNA Engine Tetrad 2).

Library purification and pooling

PCR products were purified individually using Ampure XP beads (Agencourt) at 1.8X according to manufacturers instructions. Purified PCR products were then pooled naively (i.e. equal volumes) for initial quality control or based on MiSeq sequence data considering the number of reads mapping on target per sample. To remove a low molecular-weight artifact, the PCR product pool was split across 8 wells of a 10-well pre-cast 6% polyacrylamide TBE gel (Invitrogen), run at 140V for 50 mins, and stained with 5 microliters of SYBR Gold (Invitrogen). The capture product band at ~280 bp was excised, crushed, soaked in 800 microliters of Tris-EDTA pH 8.0, and recovered from the supernatant using 100 microliters of Ampure XP beads and 700 microliters of home-made Ampure buffer (20% PEG 8000, 2.5M NaCl) according to manufacturers instructions. Pool concentration was assessed using Qubit (Invitrogen).

As an alternative strategy to reduce the time, labor, and cost required for library construction, purified PCR products were subjected to an Ampure-based size enrichment and normalization step. Twenty microliters of each purified PCR product was purified using sixteen microliters of a mixture of one part Ampure XP bead solution and four parts home-made Ampure buffer, and eluted in twenty microliters of Buffer EB (Qiagen). Two microliters of each sample was then run on a diagnostic pre-cast 6% polyacrylamide TBE gel as described above to assess relative concentrations of capture product. The gel image was analyzed for band intensity (ImageJ) and the purified PCR products were pooled according to relative band intensity and the pool was quantified via Qubit.

Sequencing and primary analysis

Samples were sequenced using the HiSeq 2000 and MiSeq (Illumina) platforms according to manufacturer instructions using custom sequencing primers (Table C.7). On the HiSeq platform, we collected two 101 nucleotide reads to determine the sequence of the gap-fill and the molecular tag and one 8 nucleotide read to determine the sequence of the sample index. On the MiSeq platform, we collected two 152 nucleotide reads and one 8 nucleotide read. Initial quality control and capture performance of an equivolume, non-size selected pool of all 55 samples was assessed using one run of the MiSeq platform. Purified PCR products were then repooled according to MiSeq data, size-selected using a PAGE gel as described above, and subjected to 2.75 lanes of HiSeq 2000 sequencing (two lanes with no other samples mixed in and one lane with 25% by moles of an unrelated library). For the

establishment of a rapid workflow, eight samples were processed as described above and subjected to one run of the MiSeq platform.

Read-pairs were assigned to samples requiring an exact match to the expected 8 nucleotide sample index sequence (Table C.7) and the first 12 nucleotides of the reverse read (corresponding to the molecular tag sequence) were stripped out and placed in the header. Read-pairs with molecular tags with homopolymers longer than 4 nucleotides were discarded. Overlapping regions of read-pairs were then reconciled to form single "raw reads" using a custom Smith-Waterman-based strategy. For positions where the read-pairs did not overlap, quality scores from the individual reads were retained. For positions where the read-pairs did overlap, quality scores for the resulting consensus calls were estimated as below for smc-reads. Only successfully overlapped raw reads were retained for downstream analysis; read-pairs that failed to merge were discarded, although subsequent implementations aimed at greater sensitivity towards large insertions such as those found in *FLT3* will retain and analyze these reads, and smMIP does not explicitly require collecting overlapping read-pairs. Raw reads were aligned to the human reference genome (hg19) using the bwasw alignment mode of the aligner *bwa* [208] (v0.5.9) with non-default parameters "-r 1". Based on expected alignment positions according to the probe design, raw reads were then assigned to individual probes, allowing one nucleotide of tolerance in each direction for the beginning of the read, which primarily accommodates insertion and deletion mutations during probe synthesis. Then, for each sample and each probe, raw reads were grouped by molecular tag sequence to form tag-defined read groups (TDRGs).

Single molecule consensus read calling

Alignments to the reference genome were used to call a consensus sequence (*i.e.* a single molecule consensus read or "smc-read") for each TDRG. Positions expected to be derived from probe targeting arms (and therefore synthetic DNA) were excluded from consideration at this step. We adopted a likelihood ratio framework to incorporate both the abundance and associated quality-scores of raw read base-calls supporting each possible nucleotide at each position. Briefly, we calculated the log-likelihood L_x of consensus nucleotide x as the difference of the log-likelihood of a model in which a given nucleotide x was underlying none of the calls and the log-likelihood of a model in which x was underlying all of the calls. This can be represented as in the equation below, where a given observation is associated with a nucleotide call o and a Phred-like quality score Q_o.

$$L_x = \sum_{(o|o \neq x) \in calls} (log_{10}(1 - \frac{10^{\frac{-Q_o}{10}}}{3}) - log_{10}(\frac{10^{\frac{-Q_o}{10}}}{3})) + \sum_{(o|o=x) \in calls} (log_{10}(10^{\frac{-Q_o}{10}}) - log_{10}(1 - 10^{\frac{-Q_o}{10}}))$$

This value (L_x) is computed for each possible consensus call (A, C, G, T, N, and deletion) at each position in the alignment based on the set of base-calls and associated quality scores in the TDRG at that position in the alignment. The consensus call is then determined as the call with the minimum (*i.e.* most negative) log-likelihood value, as this indicates the consensus nucleotide where the model assuming that nucleotide did not underlie any of the raw calls was the least likely relative to the model assuming that nucleotide underlay all of the raw calls. The final "Phred-like" quality score is calculated as the integer casting of $-10*L_x$. For example, in the event that a model against a given consensus nucleotide is 10^{-3} times as likely as a model for a given consensus nucleotide, the associated quality score was calculated to be 30. These scores were capped at 60 as we did not observe substitution rates substantially below $1x10^{-6}$ in practice. We note that a smc-read base-call that is derived from a single Q60 raw read base-call (which may have been derived from two Q40 calls at an overlapping position, for example) will be assigned an estimated quality score of 59 because of the integer casting step. In practice, therefore, only

smc-read base-calls that were supported by at least two independent raw read base-calls can attain quality 60.

For reference positions where at least one read in the alignment indicated a deletion and at least one a match, or where at least one read contained an insertion relative to reference and at least one read lacked that insertion, we applied the same framework, assigning deletion calls a Phred-like quality score of 40. We note that this framework makes simplifying assumptions: that a given nucleotide, in the event of a sequencing error, is equally likely to give rise to any of the other three substitution nucleotides, and that a single nucleotide truly underlies all calls across all reads in the TDRG. Furthermore, for interpretability, we used three as the denominator to distribute the probability when the observed nucleotide was not the candidate consensus nucleotide, though we computed this value for a total of six possible consensus calls and not four.

This strategy was also used to determine quality scores for consensus calls at overlapping positions in raw reads, which represents the simpler case of two and only two calls.

Variant calling and classification

To accommodate variants present across a wide range of frequencies, we adopted a two-pronged variant calling strategy. First, to detect variants present at higher frequencies (*i.e.* approximately 10% or higher), we used alignments of raw reads and smc-reads for each sample individually (i.e. single sample calling) as inputs to the Genome Analysis Toolkit (GATK, v1.6-5-g557da77)[209] variant caller "UnifiedGenotyper" with non-default command line options as follows:

```
-U ALLOW_UNSET_BAM_SORT_ORDER \
--output_mode EMIT_ALL_SITES \
--downsampling_type NONE \
--genotype_likelihoods_model BOTH \
--read_filter BadCigar \
--min_base_quality_score 20"
```

Variants were then filtered using the GATK tool "VariantFiltration" using the following non-default command line options:

```
--filterExpression "QD < 10.0" \
--filterName "LowQD" \
--filterExpression "DP < 30" \
--filterName "LowDP"
```

Variants flagged as "LowDP" were ignored in all analyses. Variants with low "quality-by-depth" according to GATK (*i.e.* QD<10, "LowQD"), were ignored when attempting to discover germline variation (*i.e.* for the un-mixed HapMap samples) and retained when attempting to discover somatic variation. Variants called by GATK were used to determine concordance between smMIP genotypes and 1000 Genomes genotypes, to exclude sites from consideration when quantifying substitution rates, and for the detection of somatic variation in clinical samples. Variants were annotated using the SeattleSeq webserver (http://snp.gs.washington.edu/SeattleSeqAnnotation134/).

To detect low-frequency variation and quantify substitution rates, we adopted a distinct strategy. Alignments of raw reads and smc-reads were considered directly and base-calls at putative homozygous reference sites (according to genotypes called by GATK) were tabulated, considering only very high quality calls (at least Q41 for raw reads, Q60 for smc-reads). We note that some but not all positions in raw reads can attain higher quality scores via the merging process (up to a maximum of 60 as specified by the quality score estimation process described above).

To categorize variation as putative germline or somatic, we performed several filtering steps. First, we obtained a list of sites ("ESP5400") that had been detected as variant in at least one of 5400 exomes sequenced at the University of Washington as part of the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/). Because we observed that some sites in this list were also present in the COSMIC database (CosmicMutantExport_v59_230512.tsv obtained from ftp://ftp.sanger.ac.uk/pub/CGP/cosmic/data_export/), we first filtered the ESP5400 list to remove these positions (for example, *JAK2* p.V617F). Next, variants were categorized as putative germline if they occurred at a site present in the ESP5400 list that had been filtered of COSMIC variant sites. Remaining variant sites were then compared to the COSMIC list and categorized thusly.

Bibliography

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

2 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**, 560-564 (1977).

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**, 5463-5467 (1977).

4 Kasper, T. J., Melera, M., Gozel, P. & Brownlee, R. G. Separation and detection of DNA by capillary electrophoresis. *Journal of chromatography* **458**, 303-312 (1988).

5 Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679, doi:10.1038/321674a0 (1986).

6 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**, 186-194 (1998).

7 Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8**, 175-185 (1998).

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

9 Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-1462 (1997).

10 Sulston, J. E., Waterston, R. H. & Consortium, T. C. e. S. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

11 The International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

12 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:10.1038/nature03959 (2005).

13 Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

14 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

15 Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-352, doi:10.1038/nature10242 (2011).

16 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature biotechnology* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).

17 Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, doi:10.1126/science.1219240 (2012).

18 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).

19 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

20 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology* **30**, 265-270, doi:10.1038/nbt.2136 (2012).

21 Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**, 790-793, doi:10.1038/ng.646 (2010).

22 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250, doi:10.1038/nature10989 (2012).

23 Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46, doi:10.1038/nrg2626.

24 Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature biotechnology*, doi:10.1038/nbt.2288 (2012).

25 Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, doi:10.1038/nbt.2280 (2012).

26 The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).

27 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).
28 Ames, B. N., McCann, J. & Yamasaki, E. Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. *Mutation research* **31**, 347-364 (1975).
29 Ellenberger, J. & Mohn, G. Mutagenic activity of cyclophosphamide, ifosfamide, and trofosfamide in different genes of escherichia coli and salmonella typhimurium after biotransformation through extracts of rodent liver. *Archives of toxicology* **33**, 225-240 (1975).
30 Li, J. *et al.* Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nature medicine* **14**, 579-584, doi:10.1038/nm1708 (2008).
31 Vogelstein, B. & Kinzler, K. W. Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9236-9241 (1999).
32 Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8817-8822, doi:10.1073/pnas.1133470100 (2003).
33 Wu, D. Y., Ugozzoli, L., Pal, B. K. & Wallace, R. B. Allele-specific enzymatic amplification of beta-globin genomic DNA for diagnosis of sickle cell anemia. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 2757-2760 (1989).
34 Newton, C. R. *et al.* Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic acids research* **17**, 2503-2516 (1989).
35 Bielas, J. H. & Loeb, L. A. Quantification of random genomic mutations. *Nature methods* **2**, 285-290, doi:10.1038/nmeth751 (2005).
36 Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nature genetics* **39**, 347-351, doi:10.1038/ng1975 (2007).
37 Ross, P., Hall, L., Smirnov, I. & Haff, L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature biotechnology* **16**, 1347-1351, doi:10.1038/4328 (1998).
38 Luria, S. E. & Delbruck, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491-511 (1943).
39 Kunkel, T. A. The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *The Journal of biological chemistry* **260**, 5787-5796 (1985).
40 Keohavong, P. & Thilly, W. G. Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 9253-9257 (1989).
41 Milbury, C. A., Correll, M., Quackenbush, J., Rubio, R. & Makrigiorgos, G. M. COLD-PCR enrichment of rare cancer mutations prior to targeted amplicon resequencing. *Clinical chemistry* **58**, 580-589, doi:10.1373/clinchem.2011.176198 (2012).
42 Morlan, J., Baker, J. & Sinicropi, D. Mutation detection by real-time PCR: a simple, robust and highly selective method. *PloS one* **4**, e4584, doi:10.1371/journal.pone.0004584 (2009).
43 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).
44 Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-109, doi:10.1126/science.1150427 (2008).
45 The 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534.
46 Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41**, 1061-1067, doi:10.1038/ng.437 (2009).
47 Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods* **6**, 99-103, doi:10.1038/nmeth.1276 (2009).
48 Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research* **19**, 1586-1592, doi:10.1101/gr.092981.109 (2009).
49 Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome research* **22**, 1525-1532, doi:10.1101/gr.138115.112 (2012).
50 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:10.1126/science.1197005.

51 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**, 30-35, doi:10.1038/ng.499 (2010).
52 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).
53 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060, doi:10.1038/nature09710 (2010).
54 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).
55 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).
56 Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469, doi:10.1038/nature07488 (2008).
57 Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479-486, doi:10.1038/nature08170 (2009).
58 Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature biotechnology* **29**, 607-614, doi:10.1038/nbt.1873 (2011).
59 Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).
60 Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature methods* **7**, 741-746, doi:10.1038/nmeth.1492.
61 Bettegowda, C. *et al.* Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science* **333**, 1453-1455, doi:10.1126/science.1210557 (2011).
62 Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506, doi:10.1038/nature11071 (2012).
63 Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-163, doi:10.1038/nature10725 (2012).
64 Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, doi:10.1038/nature10933 (2012).
65 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848, doi:10.1126/science.1162228 (2008).
66 Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373, doi:10.1038/nature09652 (2011).
67 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223, doi:10.1126/science.1168978 (2009).
68 Shendure, J. The beginning of the end for microarrays? *Nature methods* **5**, 585-587, doi:10.1038/nmeth0708-585 (2008).
69 Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity* **100**, 659-674, doi:10.1093/jhered/esp086 (2009).
70 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature reviews. Genetics* **7**, 85-97, doi:10.1038/nrg1767 (2006).
71 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer* **7**, 233-245, doi:10.1038/nrc2091 (2007).
72 Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology* **162**, 729-773 (1982).
73 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics* **32**, 314-331 (1980).
74 Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319-337 (1987).
75 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).
76 Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).

77 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).
78 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
79 Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327, doi:10.1016/j.ygeno.2010.03.001 (2010).
80 Idury, R. M. & Waterman, M. S. A new algorithm for DNA sequence assembly. *Journal of computational biology : a journal of computational molecular cell biology* **2**, 291-306 (1995).
81 Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 9748-9753, doi:10.1073/pnas.171285098 (2001).
82 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).
83 Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome research* **18**, 324-330, doi:10.1101/gr.7088808 (2008).
84 Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**, 810-820, doi:10.1101/gr.7337908 (2008).
85 Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123, doi:10.1101/gr.089532.108 (2009).
86 Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**, 265-272, doi:10.1101/gr.097261.109 (2010).
87 Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and correction of sequencing errors. *Genome biology* **11**, R116, doi:10.1186/gb-2010-11-11-r116 (2010).
88 Chaisson, M. J., Brinza, D. & Pevzner, P. A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research* **19**, 336-346, doi:10.1101/gr.079053.108 (2009).
89 Wu, X. *et al.* Neutralization escape variants of human immunodeficiency virus type 1 are transmitted from mother to infant. *Journal of virology* **80**, 835-844, doi:10.1128/JVI.80.2.835-844.2006 (2006).
90 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5088-5090 (1977).
91 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261-5267, doi:10.1128/AEM.00062-07 (2007).
92 Clarridge, J. E., 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews* **17**, 840-862, table of contents, doi:10.1128/CMR.17.4.840-862.2004 (2004).
93 Kalyuzhnaya, M. G. *et al.* High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature biotechnology* **26**, 1029-1034, doi:10.1038/nbt.1488 (2008).
94 Alcami, A. & Koszinowski, U. H. Viral mechanisms of immune evasion. *Trends in microbiology* **8**, 410-418 (2000).
95 Hahn, B. H. *et al.* Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* **232**, 1548-1553 (1986).
96 Mansky, L. M. & Temin, H. M. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology* **69**, 5087-5094 (1995).
97 McMichael, A. J. & Phillips, R. E. Escape of human immunodeficiency virus from immune control. *Annual review of immunology* **15**, 271-296, doi:10.1146/annurev.immunol.15.1.271 (1997).
98 Henn, M. R. *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* **8**, e1002529, doi:10.1371/journal.ppat.1002529 (2012).
99 Macalalad, A. R. *et al.* Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS computational biology* **8**, e1002417, doi:10.1371/journal.pcbi.1002417 (2012).

100 Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome biology* **12**, 227, doi:10.1186/gb-2011-12-9-227 (2011).
101 Myers, R. M., Tilly, K. & Maniatis, T. Fine structure genetic analysis of a beta-globin promoter. *Science* **232**, 613-618 (1986).
102 Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081-1085 (1989).
103 Farber, S. & Diamond, L. K. Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid. *The New England journal of medicine* **238**, 787-793, doi:10.1056/NEJM194806032382301 (1948).
104 Mukherjee, S. *The emperor of all maladies : a biography of cancer*. Large print edn, (Thorndike Press, 2010).
105 Jordan, V. C. Fourteenth Gaddum Memorial Lecture. A current view of tamoxifen for the treatment and prevention of breast cancer. *British journal of pharmacology* **110**, 507-517 (1993).
106 Druker, B. J. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nature medicine* **2**, 561-566 (1996).
107 National Cancer Institute. *Targeted Cancer Therapies*, <http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted> (2012).
108 Worley, L. A. *et al.* Transcriptomic versus chromosomal prognostic markers and clinical outcome in uveal melanoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 1466-1471, doi:10.1158/1078-0432.CCR-06-2401 (2007).
109 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).
110 Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS medicine* **2**, e73, doi:10.1371/journal.pmed.0020073 (2005).
111 Diaz Jr, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, doi:10.1038/nature11219 (2012).
112 Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, doi:10.1038/nature11156 (2012).
113 MacConaill, L. E. *et al.* Profiling critical cancer gene mutations in clinical tumor samples. *PloS one* **4**, e7887, doi:10.1371/journal.pone.0007887 (2009).
114 Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in C. elegans. *Nature methods* **5**, 183-188, doi:10.1038/nmeth.1179 (2008).
115 Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research* **19**, 1141-1152, doi:10.1101/gr.085464.108 (2009).
116 Weinstein, J. A., Jiang, N., White, R. A., 3rd, Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807-810, doi:10.1126/science.1170020 (2009).
117 Bentley, G. *et al.* High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue antigens* **74**, 393-403, doi:10.1111/j.1399-0039.2009.01345.x (2009).
118 Stover, C. K. *et al.* Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature* **406**, 959-964, doi:10.1038/35023079 (2000).
119 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
120 Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).
121 Reinhardt, J. A. *et al.* De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. *Genome research* **19**, 294-305, doi:10.1101/gr.083311.108 (2009).
122 Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods* **6**, 673-676, doi:10.1038/nmeth.1358 (2009).
123 Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res* **27**, 2369-2376 (1999).
124 Sabourin, J. C. *et al.* An intronic enhancer essential for tissue-specific expression of the aldolase B transgenes. *J Biol Chem* **271**, 3469-3473 (1996).

125 Gregori, C. *et al.* Expression of the rat aldolase B gene: a liver-specific proximal promoter and an intronic activator. *Biochem Biophys Res Commun* **176**, 722-729 (1991).
126 Gregori, C., Porteu, A., Mitchell, C., Kahn, A. & Pichard, A. L. In vivo functional characterization of the aldolase B gene enhancer. *J Biol Chem* **277**, 28618-28623, doi:10.1074/jbc.M204047200 (2002).
127 Kim, M. J. *et al.* Functional characterization of liver enhancers that regulate drug-associated transporters. *Clinical pharmacology and therapeutics* **89**, 571-578, doi:10.1038/clpt.2010.353 (2011).
128 Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature methods* **7**, 119-122, doi:10.1038/nmeth.1416 (2010).
129 Zhang, G., Budker, V. & Wolff, J. A. High levels of foreign gene expression in hepatocytes after tail vein injections of naked plasmid DNA. *Human gene therapy* **10**, 1735-1737, doi:10.1089/10430349950017734 (1999).
130 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816, doi:10.1038/nature05874 (2007).
131 Kel, A. E. *et al.* MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31**, 3576-3579 (2003).
132 Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036-1040, doi:10.1126/science.1186176 (2010).
133 Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-140 (2000).
134 Margulies, E. H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome research* **17**, 760-774, doi:10.1101/gr.6034307 (2007).
135 Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**, 158-160, doi:10.1038/ng.2007.55 (2008).
136 Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-810, doi:10.1038/ng.650 (2010).
137 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**, 901-913, doi:10.1101/gr.3577405 (2005).
138 A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
139 Botstein, D. & Shortle, D. Strategies and applications of in vitro mutagenesis. *Science* **229**, 1193-1201 (1985).
140 Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science translational medicine* **3**, 111ra121, doi:10.1126/scitranslmed.3003161 (2011).
141 Wagle, N. *et al.* High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer discovery* **2**, 82-93, doi:10.1158/2159-8290.CD-11-0184 (2012).
142 Lipson, D. *et al.* Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nature medicine* **18**, 382-384, doi:10.1038/nm.2673 (2012).
143 Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature biotechnology* **27**, 1025-1031, doi:10.1038/nbt.1583 (2009).
144 Harismendy, O. *et al.* Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome biology* **12**, R124, doi:10.1186/gb-2011-12-12-r124 (2011).
145 Forshew, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Science translational medicine* **4**, 136ra168, doi:10.1126/scitranslmed.3003726 (2012).
146 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94, doi:10.1038/nature09807 (2011).
147 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

148 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).
149 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).
150 Druley, T. E. *et al.* Quantification of rare allelic variants from pooled genomic DNA. *Nature methods* **6**, 263-265, doi:10.1038/nmeth.1307 (2009).
151 Flaherty, P. *et al.* Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic acids research* **40**, e2, doi:10.1093/nar/gkr861 (2012).
152 Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature communications* **3**, 811, doi:10.1038/ncomms1814 (2012).
153 Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic acids research* **39**, e81, doi:10.1093/nar/gkr217 (2011).
154 Fu, G. K., Hu, J., Wang, P. H. & Fodor, S. P. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9026-9031, doi:10.1073/pnas.1017621108 (2011).
155 Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 20166-20171, doi:10.1073/pnas.1110064108 (2011).
156 Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9530-9535, doi:10.1073/pnas.1105422108 (2011).
157 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods* **9**, 72-74, doi:10.1038/nmeth.1778 (2012).
158 Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1347-1352, doi:10.1073/pnas.1118018109 (2012).
159 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature methods* **6**, 315-316, doi:10.1038/nmeth.f.248 (2009).
160 Shen, P. *et al.* High-quality DNA sequence capture of 524 disease candidate genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6549-6554, doi:10.1073/pnas.1018981108 (2011).
161 Durbin, R. M. & Consortium, T. G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).
162 NHLBI Exome Sequencing Project (ESP). (2012).
163 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**, D945-950, doi:10.1093/nar/gkq929 (2011).
164 Lin, J. *et al.* Polyclonality of BRAF mutations in acquired melanocytic nevi. *Journal of the National Cancer Institute* **101**, 1423-1427, doi:10.1093/jnci/djp309 (2009).
165 Lin, J. *et al.* Polyclonality of BRAF mutations in primary melanoma and the selection of mutant alleles during progression. *British journal of cancer* **104**, 464-468, doi:10.1038/sj.bjc.6606072 (2011).
166 Yancovitz, M. *et al.* Intra- and inter-tumor heterogeneity of BRAF(V600E))mutations in primary and metastatic melanoma. *PloS one* **7**, e29336, doi:10.1371/journal.pone.0029336 (2012).
167 Vaughn, C. P., Zobell, S. D., Furtado, L. V., Baker, C. L. & Samowitz, W. S. Frequency of KRAS, BRAF, and NRAS mutations in colorectal cancer. *Genes, chromosomes & cancer* **50**, 307-312, doi:10.1002/gcc.20854 (2011).
168 Maughan, T. S. *et al.* Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet* **377**, 2103-2114, doi:10.1016/S0140-6736(11)60613-2 (2011).
169 De Roock, W. *et al.* Association of KRAS p.G13D mutation with outcome in patients with chemotherapy-refractory metastatic colorectal cancer treated with cetuximab. *JAMA : the journal of the American Medical Association* **304**, 1812-1820, doi:10.1001/jama.2010.1535 (2010).

170 Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nature medicine* **14**, 985-990, doi:10.1038/nm.1789 (2008).
171 Kitzman, J. O. *et al.* Noninvasive whole-genome sequencing of a human fetus. *Science translational medicine* **4**, 137ra176, doi:10.1126/scitranslmed.3004323 (2012).
172 Fan, H. C. *et al.* Non-invasive prenatal measurement of the fetal genome. *Nature*, doi:10.1038/nature11251 (2012).
173 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1208715109 (2012).
174 Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. *Rapid and sensitive multiplex sequencing of actionable genes in clinical cancer samples* (In submission, 2012).
175 Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research* **21**, 2224-2241, doi:10.1101/gr.126599.111 (2011).
176 Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* **7**, 709-715, doi:10.1038/nmeth.1491 (2010).
177 Goodbourn, S. & Maniatis, T. Overlapping positive and negative regulatory domains of the human beta-interferon gene. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 1447-1451 (1988).
178 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* **30**, 271-277, doi:10.1038/nbt.2137 (2012).
179 Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**, 521-530, doi:10.1038/nbt.2205 (2012).
180 Hoffmann, C. *et al.* DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic acids research* **35**, e91, doi:10.1093/nar/gkm435 (2007).
181 Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M. & Shafer, R. W. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome research* **17**, 1195-1201, doi:10.1101/gr.6468307 (2007).
182 Eriksson, N. *et al.* Viral population estimation using pyrosequencing. *PLoS computational biology* **4**, e1000074, doi:10.1371/journal.pcbi.1000074 (2008).
183 Rozera, G. *et al.* Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* **6**, 15, doi:10.1186/1742-4690-6-15 (2009).
184 Archer, J. *et al.* The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time--an ultra-deep approach. *PLoS computational biology* **6**, e1001022, doi:10.1371/journal.pcbi.1001022 (2010).
185 Hedskog, C. *et al.* Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PloS one* **5**, e11345, doi:10.1371/journal.pone.0011345 (2010).
186 Zagordi, O., Geyrhofer, L., Roth, V. & Beerenwinkel, N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology : a journal of computational molecular cell biology* **17**, 417-428, doi:10.1089/cmb.2009.0164 (2010).
187 Zagordi, O., Klein, R., Daumer, M. & Beerenwinkel, N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic acids research* **38**, 7400-7409, doi:10.1093/nar/gkq655 (2010).
188 Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing noise from pyrosequenced amplicons. *BMC bioinformatics* **12**, 38, doi:10.1186/1471-2105-12-38 (2011).
189 Zagordi, O., Bhattacharya, A., Eriksson, N. & Beerenwinkel, N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics* **12**, 119, doi:10.1186/1471-2105-12-119 (2011).
190 Prosperi, M. C. & Salemi, M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* **28**, 132-133, doi:10.1093/bioinformatics/btr627 (2012).
191 Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome research* **18**, 763-770, doi:10.1101/gr.070227.107 (2008).

192 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).
193 Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome research* **19**, 1124-1132, doi:10.1101/gr.088013.108 (2009).
194 Malhis, N., Butterfield, Y. S., Ester, M. & Jones, S. J. Slider--maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics* **25**, 6-13, doi:10.1093/bioinformatics/btn565 (2009).
195 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).
196 Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285, doi:10.1093/bioinformatics/btp373 (2009).
197 Bansal, V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* **26**, i318-324, doi:10.1093/bioinformatics/btq214 (2010).
198 Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**, 730-736, doi:10.1093/bioinformatics/btq040 (2010).
199 Vallania, F. L. *et al.* High-throughput discovery of rare insertions and deletions in large cohorts. *Genome research* **20**, 1711-1718, doi:10.1101/gr.109157.110 (2010).
200 Altmann, A. *et al.* vipR: variant identification in pooled DNA using R. *Bioinformatics* **27**, i77-84, doi:10.1093/bioinformatics/btr205 (2011).
201 Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology* **30**, 434-439, doi:10.1038/nbt.2198 (2012).
202 Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature biotechnology* **29**, 59-63, doi:10.1038/nbt.1740 (2011).
203 Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-195, doi:10.1038/nature11236 (2012).
204 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
205 Shibutani, S., Takeshita, M. & Grollman, A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349**, 431-434, doi:10.1038/349431a0 (1991).
206 Lasken, R. S., Schuster, D. M. & Rashtchian, A. Archaebacterial DNA polymerases tightly bind uracil-containing DNA. *The Journal of biological chemistry* **271**, 17692-17696 (1996).
207 Greagg, M. A. *et al.* A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9045-9050 (1999).
208 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
209 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).