Discovery and interpretation of genetic variation
generation sequencing technologies

Author: Aaron Ryan Quinlan

Boston College

The Graduate School of Arts and Sciences

Department of Biology

DISCOVERY AND INTERPRETATION OF GENETIC VARIATION WITH NEXT-GENERATION SEQUENCING TECHNOLOGIES

A dissertation

by

Aaron Ryan Quinlan

submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

May 2008

Abstract

Discovery and interpretation of genetic variation with next-generation sequencing technologies

Aaron Ryan Quinlan

Dissertation advisor: Gabor T. Marth

Improvements in molecular and computational technologies have driven and will continue to drive advances in our understanding of genetic variation and its relationship to phenotypic diversity. Over the last three years, several new DNA sequencing technologies have been developed that greatly improve upon the cost and throughput of the capillary DNA sequencing technologies that were used to sequence the first human genome. The economy of these so-called "next-generation" technologies has enabled researchers to conduct genome-wide studies in genetic variation that were previously intractable or too expensive.

However, because the new technologies employ novel molecular techniques, the resulting sequence data is quite different from the capillary sequences to which the genomics field is accustomed. Moreover, the vast amounts of sequence data that these technologies produce present novel statistical and computational challenges in order to make even the simplest observations. The focus of my dissertation has been the development of novel computational and analytical methods that facilitate genome-wide studies in genetic variation with traditional capillary sequencers and with new sequencing technologies. I present a novel method that produces more accurate error estimates for sequence data from one of these next-generation sequencing technologies. I also present two studies that illustrate the utility of two such technologies for genome-wide polymorphism discovery studies in *Drosophila melanogaster* and *Caenorhabditis elegans*. These studies accurately estimate the degree of genetic diversity in the fruitfly and nematode, respectively. I later describe how new sequencing approaches can be used to accelerate the mapping of causal genetic mutations in forward genetic screens. Lastly, I remark on where I believe these technologies will lead future studies in human genetic variation and describe their relevance to several of my future research interests.

Acknowledgements

I am forever indebted to my mentor, Gabor Marth. I had the great fortune of being the first graduate student in Gabor's laboratory. This allowed me to work side by side with Gabor from the beginning. I am thankful for his ambition, critical mind, and mostly for his irreverent sense of humor. It is impossible to enumerate everything that I have learned from you. Thank you Gabor, I hope we are friends until the end.

Katie. You're my wife, my friend, and my partner in crime. You have been my biggest fan and you've allowed me to work long nights and weekends while remaining patient, loving and supportive. I'd be lost without you. This is for you and for Madeleine, our chef-d'oeuvre.

My parents, Paula, Pete, and Dan, taught me to be curious, hard-working and kind. Mom, I'll never forget memorizing all the bones in the human body with you in sixth grade. I love you for being the incredible mother you are. You taught me to be compassionate, kind and hard-working. Dad, you always made me believe I was the best even when I was far from it. You're gone, but what you taught me is not. Pete, you are a model husband and a model human. Thanks for always being patient and for being a great friend, father and husband to mom.

I thank Chip Stewart for his efficient mind, patient demeanor and his willingness to help me. I thank Michael Stromberg for the work we've done together and for all he has taught me about software development. I am also indebted to Peter Clote for his part in the Presidential Fellowship that I received from Boston College. Lastly, I thank Dr. Mitch Chernin for being the best professor I've ever known and for teaching a computer scientist a little molecular biology.

There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

-Mark Twain

1. Introduction

The 19th century saw both Darwin and Wallace posit the then-novel theory of evolution, describing the forces that cause speciation via the creation and elimination of genetic traits. Separately, but during the same period, Gregor Mendel's work with pea plants established a mathematical framework for understanding the relative frequencies of unlinked genetic traits. Collectively, the work of Darwin, Wallace and Mendel established a foundation for future research investigating the evolution of species, the causes and degree of genetic diversity within populations, and the genetic profiles that predispose individuals to disease.

The one hundred and fifty years since have been host to hundreds of groundbreaking experiments that have formed our current understanding of genetic diversity. Using the fruitfly *Drosophila melanogaster* as a model, T.H. Morgan demonstrated that genes are found on chromosomes, providing a molecular mechanism for inheritance. Avery, McCarty and MacLeod later established that DNA, not proteins, carries genetic information. Soon after, Watson, Crick, Wilkins and Franklin deciphered the structure of DNA, and in so doing, elucidated the mechanism of DNA replication and inheritance. By 1977, both Frederick Sanger and Walter Gilbert had developed reliable, albeit painstaking, methods for sequencing DNA. These methods led to discovery of the first known gene sequences and the first sequenced genome: Sanger's laborious sequencing of the bacteriophage ΦX174 genome.

Subsequent technological advances such as the miniaturization of electrophoresis, the use of fluorescently-labeled nucleotides in the Sanger reaction, and Kary Mullins' ingenious vision of the polymerase chain reaction, enabled research departments and smaller laboratories to reliably sequence large sections of DNA from human and model genetic organisms with increasing economy. Ultimately, DNA sequencing technology progressed to the point where the complete genome sequencing of well-studied model organisms (Adams et al. 2000; The *C. elegans* Sequencing Consortium 1998) and the completion of a high-quality draft of the human genome were possible by the turn of the millennium (Lander et al. 2001; Venter et al. 2001). These complete genome sequences provide a point of reference to which one can compare genetic information from other individuals of the same species. In principle, comparisons of this sort can, when performed among many individuals from the same species, describe the landscape of genetic variation within that species and potentially identify those variations that predispose certain individuals to disease.

The Human Genome Project and the research that has followed have estimated that the genomes of any two humans are at least 99.5% identical (Lander et al. 2001; Levy et al. 2007; The International HapMap Consortium 2005; Venter et al. 2001). Moreover, comparative estimates resulting from the same studies place the similarity of any human and any chimpanzee genome at roughly 99%. While such similarity may possibly seem high, even a 0.1% difference allows for over three million genetic differences between any two humans. Clearly, this degree of genetic diversity has the potential to explain, in large part, the breadth of phenotypic diversity present in the human population. A major impetus behind the Human Genome Project was the assumption that one could correlate observed genetic differences with disease phenotypes and thus facilitate and expedite the understanding and treatment of disease. As part of the comparative analyses conducted during the Human Genome Project, over 1.4 million human single-nucleotide polymorphisms (SNPs) were discovered (Sachidanandam et al. 2001). These variations provided the initial foundation for several preliminary investigations into disease predisposition and catalyzed subsequent research to describe the extent of human genetic variation.

Single-nucleotide polymorphisms

Single-nucleotide polymorphisms have been by far the most widely studied type of genetic variation (Altshuler et al. 2000; Braverman et al. 1995; Collins et al. 1998a; Collins et al. 1998b; Freedman et al. 2004; Fu 1995; Gibbs 2003; Goldstein and Weale 2001; Ke et al. 2004; Kwok and Chen 1998; Kwok and Gu 1999; Marnellos 2003; Marth et al. 2001; Marth et al. 1999; Mullikin et al. 2000; Sachidanandam et al. 2001; Sherry et al. 1999; Sunyaev et al. 2000; Tajima 1989; Weber 1990). SNPs are of such broad interest for several reasons. First, there are several diseases such as sickle-cell anemia where specific polymorphic alleles are highly-correlated or directly causal for the disease phenotype (O'Donald 1967; Packard et al. 2007). Second, they are quite common in the human genome, where even conservative estimates place the pair-wise SNP rate between any two humans at approximately one SNP per every thousand base pairs, on average

(Chakravarti 1999). Third, they are comparatively simple to identify and experimentally validate (Marth et al. 1999; Stephens et al. 2006; Zhang et al. 2005). Finally, they serve as a genetic landmark with which to track the evolution of species (Marth et al. 2003).

We typically classify SNPs by type as transitions: nucleotide changes from purines to purines or pyrimidines to pyrimidines, and transversions: nucleotide changes from purines to pyrimidines or vice versa. While transitions theoretically represent only one-third of all possible polymorphisms, they have been observed to represent two-thirds of all actual polymorphisms (Petrov and Hartl 1999; Zhang and Gerstein 2003; Zhao and Boerwinkle 2002). This phenomenon is attributable to the high rate of mutation from 5-methyl cytosine to thymine through deamination. SNPs are often further classified by function, according to how they affect the function of protein-coding genes. Whereas “synonymous” SNPs do not alter the encoded amino acid despite the nucleotide change, “non-synonymous” SNPs do. Moreover, “nonsense” mutations cause the codon to become a stop codon, thus introducing a premature end to the encoded protein. SNPs that do not occur in gene coding sequences have typically not been classified by function. However, recent studies have illustrated that the canonical “beads on a string” paradigm of genome organization may oversimplify the complexity of genome expression (Birney et al. 2007). Consequently, it is likely that polymorphisms lying outside of gene coding regions may in fact alter the regulatory function of small inhibitory RNAs, as well as the promoters, introns, and untranslated exons of protein-coding genes (He et al. 2007; Ibarra et al. 2007; Kim et al. 2007).

SNPs have been shown to be non-randomly distributed throughout the human genome (Altshuler et al. 2000; Mullikin et al. 2000; Sachidanandam et al. 2001; Venter et al. 2001) (**Figure 1.1**). This is perhaps not surprising, given our understanding of the forces of selection and the mechanisms of recombination. One expects that highly conserved genes would be under purifying selective pressure that would reduce novel alleles (and thus polymorphism) in genes that are crucial to, for example, cell function (i.e., cell structure genes, ribosomal genes, etc.). A corollary of this hypothesis is that genes that benefit from allelic diversity (e.g., genes involved in immune system response) would be under positive selection to allow for the creation of beneficial alleles. Of equal interest is the variability of regional SNP density in non-protein coding DNA in the genome, as such estimates provide a proxy for the degree of linkage disequilibrium (and thus recombination) throughout a genome. Studies to date have shown that SNP densities vary substantially across the genome and that increased SNP density is correlated with increased recombination rates (The International HapMap Consortium 2005). Similarly, it has also been observed that the human genome is defined by frequent and often large haplotype blocks where the same SNP alleles are observed to be in linkage disequilibrium (LD) (Daly et al. 2001; Reich et al. 2001). In other words, in regions of high LD, SNP alleles are not randomly assorted as they would be were there no local linkage. Thus, the allele present for one SNP in a given haplotype block is predictive of another SNP allele in the same haplotype block (**Figure 1.2**). This means that one can theoretically use SNPs as markers that predict the local haplotype of an individual and can thus help to reduce the complexity of determining their genotype.

For these reasons, the seven years following the completion of the human genome sequence have seen substantial efforts to catalog all of the common polymorphisms in the human genome. As part of this process, there has been a concerted effort to improve the molecular and computational methods used to discover and genotype SNPs.

Insertion-deletion polymorphisms

Insertion-deletion polymorphisms (so called INDELs) are small insertions or deletions in the DNA of one individual or chromosome relative to the DNA of another individual or chromosome. INDELs are traditionally restricted to short (less than or equal to twenty base pairs) insertions or deletions and have recently been found to occur much more frequently in the human genome than previously thought (Levy et al. 2007). In coding regions, INDELs with lengths that are not a multiple of three cause frameshift mutations in the DNA and thus may alter the resulting protein. INDELS that are multiples of three can either add or remove entire codons, which also alter the coding sequence. One such famous mutation is the deletion of a TTT codon (phenylalanine) in the CFTR gene in humans, which, if homozygous, is a primary cause of cystic fibrosis (Audrezet et al. 2004). Short INDEL polymorphisms of one or two nucleotides are traditionally very difficult to identify because of a lack of sufficient nucleotide resolution in Sanger capillary sequences. Therefore, previous studies have been fraught with high false discovery rates (Weber et al. 2002). Given that short (length less than or equal to four) INDELs occur substantially more frequently than longer INDELS (**Figure 1.3**) and that such polymorphisms are the second most frequent polymorphism in the human genome (representing roughly half as many nucleotide differences as SNPs) (Clark et al. 2007), there is great interest in the development of reliable methods for short INDEL detection. Moreover, INDELs and SNPs are similarly distributed throughout the human genome; therefore, they could theoretically be used as genetic markers in a manner similar to SNPs. The advent of several new sequencing technologies with improved nucleotide resolution have illustrated that improved INDEL discovery is possible. The results achieved with such technologies are discussed in Chapter 5.

Copy-number and structural variation

Copy-number and structural variations include small to large insertions and deletions in one or both chromosomes, as well as chromosomal rearrangements such as translocations and inversions. Prior to two seminal studies reported in 2004, the frequency and complexity of copy-number variation (CNV) and structural variation (SV) in the human genome was drastically under-appreciated (Iafrate et al. 2004; Sebat et al. 2004). This is somewhat surprising given that, for many years, it was known that large-scale genomic deletions, duplications, and inversions were observed in many diseases such as cancers, trisomy syndromes, and Prader-Willi syndrome. However, previous studies of so called 'genomic disorders' were limited to what could be detected with very low-resolution methods such as microscopy, karyotyping and fluorescent *in situ* hybridization (FISH). As **Figure 1.4** illustrates, the size range of CNVs is extremely broad, and therefore, the methods used to detect them must ideally be suitable across the same spectrum.

The advent of higher resolution technologies such as array-CGH (comparative genomic hybridization), representational oligonucleotide microarray analysis (ROMA) and dense SNP genotyping chips have allowed researchers to detect CNVs (henceforth referred to solely as CNV) at a minimum of ten kilobases. Collectively, studies using these technologies have provided substantial evidence that CNVs amount to at least 4 Mb of genetic difference between any two humans (Carter 2007; Conrad et al. 2006; Fiegler et al. 2006; Redon et al. 2006). Other, less conservative studies place this estimate somewhere between 5 and 24 Mb (Redon et al. 2006). Thus using even the most conservative estimates, CNV likely accounts for at least as much overall genetic variation in the human genome as SNPs.

Given the extent of CNV observed with predominantly chip-based methods, it is likely that we will observe even more such variation with higher resolution methods such as high-throughput sequencing. A recent study with the 454 Life Sciences sequencing technology suggests that despite existing methodological challenges, the new, high-throughput sequencing technologies are well-suited to large-scale CNV discovery (Korbel et al. 2007). In 2007, Jonathan Sebat and colleagues suggested that autism is highly correlated with frequent *de novo* CNV (that is, occurring in one or both parental gametes) changes (Sebat et al. 2007). This observation further illustrates the need for efficient and comprehensive CNV detection technologies in order to discern any potentially systematic character of *de novo* CNV mutations in such diseases. Moreover, this and other studies indicate that the impact of CNV on human disease has been underappreciated.

Genetic variation: why should we care?

A primary motivation behind the study of genetic variation is the assumption that most phenotypes, disease-related or otherwise, can be attributed either to a single, rare allele (i.e., Mendelian traits) or to the combined effects of multiple alleles. Indeed, many researchers assume that there are sets of common alleles that are highly correlated with various common diseases: the so-called “common-disease, common-variant” hypothesis. If this hypothesis is valid, correlated variations can be used as markers for disease detection. It is also assumed that either such variations must themselves be the causal variations or they must be linked to the causal variations. If so, they can therefore be used to understand disease etiology and drive the development of appropriate therapies.

Large-scale studies founded upon these assumptions require dense variation maps of the entire human genome in order to compare the genetic profiles of healthy and diseased individuals. By 2001, there were nearly 1.5 million known SNPs in the human genome, yielding, on average, one marker every 2,000 base pairs (Sachidanandam et al. 2001). Because of linkage, we know that the closer any two markers are to one another, the less likely it is that recombination will occur between the two markers. Thus, the alleles observed at any two unlinked markers should be random, whereas the alleles observed at linked loci should be relatively consistent. As previously mentioned, the degree of this consistency is known as linkage disequilibrium. SNP markers that are “in LD” can serve as proxies for one another (that is, the allele of one marker is highly predictive of the allele of a second marker) while ostensibly still elucidating relationship between disease phenotype and the underlying genotype (Daly et al. 2001; Reich et al. 2001).

These were the motivations behind The International HapMap Project (Frazer et al. 2007; The International HapMap Consortium 2003; The International HapMap Consortium 2005), which sought to describe the structure and degree of linkage disequilibrium in the human genome among four representative human sub-populations (Utah residents with ancestry from Northern and Western Europe, Yorubans from Nigeria, Japanese in Tokyo and Han Chinese from Beijing) totaling 270 presumably healthy individuals. Owing to the high cost of complete human genome resequencing and the comparatively low cost of chip-based SNP genotyping experiments, it was thought that researchers could exploit the haplotype block structure of the human genome to reduce the complexity (and cost) of large-scale disease association studies. It was believed that one could use the genotype for a smaller set of SNPs to extrapolate the genotype of nearly the entire genome, given the extent of linkage disequilibrium observed. Hypothetically, such a reduction in complexity would enable the statistical comparison of the genotypes of large panels of individuals with a disease ("cases") and those without the disease ("controls") in order to elucidate those alleles that are highly correlated with a given disease of interest. This hypothesis assumes that: a) the disease in question is largely caused by a combination of common polymorphisms, and b) nearly all of the genetic causes of a given disease are detectable by SNP alleles. To date, few disease association studies have identified sets of alleles that describe more than 5% of the risk for a given disease.

Several factors are likely involved in the relative lack of success of most association studies thus far. One likely problem is that association studies that use SNP markers from the HapMap are limited by the fact that the HapMap is restricted to common polymorphisms having a minor allele frequency (that is, the frequency of the less common SNP allele among observed chromosomes) of at least 5%. Therefore, the ability to detect any disease associations with combinations of less common alleles is diminished in these studies. For this reason, beginning in 2008, a new international effort known as "The 1000 Genomes Project" will seek to uncover all SNPs with a minor allele frequency as low as 1% in the entire genome and as low as 0.1% in protein coding regions. Current association studies may also be limited in power by the fact that only

SNP markers are used to detect disease correlations. The HapMap project accurately described the patterns of linkage disequilibrium among SNPs in the genomes of the 270 individuals studied. However, it is unclear to what degree SNPs are in LD with other polymorphisms. If there is less LD between SNPs and INDELs or other structural variations, then association studies that focus solely on SNP markers may be blind to existing associations to non-SNP markers. Lastly, poorly defined qualitative phenotypes or quantitative phenotypes that are classified by highly variable assays may also perturb disease association studies. For example, imagine an association study that seeks to uncover alleles that are associated with hypercholestoremia in a cohort of 500 cases and controls. Existing associations could potentially be weakened by variability in the classification of hypercholestoremia, and the assays used to measure cholesterol levels, as well as environmental contributions to the disease phenotype. For these reasons, accurate and quantitative patient records will likely improve the power of association studies.

A more complete understanding of the degree and types of genetic variation among healthy and diseased humans will undoubtedly improve our understanding of the genetic mechanisms of disease predisposition. As DNA sequencing and probe-based technologies improve and their costs continue to decline, more and more investigators will be able exploit the technologies for large scale studies of genetic variation. Yet as the throughput and economy of these technologies increase, so does the burden placed upon reliable computational methods for the analysis and discovery of genetic variation. While it is feasible to manually inspect and validate polymorphisms discovered in a handful of loci, it is impossible to confirm manually the thousands of variations that will inevitably come out of large-scale, whole genome studies in the near future. Thus, the development of highly-accurate computational methods must keep in step with the molecular methods they are designed to analyze.

Methods for polymorphism discovery with capillary sequencing technologies

The complete (or nearly so) genome sequences of humans and other model organisms provide frames of reference to which sequences from individuals of the same species can be compared. Such comparisons facilitate the identification of single-nucleotide and insertion-deletion polymorphisms relative to the reference genome sequences (**Figure 1.5**). This is the basic approach employed by the majority of the commonly used polymorphism discovery methods (Marth et al. 1999; Stephens et al. 2006; Zhang et al. 2005). These methods typically differ in the way they attempt to segregate variation caused by sequencing error from *bona fide* genetic variation. The accuracy of differentiating sequencing errors from true variation is highly dependent on the accuracy of the "base calling program". Base calling programs such as *Phred* (Ewing and Green 1998; Ewing et al. 1998) interpret the raw sequencing read output from capillary DNA sequencing machines (e.g., Applied Biosystems 3730) and convert them into a called nucleotide sequence (e.g., 5'-AACTGGCATT-3'). In addition, base calling programs estimate, for each of the bases they call, the likelihood that the call was, in fact, wrong. These estimates are generally referred to as base quality values and typically conform to the *Phred* base quality paradigm defined by Phil Green and colleagues. In this framework, a quality value (Q) is defined as $Q = -10 * \log_{10} (P)$, where P is the probability that the given base was called in error. For example, if a base calling program estimates that the error likelihood for a called base is 0.1, then the base quality value Q would be 10. Similarly, if the error likelihood were estimated to be 0.01, the base quality value would be 20. Thus, higher base quality values indicate bases that were called with greater confidence. Accurate base calls and base quality values are essential for polymorphism discovery, as they are the usually the only means to distinguish true polymorphism from sequencing error.

Earlier polymorphism discovery projects screened for polymorphisms among protein-coding sequences by creating cDNA libraries from mature mRNA sequences. These cDNA clones (which are inherently from a single chromosome) were sequenced with capillary sequencing technologies and the sequencing reads were base-called with *Phred*. The sequence reads were then aligned to a reference genome sequence and screened for polymorphisms by software such as *PolyBayes* (Marth et al. 1999). Because the aligned alleles are from a single chromosome, the polymorphism discovery software seeks to determine whether or not there is sufficient evidence of polymorphism among the aligned alleles and quality values. Similarly, the vast majority of SNPs known today were discovered via whole genome shotgun sequencing protocols, which compared haploid sequence reads to the human reference sequence and to each other for detecting alternate alleles.

However, more recent studies aimed toward discerning the complete genotype of an individual in a genomic region of interest (e.g., a gene) involve the targeted sequencing of such regions via PCR. In such cases, PCR primer-pairs are designed to amplify the region of interest, and the resulting amplicons are sequenced on capillary sequencing machines. Since PCR inherently amplifies the DNA on both chromosomes (in a diploid organism), the resulting sequencing reads reflect the alleles present on both chromosomes. Therefore, homozygous base pairs appear as a single peak, which reflects the fact that the individual has the same allele on both chromosomes. In contrast, heterozygotes should ideally manifest as two peaks for the same base pair, where each peak is roughly the same height (assuming unbiased PCR amplification) and each peak is approximately 50% of the height of a homozygous base pair. Because of the limitations of capillary sequencing machines, it is often difficult to determine whether two observed sequencing peaks at a given locus reflect true heterozygosity or whether they are merely a sequencing artifact. Consequently, early attempts at identifying heterozygotes in such sequence reads were fraught with high error rates. Chapter 2 discusses my research into an improved method for detecting heterozygotes in these reads using a machine learning approach.

The reliable discovery of heterozygotes using PCR amplicon sequencing is further impeded by previously unknown heterozygosity within chromosomal sequences homologous to the PCR primers. In such cases, the chromosome that exactly-matches the PCR primer will be preferentially amplified relative to the imperfectlymatched chromosome. As a result, the expected 50/50 ratio for the two heterozygous alleles is frequently affected (more specifically, it is greatly skewed in favor of the preferentially amplified allele), which in turn, often prevents the heterozygote from being detected by SNP discovery software (Quinlan and Marth 2007). This dilemma often affects so-called "medical re-sequencing" projects in which researchers sequence a large genomic region (e.g., 100 Kb) among many individuals with and without a disease. In these studies, the goal is to uncover all of the polymorphisms that exist among the cohort. Therefore, attempts to avoid existing SNPs during PCR primer design are often in vain because inherently little is known about the extent of variation within the studied region. In Chapter 2, I describe the under-appreciated effects of this problem in medical resequencing projects and propose a rational approach to mitigating the effects of heterozygosity within PCR primers.

Despite this limitation, many polymorphism discovery studies with capillary sequences using these methods have been conducted over the past ten years with relatively high accuracy. Existing polymorphism discovery methods, while not perfect, have become reliable enough to facilitate genome-wide polymorphism discovery in organisms with smaller (<= 15Mb) genomes. The major limitation to such studies have been the high cost and relatively low throughput of the capillary sequencing technologies, relative to the next-generation sequencing technologies that have recently been developed.

Next-generation sequencing technologies

Until three years ago, capillary sequencers such as those made by Applied Biosystems were the only realistic option for large sequencing projects. Since then, the genomics field has seen the development of several novel sequencing technologies that have greatly improved the economy of large-scale sequencing projects. The staggering throughput of these technologies relative to traditional capillary sequencing technologies has enabled rapid, cost effective polymorphism discovery projects spanning the entire genome of several model organisms (Hillier et al. 2008). Similarly, the dramatically increased throughput has led to the use of sequence-based methods in lieu of, or in addition to, chip-based methods for epigenetic, gene expression and structural variation studies. As of 2007, three such next-generation sequencing platforms are available from

454 Life Sciences, Illumina, and Applied Biosystems. Other technologies are expected to become available over the next few years (e.g., Helicos BioSciences, Pacific Biosciences, and Visigen Biosciences among others).

The 454 Life Sciences DNA sequencing machine employs a “sequencing-by-synthesis” chemistry, using a pyrosequencing method which cyclically tests for the incorporation of the four DNA nucleotides (Margulies et al. 2005; Ronaghi et al. 1996). The light emitted by luciferase in the subsequent reaction is recorded and is used to detect nucleotide incorporation. Multiple bases in homopolymer runs are incorporated in a single nucleotide test; therefore the number of incorporated bases must be determined from a single scalar intensity measurement (**Figure 1.6**). This causes the 454 sequencing reads to be characterized by nucleotide over-calls (that is, deciding there were too *many* bases incorporated from the observed signal) and under-calls (that is, deciding there were too *few* bases incorporated from the observed signal). Consequently, insertions and deletions are the dominant error types in 454 reads (Huse et al. 2007; Margulies et al. 2005; Quinlan et al. 2008).

Because of the variable number of nucleotides that are incorporated in each successive test, the resulting sequencing reads have variable lengths. The first sequencer model (known as the GS20) produces, on average, 100-base reads. The more recent FLX model generates on average, 250 bp reads. The total throughput per machine run is ca. 125 Mb. A 500 bp read-length model is awaiting release and paired-end read protocols are now available. Paired-end sequencing refers to the sequencing of short stretches of DNA on either end of a much longer DNA fragment of a known length (e.g. the sequencing of 25 bp on either end of a 1 Kb DNA fragment).

The Illumina 1G short-read sequencer, which is based on Solexa technology, also employs “sequencing-by-synthesis” chemistry, but unlike the 454 technology, it uses a modification of the Sanger dideoxy method (Sanger et al. 1977) to allow for the addition of a single complementary nucleotide analog in each sequencing cycle (**Figure 1.7**). Since the nucleotides are fluorescently labeled with different color dyes, it is possible to determine which of the four nucleotides was incorporated in a given cycle. The Illumina sequencer currently produces approximately 1 Gb per run from 25-50 bp, fixed-length reads. Paired-end read protocols are available for short fragment sizes (up to 2000 bp).

The SOLiD technology from Applied Biosystems employs a ”sequencing-by-ligation” chemistry, which serially tests for the ligation of fluorescently-labeled, di-base encoding, oligonucleotide probes. Consequently, each probe detects dinucleotides, as opposed to individual bases. The machine reads a degenerate “color-space” alphabet that can be translated into actual nucleotide sequence through the first known base in the read. The SOLiD machine currently produces over 4Gb of raw sequence per machine run with 25-70 bp fixed-length reads. Paired-end read protocols are available for the 1-10 kb DNA fragment size range with ~2.5 kb fragment length representing the best compromise between library complexity and fragment length.

In addition, there are several other companies that are in various stages of developing even higher-throughput sequencing technologies. Two of the most promising companies are Helicos Biosciences and Pacific Biosciences. Both companies seek to develop sequencing technologies that are capable of sequencing single molecules of DNA—thus obviating the need to amplify the starting material. As of February 2008, Helicos sold its first instrument whose throughput appears to be roughly 30 Gb per sequencing run. However, the accuracy of the sequencing reads is still unclear, and the reads remain quite short (<= 30 bp on average). Pacific Biosciences presented preliminary sequencing results at the 2008 Advances in Genome Biology and Technology Meeting on Marco Island, Florida. Because of the imaging and chemical approach, this technology seems extremely promising. They anticipate the ability to sequence a human genome in less than an hour for $1000. However, the technology will likely not be available until 2010.

Owing to the sheer volume, shorter read length and different error profiles of these new technologies, traditional sequence analysis methods (e.g., base-calling, sequence mapping and sequence alignment) have proven to be inadequate for large-scale studies employing these new data. In addition, the sequencing technologies themselves continue to evolve as throughput and read lengths increase and library preparation methods improve. Therefore continued development of novel computational and experimental methods for complex genomic studies using these new technologies will be required.

Studying genetic variation with new sequencing technologies

The unprecedented throughput and economy of the new sequencing technologies has allowed several large-scale studies investigating genome-wide variation (Hillier et al. 2008; Korbel et al. 2007; Owen-Hughes and Engeholm 2007; Sebat et al. 2007).

However, because of the novel sequencing chemistries employed by the new technologies, the error profiles of the sequence reads differ substantially from those of traditional capillary sequence reads. As a result, the base quality values assigned by the base calling programs that are used in conjunction with these sequencing machines fail to accurately reflect the true accuracy of the called bases. This poses a problem for polymorphism discovery programs such as *PolyBayes*, because they rely on accurate base quality values to identify true polymorphisms. Unreliable base quality values will lead to high false positive (that is, spurious polymorphism calls) and false negative (that is, true polymorphisms that are missed) rates. Earlier polymorphism discovery projects using the new sequencing technologies have relied on either deep sequence coverage or specialized error estimates to overcome this problem.

Yet single-end sequence protocols from both the Illumina and 454 technologies have recently been shown to be suitable for accurate SNP discovery, even with shallow sequence coverage (Hillier et al. 2008; Quinlan et al. 2008). While the 454 technology faces inherent limitations for small INDEL discovery because of the frequency of nucleotide over- and under-calls, it appears that the Illumina technology enables accurate small INDEL discovery because of its low insertion-deletion error rate.

The major limitation in studying genetic variation with the next-generation technologies is that the relatively short sequence reads they produce are susceptible to improper genomic mapping. However, as mentioned, all of the extant technologies have developed paired-end sequencing technologies that reflect short sequences from the ends of much larger DNA fragments. Paired-end sequences therefore serve as a proxy for much longer DNA sequences that are much less susceptible to improper genome mapping because of sequence paralogy. Moreover, because of the much larger fragment length, paired-end reads are suitable for identifying copy number and structural variations in the genome. Assuming that the standard deviation of the paired-end fragment distribution is relatively small, insertions and deletions that are significantly larger or smaller than the tails of the fragment distribution can be confidently identified in a given genome. Several studies employing this approach are currently underway and seek to describe the landscape of structural and copy number variation in healthy and disease-correlated genomes.

Summary of dissertation

My research in the Marth laboratory has focused on the development of computational and experimental methods to facilitate discovery and characterization of genetic variation. Chapter 2 describes a systematic bias in traditional, PCR-based resequencing studies that hinders the discovery of medically important rare alleles. We illustrate that this bias is caused by cryptic heterozygosity in PCR primer binding sites and describe a rationale and effective resolution to the problem. Chapter 3 describes the novel base calling algorithm, *Pyrobayes*, that I developed in order to assign base quality values that more precisely reflect the accuracy of the called bases in pyrosequences from 454 Life Sciences sequencing machines. We show that the quality values assigned by Pyrobayes are more accurate than those produced by the manufacturer-supplied software. As described in Chapters 3 and 4, the improved base quality values produced by Pyrobayes enable sensitive SNP calling even among single 454 sequence read coverage. Chapters 4, 5 and 6 describe large-scale, whole genome polymorphism discovery projects in *D. melanogaster*, *C. elegans* and *P. stipitis*. These studies required the development of an improved sequence alignment method (*Mosaik*, Michael Stromberg) and an efficient version of the *PolyBayes* polymorphism discovery algorithm (developed by Gabor Marth) that is suited to the vast throughput of the new sequencing technologies. These studies involved each of the three primary next-generation sequencing technologies and illustrated their respective strengths and weaknesses for studying genetic variation. Collectively, these studies have established a framework for whole-genome polymorphism discovery in human and for the rapid mutational profiling of model organisms using next-generation sequencing technologies. Chapter 7 discusses some of the pitfalls of next-generation sequencing technologies, makes predictions about where the genomics field is heading in the next five years, and describes several new lines of research that I would like to undertake that build upon the experience I have gained during my dissertation work.

Chapter 1 Figures

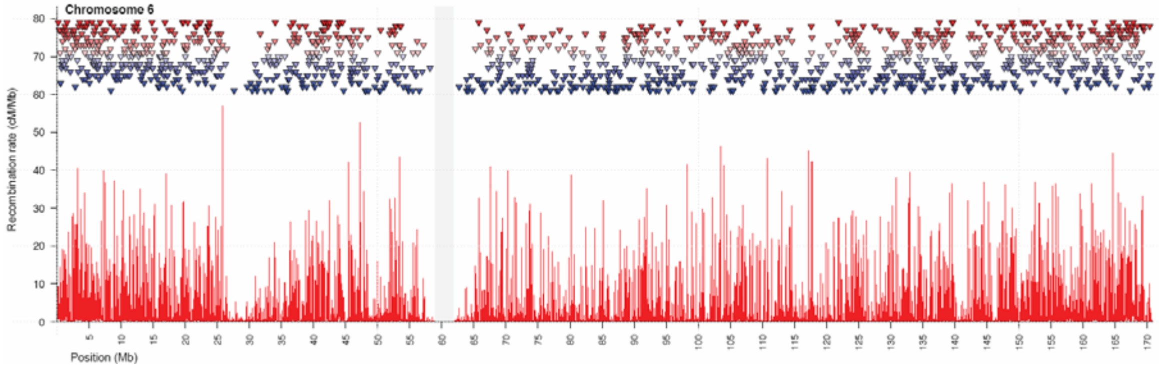


Figure 1.1. Recombination rates and SNP density on human chromosome 6 (modified from The International HapMap Consortium, 2005)**.** The magnitude of recombination rates (red columns) and the density of SNPs (blue triangles) vary across the genome, and their rates are correlated.

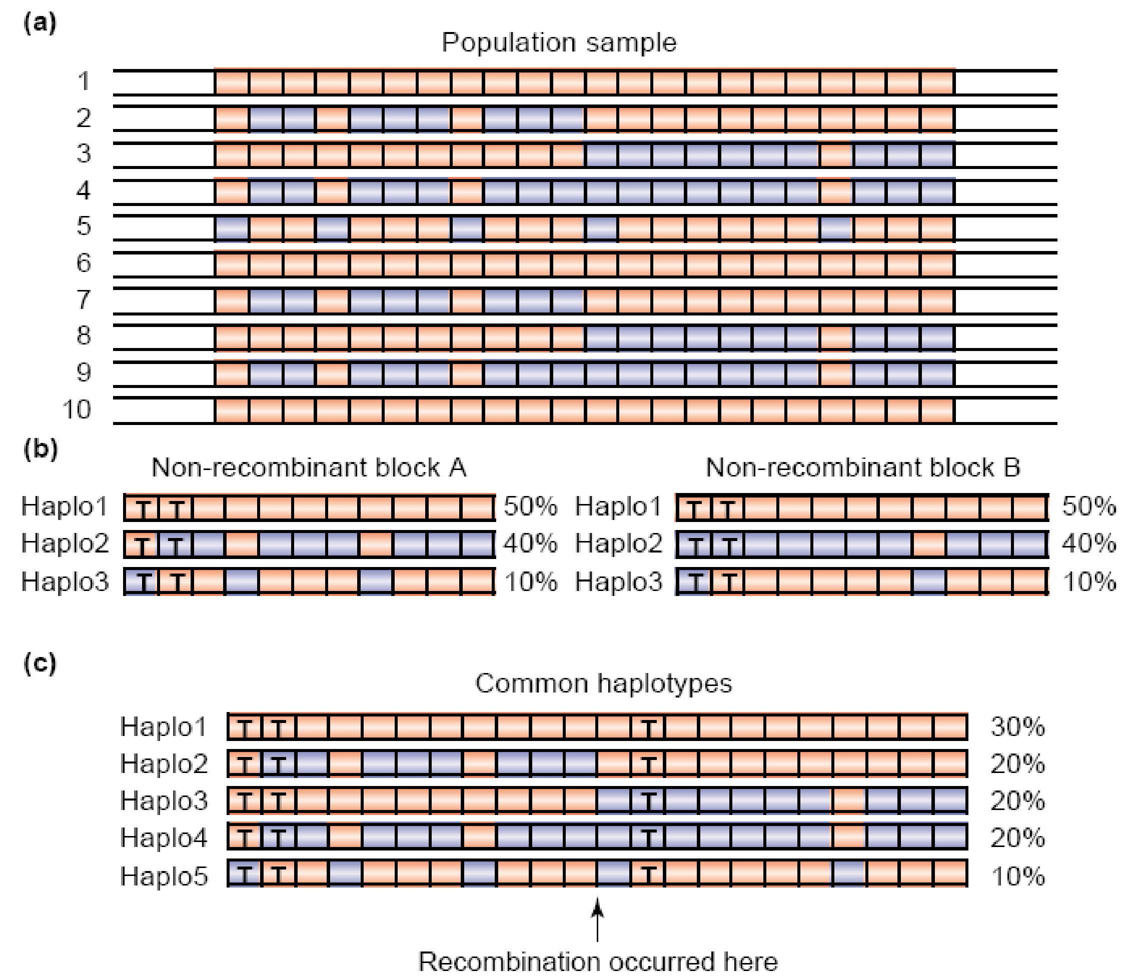


Figure 1.2. The haplotype structure of human chromosomes (modified from Cardon et al, 2002). (a) Hypothetical chromosomes from a population are shown with common alleles in red and alternate alleles in blue. (b) The haplotype variation in (a) can be summarized by two sets of haplotype blocks. Each set has three haplotypes that show no evidence of recombining. Each haplotype block can be inferred by genotyping two SNPs (marked with a "T"). (c) However, when considering the variation observed in each block in the population, only three SNPs are needed to determine the entire haplotype instead of four.

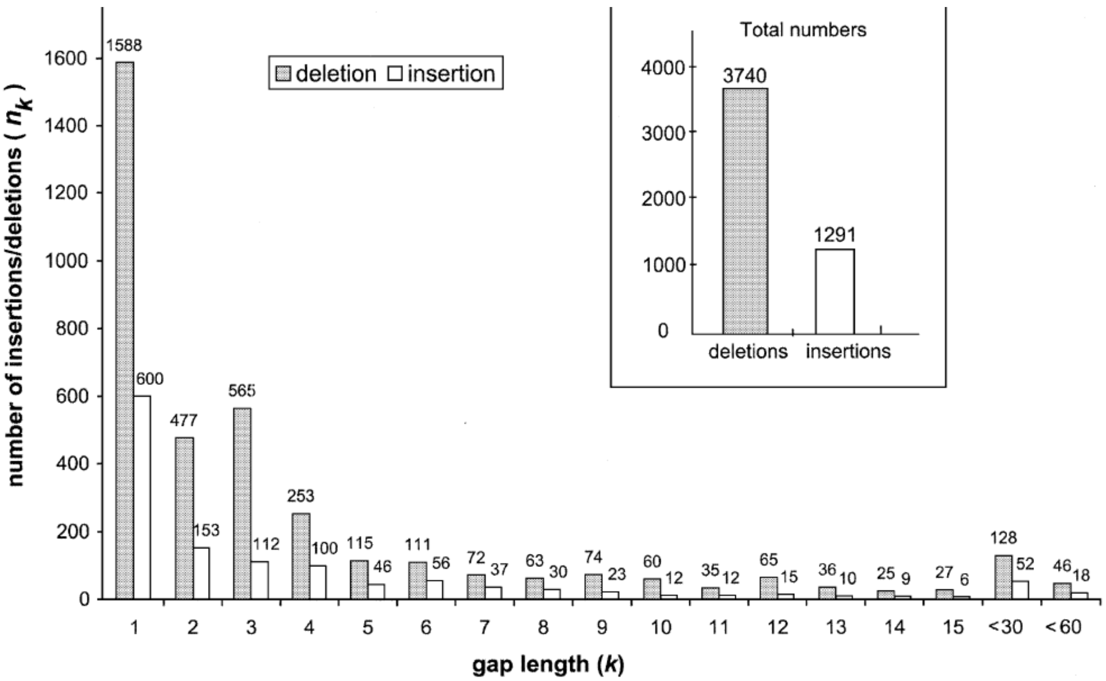


Figure 1.3. Length distribution of INDELs in human pseudogenes. (Modified from Zhang et al, 2003). The frequency of deletions (gray) and insertions (white) are shown as a function of their length in human pseudogenes. The vast majority of INDEL polymorphisms are <= 3 bp in length.

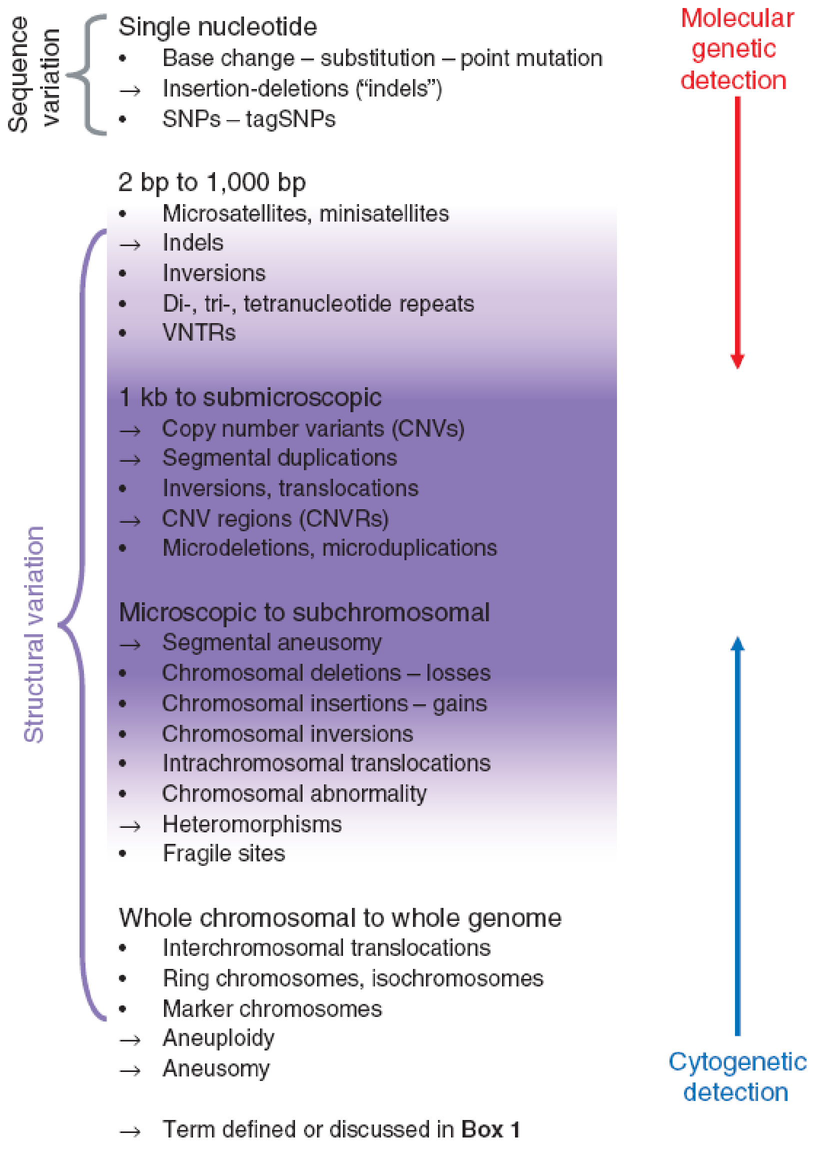

Sequence variation
Single nucleotide
• Base change – substitution – point mutation
→ Insertion-deletions (“indels”)
• SNPs – tagSNPs
2 bp to 1,000 bp
• Microsatellites, minisatellites
→ Indels
• Inversions
• Di-, tri-, tetranucleotide repeats
• VNTRs
1 kb to submicroscopic
→ Copy number variants (CNVs)
→ Segmental duplications
• Inversions, translocations
→ CNV regions (CNVRs)
• Microdeletions, microduplications
Microscopic to subchromosomal
→ Segmental aneusomy
• Chromosomal deletions – losses
• Chromosomal insertions – gains
• Chromosomal inversions
• Intrachromosomal translocations
• Chromosomal abnormality
→ Heteromorphisms
• Fragile sites
Whole chromosomal to whole genome
• Interchromosomal translocations
• Ring chromosomes, isochromosomes
• Marker chromosomes
→ Aneuploidy
→ Aneusomy
Structural variation
Molecular genetic detection
Cytogenetic detection
→ Term defined or discussed in **Box 1**

Figure 1.4. The spectrum of genetic variation (modified from Scherer et al, 2007). The various types of genetic variation are shown from top to bottom in increasing size. The smallest variation types (SNP and INDELs) are the most frequent yet the larger structural variations account for a

more genetic variation in terms of the number of base pairs that vary between any two individuals.

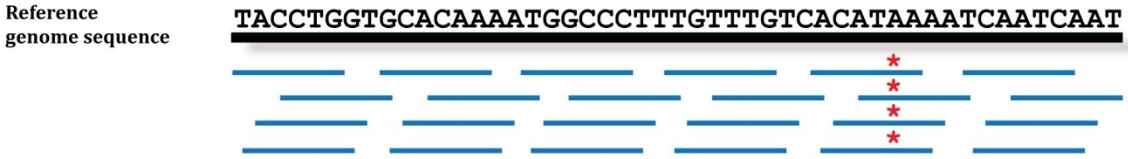

Reference genome sequence
TACCTGGTGCACAAAATGGCCCTTTGTTTGTCACATAAAATCAATCAAT

Figure 1.5. Polymorphism discovery via resequencing strategies. Sequence reads (blue) are aligned to the reference genome (black) of the same species. Polymorphisms are identified by screening for aligned alleles that differ (red asterisks) from the reference sequence.

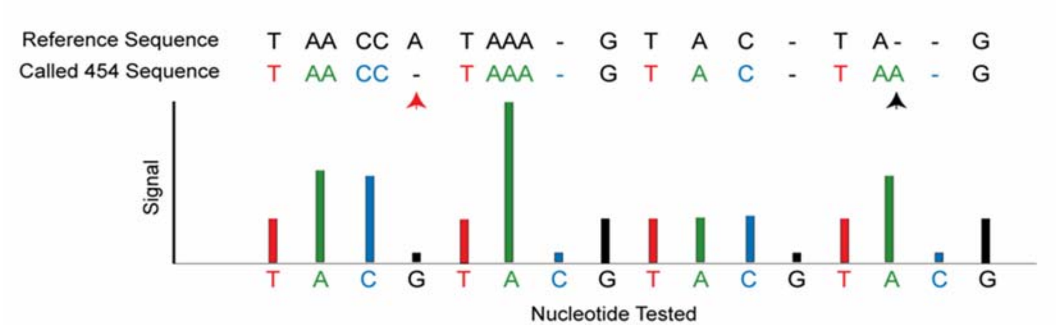


Figure 1.6. Pyrosequencing reads from the 454 Life Sciences technology. The pyrosequencing reads from the 454 Life Sciences technology are the result of cyclical tests (x-axis) for the incorporation of adenine, cytosine, guanine and thymine. The light observed in each test (y-axis) is theoretically proportional to the number of incorporated nucleotides. However, because of ambiguous signals, it is often difficult to determine exactly how many nucleotides were incorporated. Consequently, insertion (black arrow) and deletion (red arrow) errors are common.

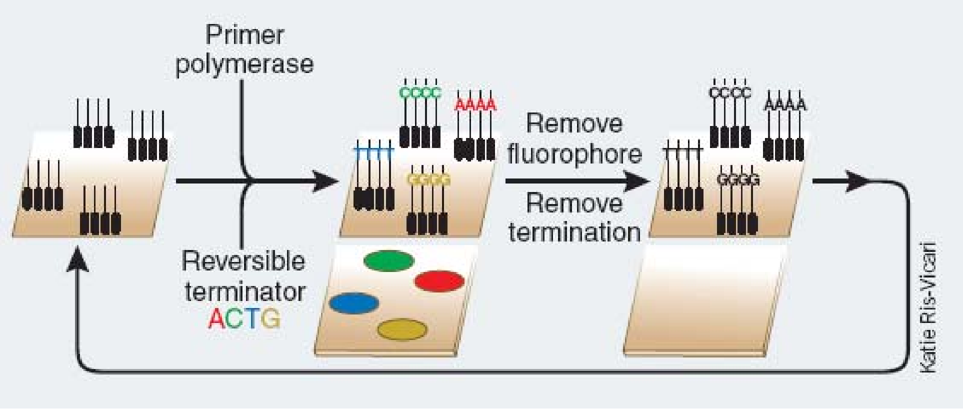


Figure 1.7. The Illumina sequencing-by-synthesis process (adopted from Rusk et al, Nature Methods 2007)**.** Single-stranded DNA fragments are clonally-amplified on a solid chip surface. The colonies are sequenced by the addition of polymerase and fluorescently-labeled nucleotides that, using a reversible terminator, allow exactly one nucleotide to be added to the complementary strand in each sequencing cycle. Fluorescence is detected in each cycle in order to determine which nucleotide was incorporated.

I want to stay as close to the edge as I can without going over. Out on the edge you see all kinds of things you can't see from the center.

-Kurt Vonnegut

2. The negative impact of heterozygosity within PCR primer binding sites on SNP discovery

Abstract

Despite recent advances in sequencing technology, Sanger-principle capillary sequencing of custom PCR amplicons from diploid genomic DNA remains the standard medical resequencing method for individual mutation detection targeting specific genome regions. In this chapter, we describe a systematic error caused by heterozygosity within the PCR primer hybridization sites. Such heterozygosity causes disproportionate amplification between the matched and mismatched chromosomes and leads to missed heterozygotes in the sequence traces. Although this phenomenon has been known for some time, its magnitude has not been estimated or appreciated. This analysis of ten deeply-resequenced ENCODE regions reveals that nearly one in six amplicons contains a SNP in its primers. In such amplicons, one quarter of heterozygotes are miscalled and many existing rare mutations are completely missed. This

phenomenon affects the amplified DNA template directly and therefore neither mate-pair sequencing nor manual trace review reveals it. Moreover, avoiding known SNPs during primer design does not account for novel SNPs in the resequenced individuals. We show that sequencing every nucleotide from more than a single amplicon dramatically reduces the rate of missed heterozygotes and SNPs. We suggest that this strategy should therefore be immediately adopted into PCR-based resequencing protocols.

Introduction

As part of my effort to extend the Marth laboratory's SNP discovery program *PolyBayes* (Marth et al. 1999) for the detection of heterozygotes in capillary sequence traces, we analyzed the deep resequencing data produced by the HapMap project in ten ENCODE (ENCyclopedia Of DNA Elements) regions (Birney et al. 2007; The ENCODE (ENCyclopedia Of DNA Elements) Project Consortium 2004; The ENCODE Project Consortium 2003). Each ENCODE region was sequenced in 48 individuals from roughly a thousand partially overlapping PCR amplicons, resulting in over 761,000 sequence traces (see **Table 2.1** for a detailed description of the data). This extensive dataset is appropriate for method development because it was produced using the same molecular strategy that is the standard for medical resequencing (Mackelprang et al. 2006; Sjoblom et al. 2006). It is ideal for software testing because highly accurate, chip-based genotypes (Hinds et al. 2005; Matsuzaki et al. 2004) are available for 39 of the 48 individuals from the HapMap data. These HapMap genotypes serve as a reference to which we can compare trace-based genotype calls. When making these comparisons, we found SNPs at which many heterozygous individuals in the HapMap appeared as homozygous in the traces from a given amplicon (**Figure 2.1a**). Often, the same SNP was also sequenced from a second overlapping amplicon, where the same individuals appeared as heterozygotes in the traces (**Figure. 2.1b**). The only systematic difference that we could find between such overlapping amplicons was that those in which the HapMap heterozygotes were absent typically contained one or more SNPs in their PCR primers. On the other hand, those amplicons in which the heterozygotes were present rarely had SNPs in their primers.

Results

In order to investigate whether the presence of SNPs in the PCR primer accounts for the homozygote/heterozygote discrepancy, we first identified the amplicons that contained at least one SNP in their primers (we will now refer to this as a primer SNP) (**Methods**). We found that 15.4% of the amplicons (1,440 of 9,347) had a primer SNP. To ensure that missed heterozygotes were not artifacts of our new software we employed a validated heterozygote detection method, *PolyPhred* (Stephens et al. 2006). We evaluated traces on a per-amplicon basis and tabulated candidate SNPs and the genotype call for every individual with sufficient trace quality. We considered each HapMap SNP in every amplicon in which it appeared (that is, possibly multiple times): we tabulated 19,009 such SNP positions at which there were a total of 132,638 heterozygous HapMap genotypes. *PolyPhred* detected 14,604 of the SNPs and made a genotype call for 78,340 of the heterozygotes (**Table 2.2**).

Missed heterozygote rates in amplicons with and without primer SNPs.

PolyPhred missed 7,057 of the 78,340 heterozygotes, which results in a missed heterozygote rate (MHR) of 9.1%. We then recalculated these rates for the amplicons with primer SNP(s) and those without. The MHR was much higher, 22.2% (3,155 of 14,186), in the amplicons with primer SNP(s) and much lower, 6.1% (3902 of 64154), in the amplicons without. This means that 44.6% of all missed heterozygotes occurred in the 15.4% of the amplicons containing primer SNPs, indicating that SNPs inside PCR primers are strongly correlated with missed heterozygotes. We also calculated the rate of miscalled homozygotes in the same two sets of amplicons, and found no significant difference between amplicons with primer SNPs (1.4%) and amplicons without (1.5%). The fact that heterozygotes are missed much more frequently in amplicons where the PCR primer site is polymorphic led us to hypothesize that the higher MHR is driven by unequal PCR amplification in individuals heterozygous at the primer SNP (**Figure 2.2**). We found that the MHR was 58.4% (1,830 of 3,132) in individuals with a heterozygous HapMap genotype at the primer SNP. In contrast, the rate was 6.1% (516 of 8,499) in individuals with a homozygous HapMap genotype at the primer SNP. This means that in amplicons with a primer SNP as much as 78.0% of missed heterozygotes were found in the 26.9% of individuals who were heterozygous at that primer SNP.

Haplotype prediction experiments.

When the SNP detection algorithm miscalls a heterozygote at the SNP inside the amplicon (the amplicon SNP), it makes a homozygous call for one or the other allele. If these errors are random with respect to the primer SNP genotype, one allele is just as likely to be called as the other (the null model). If, however, the PCR primers preferentially amplify the chromosome that has the matching primer sequence, it should (incorrectly) call the allele on that same chromosome. Therefore if we have the phased haplotypes of individuals heterozygous both at the primer SNP and at the amplicon SNP and we know which allele matches the primer sequence, we should be able to predict the identity of the erroneous homozygous call at the amplicon SNP. This procedure made the correct prediction 93.1% of the time, which is statistically significant when compared to the null model (p-value 2.3E-220 based on the Pearson's chi-squared test, n=1,354).

The effect on rare genetic variants.

Rare genetic variants are typically discovered as a very small number of heterozygous individuals among a larger cohort (e.g. one or two heterozygotes among 100 individuals). We find that such SNPs are missed at a high overall rate: 14.5% (463 of 3,193) of single-heterozygote SNPs, and 8.7% (117 of 1,340) of double-heterozygote SNPs were missed entirely in one amplicon. Unequal amplification caused by heterozygosity in the PCR primer should exacerbate these rates. Indeed, single-heterozygote HapMap SNPs were missed 24.5% of the time (109 of 444) within such amplicons, as compared to 12.9% of the time (354 of 2,749) in amplicons without primer SNPs. Similarly, SNPs with two heterozygotes were missed 15.0% of the time (27 of 180) in amplicons with primer SNPs, as compared to only 7.8% of the time (90 of 1,160) in amplicons without.

The impact of primer SNP location and heterozygote frequency.

Resequencing protocols with provisions to avoid SNPs during primer design focus on the 3′ end of the primer sequence (Ikegawa et al. 2002), presumably because this is where imperfect annealing due to allelic mismatch would be most likely to inhibit polymerase binding. We find that biased amplification occurs without regard to the position of the primer SNP relative to the 3′ end (**Figure 2.3**). SNPs in the primer can cause a departure from Hardy-Weinberg Equilibrium (HWE) by reducing the fraction of heterozygotes from what is expected based on the allele frequencies (Balding 2006; Ikegawa et al. 2002). We have quantified this phenomenon by measuring the reduction of heterozygote frequency at amplicon SNPs as a function of heterozygote frequency at the primer SNP (**Figure 2.4**). This reduction is most pronounced at higher primer

SNP heterozygote frequencies, where the reduction is nearly 50%. At a minimum, this skews the allele frequency estimate. At worst, departure from Hardy-Weinberg equilibrium forces quality control procedures to discard the SNP as the potential result of paralogous amplification.

Sequencing a region from an additional amplicon reduces the MHR dramatically.

The MHR for SNPs that were sequenced in a single amplicon with primer SNP(s) is 20.9%. This rate drops five-fold, to 3.7%, for SNPs sequenced both in an amplicon with primer SNP(s) and an additional amplicon without (**Figure 2.5**). Double amplicon coverage also reduces the overall MHR by four-fold (from 8.9% to 2.2%), and triple coverage reduces it by nearly twenty-fold (to below 0.5%). The genotyped HapMap SNPs do not represent a full catalog of all single-nucleotide variation in the ENCODE regions, because of incomplete ascertainment caused by failed or low quality traces, SNPs that were discovered but for which no genotyping assay could be designed, and genotyping failures (The International HapMap Consortium 2005). Therefore one cannot accurately calculate the fraction of missed SNPs in the traces. However, one can estimate the fraction of SNPs that were discovered in two overlapping amplicons but would have been missed in a single amplicon (**Methods**). This rate is 28.7% in amplicons with a primer SNP (264 of 921) and 13.1% (471 of 3,608) in amplicons with no primer SNP(s).

Discussion

We have demonstrated that heterozygosity within the primer hybridization site is an important systematic cause of missed heterozygotes. Disproportionate amplification due to primer SNP(s) increases the overall MHR by as much as 50%. In affected amplicons, one misses 22% of heterozygotes and over 20% of rare SNPs. Allele frequency estimates for SNPs in these amplicons will be miscalculated and can appear to deviate from HWE. Missed heterozygotes and rare SNPs cluster in amplicons with primer SNPs, potentially in exonic DNA. For example, in the ten ENCODE regions we found 20 exons (a total of 3,847 bp) in 15 distinct genes that were only sequenced from amplicons with primer SNPs (**Methods**). Specifically, HOXA1, HOXA2 and HOXA6 each had such an exon. An additional 36 exons (15,115 bp) in 18 unique genes were sequenced mainly (but not exclusively) from amplicons with primer SNPs. The impact of potential missed SNPs will be even greater on medical resequencing projects that exclusively target coding and regulatory regions of important candidate genes.

Our results indicate that multi-amplicon coverage dramatically reduces both the rate of missed heterozygotes and missed SNPs, and suggests that it not only mitigates the systematic bias due to unequal PCR amplification but also remedies missed heterozygotes from software errors (Stephens et al. 2006; The International HapMap Consortium 2005; Zhang et al. 2005) and other sequencing artifacts not addressed herein. Therefore we recommend that resequencing projects adopt an amplicon design strategy that, in addition to avoiding primers that overlap known SNPs, requires that each nucleotide position within the region is sequenced from at least two amplicons. This strategy necessitates the comparison of genotype calls from overlapping amplicons and the resolution of genotype discrepancies. This is best accomplished by informatics techniques that align the traces to the reference genome sequence. Although additional coverage increases project costs, it ensures much more accurate genotypes and near-complete discovery of rare mutations. At the cost of increasing amplicon coverage to three-fold, the accuracy of sequence-based genotyping approaches the accuracy of chip-based platforms, offering an economical alternative for genotype confirmation.

Methods

Data acquisition.

We downloaded all resequencing traces and the associated trace information (e.g. PCR primer pair and HapMap individual identifier) from the Trace Archive at the NCBI. We downloaded the HapMap genotypes for each ENCODE region, and for the 39 individuals shared between the genotyping and the resequencing project from the International HapMap Project website (http://hapmap.org/genotypes/2005-10/non-redundant/, October 2005, Phase II release, non-redundant set). We also downloaded the phased haplotypes for the 39 individuals (http://hapmap.org/downloads/phasing/2005-03_phaseI/ENCODE/).

Trace alignment and SNP discovery.

We mapped traces and primers to their appropriate amplicon and aligned to the corresponding genome reference sequence (build 34) from the NCBI (ftp://ftp.ncbi.nlm.nih.gov) using the anchored multiple alignment algorithm implemented in the *PolyBayes* SNP discovery program, one amplicon at a time.

We ran the heterozygote detection software *PolyPhred* (version 5) on each assembled amplicon. To reduce the number of falsely discovered SNPs and genotypes, we only accepted *PolyPhred* SNP candidates with a score >= 70. (–score 70). We used the option that instructs *PolyPhred* to include the reference sequence in the SNP detection as a separate allele (–refcomp). We also provided *PolyPhred* with information to integrate multiple traces from the same individual (–source option), and used the resultant individual genotype calls.

Comparing HapMap genotypes and trace-based genotype calls.

The alignment of the traces to the reference genome sequence placed discovered SNPs in reference chromosome coordinates. This allowed us to compare *PolyPhred* calls directly to the HapMap genotypes. We could only compare such genotypes where, on one hand, traces were aligned and analyzable (as determined by *PolyPhred*), and, on the other hand, conclusive HapMap genotypes were available. To avoid SNP calls in amplicons that may have amplified multiple paralogous regions, we excluded amplicons where *PolyPhred* called 20 or more SNPs. We calculated missed heterozygote rate (MHR) as the fraction of HapMap heterozygotes that *PolyPhred* miscalled as a homozygote.

Haplotype comparisons.

The HapMap provides phased haplotypes (Stephens and Donnelly 2003; Stephens et al. 2001) for each ENCODE individual at each SNP in the ENCODE regions. For individuals that were heterozygous at a primer SNP, we compared both haplotypes at the primer SNP to the allele present in the primer sequence. We used the haplotype that exactly matched the primer sequence allele to predict the identity of the allele that *PolyPhred* erroneously called as a homozygote.

Integrating gene annotations.

Gene annotations consisting of gene and exon coordinates were downloaded from the NCBI ftp site:
(ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.34.3/mapview/seq_gene.md.gz). We determined regions of overlap between gene features and amplicons with custom scripts.

Chapter 2 Tables

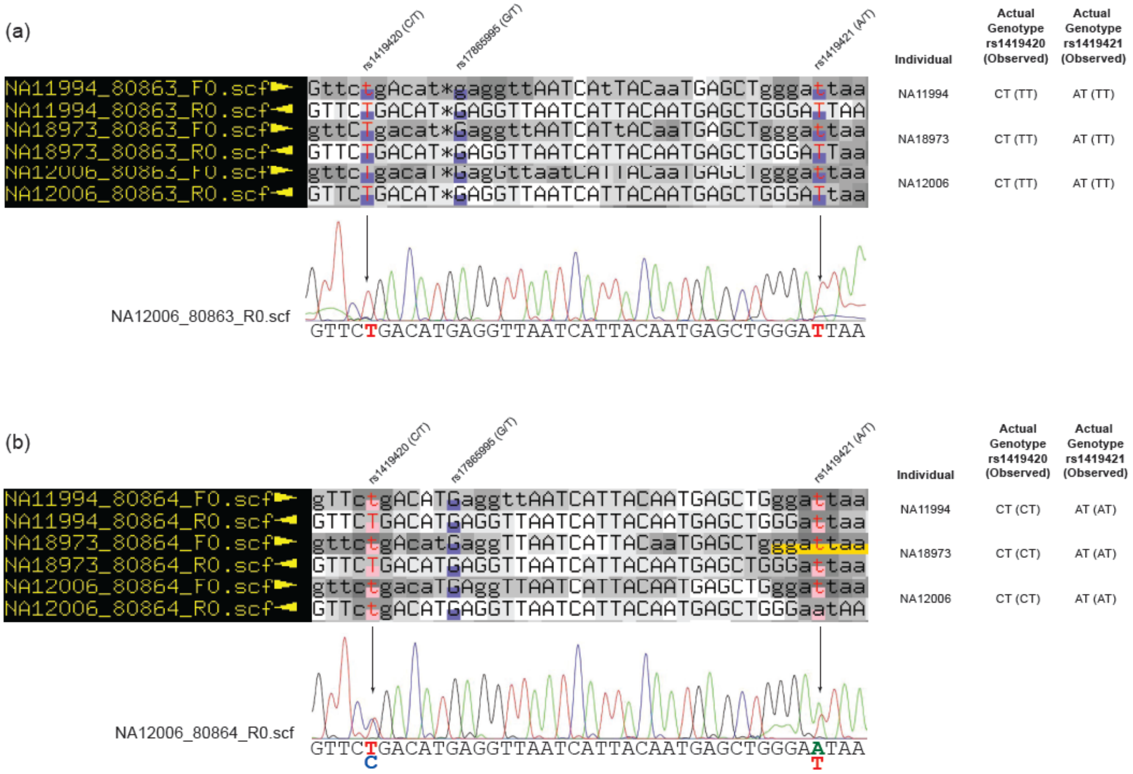
ENCODE Region	Number of Amplicons Analyzed	Average Number of Reads Per Amplicon	Average Amplicon Length	Fraction of Nucleotides Sequenced From 1 Amplicon	Fraction of Nucleotides Sequenced From >=2 Amplicons	Fraction of Amplicons With >=1 HapMap Primer SNP	Number of Polymorphic HapMap SNPs	Number of HapMap Heterozygotes Across Amplicons	Number of POLYPHRED Genotype Calls For HapMap Heterozygotes	Fraction of Missed Heterozygotes In Amplicons With a Primer SNP
ENr113	993	98.0	857.6	0.505	0.495	0.224	1753	17743	12463	0.491
ENm014	1083	94.9	749.9	0.455	0.545	0.172	2321	15946	11938	0.407
ENr112	1079	95.5	750.0	0.460	0.540	0.139	1316	16603	11913	0.529
ENr131	989	85.3	747.5	0.492	0.508	0.133	1251	14931	9200	0.379
ENm013	1040	96.8	748.4	0.478	0.522	0.180	2194	13817	10025	0.379
ENm010	829	69.7	932.0	0.638	0.362	0.115	958	8615	3927	0.423
ENr123	818	51.1	752.8	0.530	0.470	0.145	1465	15418	4183	0.395
ENr213	864	81.6	781.2	0.710	0.290	0.142	1186	10128	5860	0.442
ENr232	820	53.0	746.6	0.826	0.174	0.115	999	8209	3661	0.341
ENr321	832	71.8	767.7	0.787	0.213	0.159	1271	11228	5170	0.584
Average	***934.7***	***79.8***	***785.4***	***0.588***	***0.412***	***0.154***	***1471***	***13264***	***7834***	***0.446***

Table 2.1. Summary of data used from each ENCODE region for genotype comparisons.

Region	Number of Polymorphic HapMap Genotyped SNPs	Number of Polymorphic HM SNPs Counted Multiply Across Amplicons	Number of Polymorphic HapMap SNPs Discovered by POLYPHRED Across Amplicons	Number of HapMap Heterozygotes Counted Multiply Across Amplicons	Number of POLYPHRED Genotype Calls for the HapMap Heterozygotes Across Amplicons
ENr113	1753	2639	2181	17743	12463
ENm014	2321	2493	2048	15946	11938
ENr112	1316	2088	1748	16603	11913
ENr131	1251	1862	1535	14931	9200
ENm013	2194	2200	1805	13817	10025
ENm010	958	1323	929	8615	3927
ENr123	1465	2151	1175	15418	4183
ENr213	1186	1518	1297	10128	5860
ENr232	999	1182	757	8209	3661
ENr321	1271	1553	1129	11228	5170
Total	***14714***	***19009***	***14604***	***132638***	***78340***

Table 2.2. Summary of HapMap SNPs and heterozygotes compared to SNPs discovered and genotyped by *PolyPhred*.

Chapter 2 Figures



```
NA11994_80863_F0.scf  Gttctg Acat*gaggttAATCAtTACaaTGAGCTgggattaa
NA11994_80863_R0.scf  GTTCTGACAT*GAGGTTAATCATTACAATGAGCTGGGATTAA
NA18973_80863_F0.scf  gttCTgacat*GaggttAATCATtACaaTGAGCTgggattaa
NA18973_80863_R0.scf  GTTCTGACAT*GAGGTTAATCATTACAATGAGCTGGGATTaa
NA12006_80863_F0.scf  gttctgacaT*GagGttaatCATTACaaTGAGCTgggattaa
NA12006_80863_R0.scf  GTTCTGACAT*GAGGTTAATCATTACAATGAGCTGGGATtaa
```

Individual	Actual Genotype rs1419420 (Observed)	Actual Genotype rs1419421 (Observed)
NA11994	CT (TT)	AT (TT)
NA18973	CT (TT)	AT (TT)
NA12006	CT (TT)	AT (TT)

```
NA11994_80864_F0.scf  gTTctgACATGaggttAATCATTACAATGAGCTGggattaa
NA11994_80864_R0.scf  GTTCTGACATGAGGTTAATCATTACAATGAGCTGGGattaa
NA18973_80864_F0.scf  gttctgAcatGaggTTAATCATTACaaTGAGCTgggattaa
NA18973_80864_R0.scf  GTTCTGACATGAGGTTAATCATTACAATGAGCTGGGattaa
NA12006_80864_F0.scf  gttctgacaTGAggTTAATCATTACAATGAGCTGggattaa
NA12006_80864_R0.scf  GTTctgACATGAGGTTAATCATTACAATGAGCTGGGaatAA
```

Individual	Actual Genotype rs1419420 (Observed)	Actual Genotype rs1419421 (Observed)
NA11994	CT (CT)	AT (AT)
NA18973	CT (CT)	AT (AT)
NA12006	CT (CT)	AT (AT)

Figure 2.1. Sequence trace-based individual genotypes that disagree between amplicons. Sequence traces from individual NA12006 are displayed from two overlapping amplicons. (a) The homozygous *PolyPhred* genotype calls (blue tags) from amplicon 80863 disagree with the heterozygous HapMap genotypes. (b) The heterozygous *PolyPhred* genotype calls (pink tags) from amplicon 80864 agree with the heterozygous HapMap genotypes. The relative color intensities in each traces support the respective *PolyPhred* genotype calls. The middle SNP location (rs17865995) was not genotyped by the HapMap and therefore was not addressed here. (Yellow tag: insufficient trace quality. Lower-case alleles: low quality sequence. Upper-case alleles: high quality sequence.)

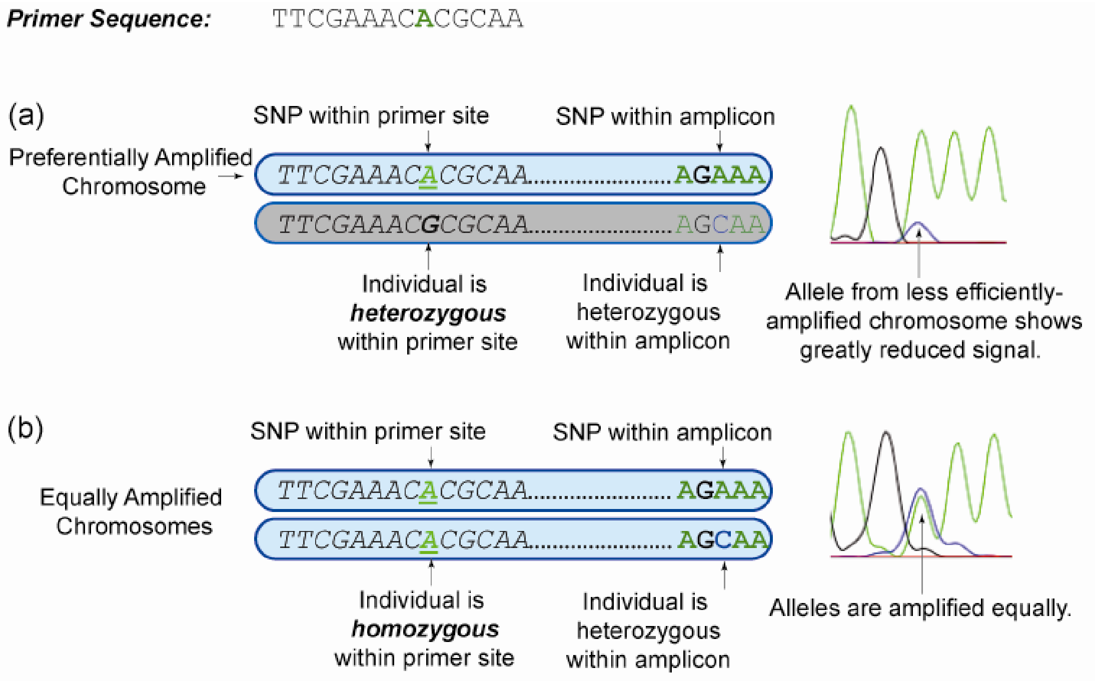

Primer Sequence:
TTCGAAACACGCAA
(a)
SNP within primer site
SNP within amplicon
Preferentially Amplified Chromosome
TTCGAAACACGCAA.......................AGAAA
TTCGAAACGCGCAA.......................AGCAA
Individual is heterozygous within primer site
Individual is heterozygous within amplicon
Allele from less efficiently-amplified chromosome shows greatly reduced signal.
(b)
SNP within primer site
SNP within amplicon
Equally Amplified Chromosomes
TTCGAAACACGCAA.......................AGAAA
TTCGAAACACGCAA.......................AGCAA
Individual is homozygous within primer site
Individual is heterozygous within amplicon
Alleles are amplified equally.

Figure 2.2. Disproportionate chromosomal amplification due to heterozygosity at a SNP in the PCR primer. (a) For an individual heterozygous at a primer SNP the chromosome matching the primer sequence is amplified more efficiently than the chromosome containing the mismatched allele. Sequencing results in reduced color intensity for the allele on the mismatched chromosome. (b) For an individual homozygous at the primer SNP amplification and allele-specific color intensities are balanced.

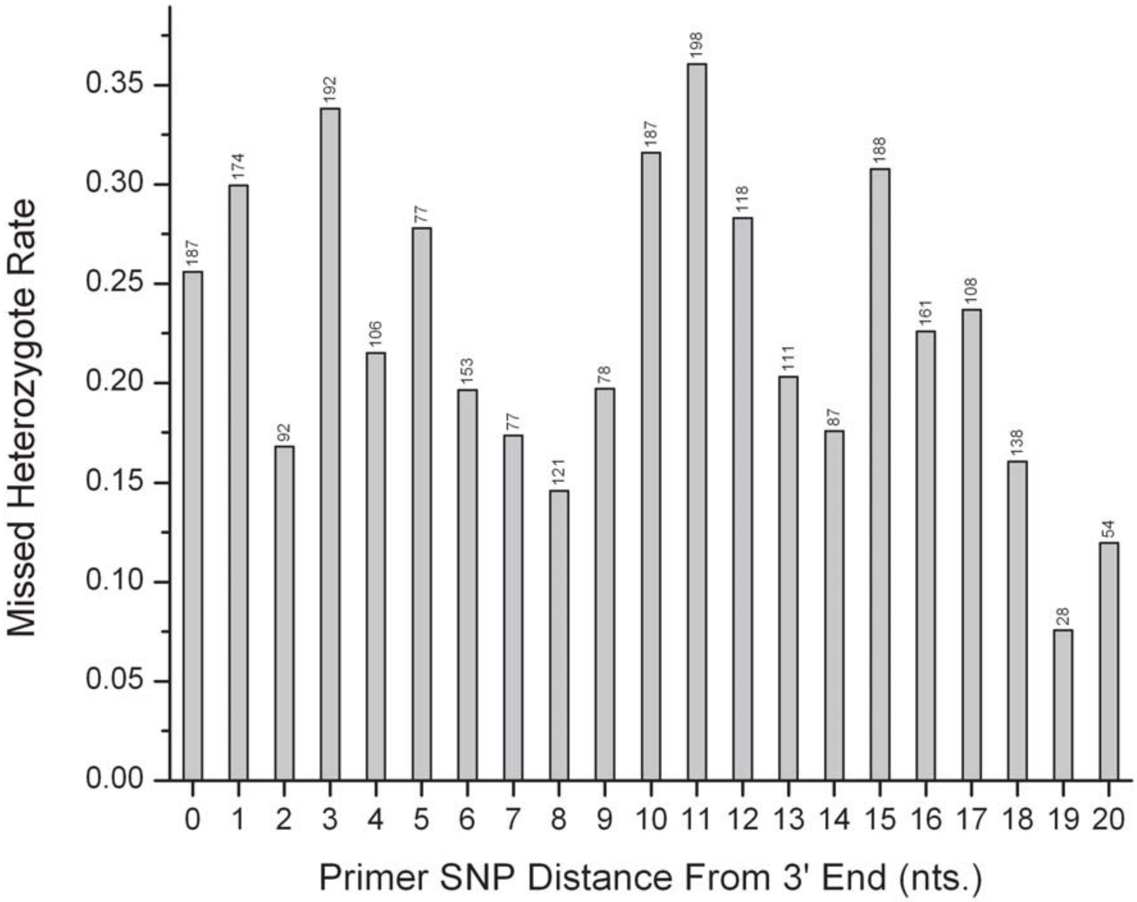


Figure 2.3. SNP location within the primer does not significantly affect MHR. MHR at amplicon SNPs is shown as a function of primer SNP distance from the 3′ end of the PCR binding site. The number of missed heterozygotes is shown above each column.

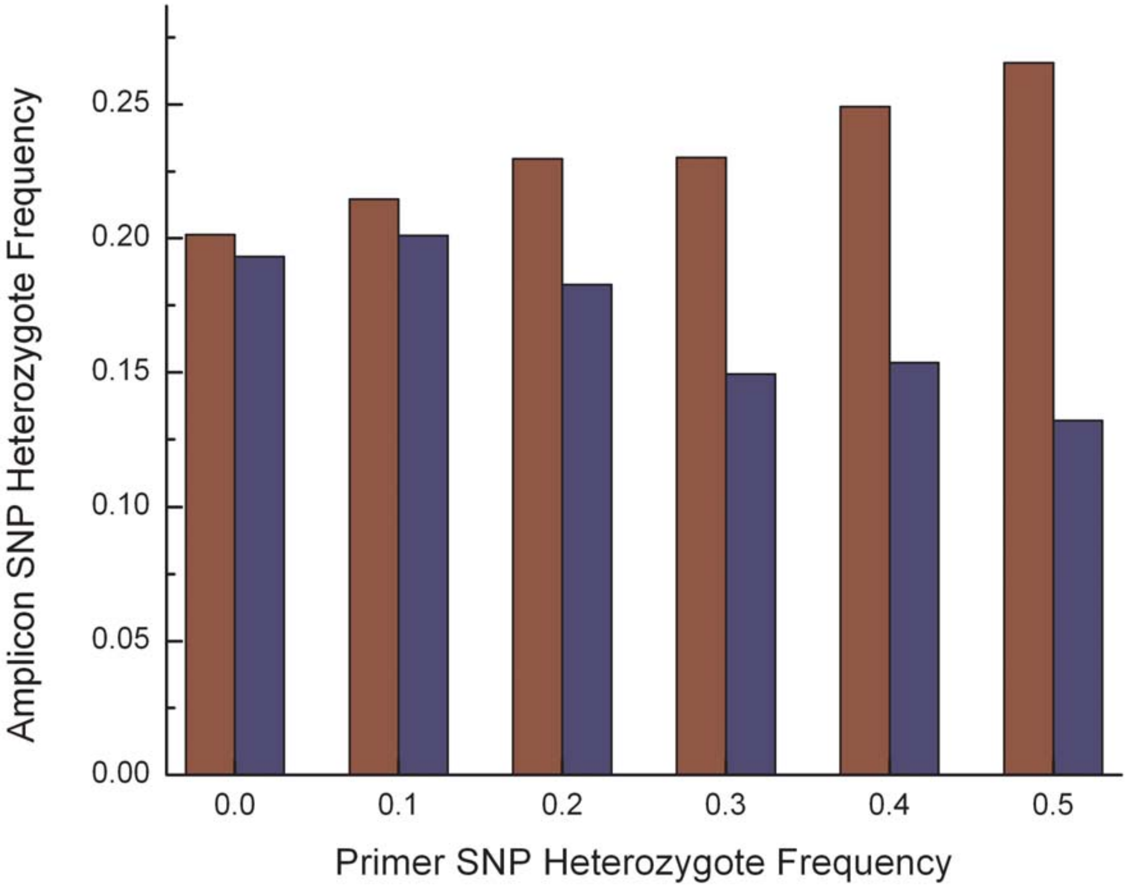


Figure 2.4. As the heterozygote frequency at the primer SNP increases, so does the fraction of missed heterozygotes at the amplicon SNP. Amplicon SNP heterozygote frequency as a function of primer SNP heterozygote frequency. Blue: *PolyPhred* calls. Red: HapMap genotypes.

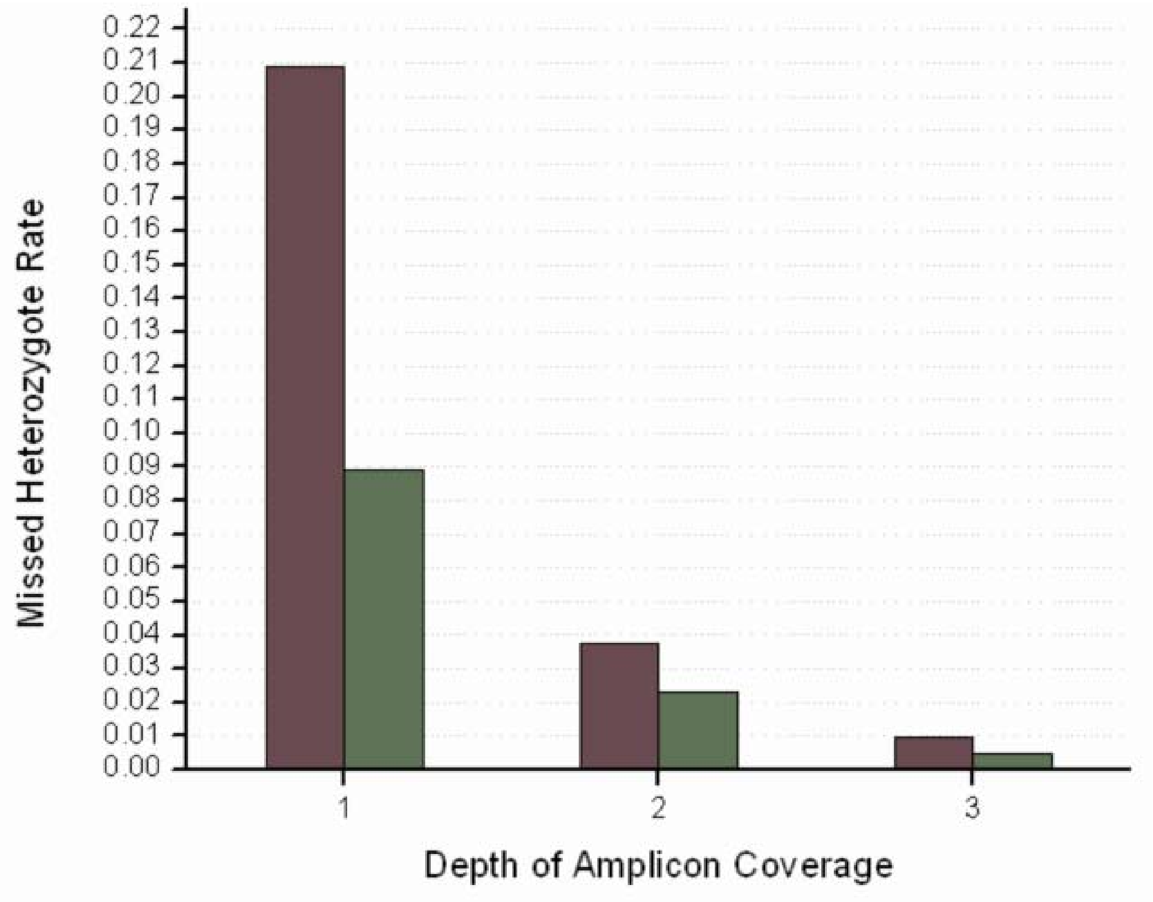


Figure 2.5. Increasing amplicon coverage decreases missed heterozygote rate. Missed heterozygote rates as a function of amplicon coverage are shown. Red: amplicons with primer SNP(s); green: all amplicons.

You can't have a light without a dark to stick it in.

-Arlo Guthrie

I do not mind lying, but I hate inaccuracy.

-Samuel Butler

3. Pyrobayes: An improved base caller for SNP discovery in pyrosequences

Abstract

There are now several next-generation sequencing technologies with greatly improved economy and sequence throughput relative to traditional Sanger-based capillary methods (Bentley 2006; Margulies et al. 2005; Shendure et al. 2004). These machines will likely replace capillary sequencers for resequencing and de novo sequencing, and will facilitate other high-throughput biology applications (Barski et al. 2007; Mikkelsen et al. 2007; Ng et al. 2006). However, utilizing sequencing reads from the new technologies is not trivial because they are generally shorter and their sequencing error profiles are different from that of traditional capillary reads. Pyrosequences from 454 Life Sciences are prone to errors caused by base number over- and

under-calls, resulting in apparent insertion and deletion errors in the reads. To date, reported applications of this technology have relied upon deep sequence coverage to resolve these errors. However, such over-sampling comes at a cost, and it is clearly advantageous to use the available sequence data economically. Here we report the development of *Pyrobayes*, a novel base-caller for pyrosequences from the 454 Life Sciences machines. *Pyrobayes* generates base quality values that correspond to the actual base accuracy significantly better than those produced by the native base caller. As a result, a larger fraction of the bases are assigned high base quality values. We illustrate that *Pyrobayes* permits accurate SNP calling in resequencing applications, even in shallow, single-read coverage.

Introduction

Despite its lower throughput relative to other new sequencing technologies, the 454 Life Sciences pyrosequencer is now in wide use because of its comparatively longer read length (producing 100-250 bp medium-length reads as opposed to 25-50 bp short reads). The reads produced by these machines are the result of cyclical nucleotide incorporation tests. If the tested nucleotide is incorporated, the intensity of the light emitted from subsequent chemical reactions is recorded. This technology is unique in that all nucleotides within a homopolymer run (e.g. AAA) are incorporated in a single test. Thus, the intensity signal produced is proportional to the number of incorporated bases. Moreover, a nonzero signal is typically observed even when no base is incorporated (i.e. “noise”). For these reasons, the biggest challenges for 454 base-callers are determining the correct number of bases, and deciding whether a base needs to be called at all in a given nucleotide test. As a result, 454 reads are characterized by nucleotide over-calls and under-calls that appear as insertion or deletion type errors (Girard et al. 2006; Thomas et al. 2006; Velicer et al. 2006) when compared to a reference sequence of the organism. True substitution errors in the sequence reads are rare because they must result from failing to call the correct base in one nucleotide test and then calling an extra base in a separate nucleotide test. Given the preponderance of insertion/deletion type sequencing errors, it is challenging to describe 454 base accuracy with *Phred* base quality values (Ewing and Green 1998; Ewing et al. 1998), which were developed for slab gel and capillary reads where the majority of errors stem from calling the wrong nucleotide. In contrast, for 454 reads, the base quality value represents the likelihood that the called base is, in fact, part of the sequenced DNA template, as opposed to a base overcall.

Yet the use of accurate *Phred*-like quality estimates is important as most existing sequence analysis software was designed to work with such quality values (Marth et al. 1999; Stephens et al. 2006; Zhang et al. 2005).

Results

The necessity of accurate base quality estimates.

Accurate base quality values are especially important for re-sequencing applications where we must decide whether apparent sequence differences between the resequenced DNA and the reference genome sequence is true allelic variation or sequencing error. To achieve a low false positive SNP calling rate, the sequencing error rate must be substantially lower than the expected polymorphism rate. This means that e.g. for human polymorphism detection no SNP calls should be made from bases with a base quality value lower than 30 (1 in 1,000 bp error rate). If, on the other hand, the majority of bases in resequencing reads are of sufficiently high quality for SNP detection, it is possible to conduct economical genome surveys at low sequence coverage, because SNPs can then be detected with high confidence between singly-aligned reads and the reference genome. Unfortunately, the majority of the native 454 base quality values are not sufficiently high for SNP calling: we find that only 24% of the native 454 base calls are above 30, and no base calls are assigned base quality values above 40 (**Figure 3.1a**). Our assessment (**Methods**) shows, however, that the reason for this is not that 454 reads cannot be called accurately but that the native base quality values grossly underestimate the actual base accuracy (**Figure 3.1b**). We developed a new base-calling program for 454 pyrosequences, *Pyrobayes*, which produces more accurate base quality values and as a result, calls more high-quality bases. We demonstrate that the higher-quality base-calls produced by our new software make SNP calling in low read coverage possible, thereby enhancing the utility of 454 reads for resequencing, the main application area for next-generation DNA sequencers.

The Bayesian base-calling strategy.

Since the main source of sequencing error in 454 sequences is nucleotide over- and under-calls, our base-calling strategy is to estimate the correct number of bases incorporated in each nucleotide test. Our Bayesian strategy (**Methods**) requires data likelihoods i.e. the distribution of the incorporation signal for each known homo-polymer length (**Figure 3.2a, Figure 3.3**). We estimated these distributions by collecting shotgun resequencing data with the 454 Life Sciences

GS20 instrument from a mouse BAC (bacterial artificial chromosome) clone, and aligning the resulting 661,481 reads (a total of 65,875,710 bases) to the finished reference sequence of the same BAC clone (**Methods**). It also requires the prior expectations for how often we see homo-polymeric runs of different lengths (**Figure 3.2b**). We determined these prior expectations for several genomes and found that the relative frequencies of nucleotide runs of length *n* are substantially different from the theoretical expectation that they are proportional to $1/4^n$. However, the actual frequency distributions were similar enough across the genomes we analyzed not to warrant organism-specific priors for our software. Using the data likelihoods and the prior distributions we calculated the Bayesian posterior probability for each possible homo-polymer length as a function of the observed nucleotide incorporation signal (**Figure 3.2c**). The most likely number of bases is the homo-polymer length with the highest posterior probability.

Pyrobayes produces the called base sequence by concatenating the most likely number of bases for every consecutive incorporation test (adding no base if the most likely base number is zero). The base quality value assigned to each base is the probability that the base in question is not an over-call relative to the true

DNA sequence (see **Methods**, **Figure 3.2d** for details). We found it also useful to call one extra base, in addition to the most likely number of bases, as long as the presence of that base is above a minimum probability (see **Discussion** for the merits of this approach).

Utility for SNP calling in low sequence coverage.

To evaluate base-calling accuracy with our method we collected 299,654 454 reads (GS20 model) from the inbred reference (*iso-1*) strain of *Drosophila melanogaster*. We re-called these reads with *Pyrobayes*. We aligned both the original and the re-called sequences to the reference genome sequence (**Methods**). Only counting reads that could be uniquely aligned with our stringent alignment criteria, over 20Mb of sequence was aligned for each method. The overall base accuracy (**Methods**) is quite high both for *Pyrobayes* and the native 454 base caller (99.60% vs. 99.61% base calling accuracy, respectively). As **Figure 3.4a** shows, the *Pyrobayes* insertion error rate is higher (0.29%) than that of the native base caller (0.24%), but the *Pyrobayes* deletion rate is lower (0.09% vs. 0.10%). Most importantly for SNP discovery, the *Pyrobayes* substitution error rate is 60% lower (0.017% vs. 0.042%) than that of the native base caller.

Comparison to the native base caller.

Moreover, the base quality values assigned by *Pyrobayes* represent the true accuracy of the 454 reads more closely than the native base quality values (**Figure 3.1b**). As a result, *Pyrobayes* base quality values are, in general, higher (**Figure 3.1c**). For example, 56% of the *Pyrobayes* bases are assigned base quality values of 30 or higher, as compared to 24% of the native calls (**Figure 3.1a**). Additionally, *Pyrobayes* produces quality values up to 50 (1 in 100,000 bp error rate), whereas the native base caller does not produce quality values above 40 (1 in 10,000 bp error rate).

We investigated the effect of *Pyrobayes*'s reduced substitution error rate and higher overall base quality values on SNP (single-nucleotide polymorphism) detection. We searched for single base pair differences between the 454-sequenced *iso-1* reads and the *iso-1 Drosophila* reference sequence (**Methods**). Since these sequences are from the same inbred melanogaster strain, and the overall accuracy of the Drosophila genome sequence is high, we expect few, if any, true polymorphisms. SNPs discovered in this comparison therefore estimate the false SNP discovery rate. This rate was 1.22 per 10,000 base pairs using the native base calls, but only 0.97 per 10,000 base pairs using the *Pyrobayes* base calls. It is important to consider that the actual false positive SNP rate depends on the polymorphism rate in the resequenced organism. For example in *Drosophila*, where the pair-wise polymorphism rate has been observed to be as high as ~ 1/200 base pairs (Hoskins et al. 2001), our estimated false SNP discovery rate would correspond to a false positive SNP rate of roughly 1.9%.

To compare the missed SNP rate with the respective base callers and estimate the false positive SNP rate directly, we analyzed a 454 dataset from a genome survey of inbred geographic isolates of *Drosophila melanogaster*. For our tests we used a single 454 run from an isolate from Malawi for which experimental SNP validation data was available. Using the *Pyrobayes* base calls we found 1,118 SNP candidates with a *PolyBayes* SNP probability (Marth et al. 1999) cutoff value of 0.7 or higher (**Methods**). The validation rate for these candidates was 93% (1,036 of 1,118). The corresponding 7% false positive SNP rate is a composite effect of incorrectly called SNPs and the usual artifacts associated with capillary sequence validation experiments (Quinlan and Marth 2007). The fraction of SNPs missed in the 454 reads at this cutoff was 14.8% (**Methods**). Using the native base calls, the validation rate remained unchanged but 30.0% of the SNPs were missed.

A new, higher throughput version of the 454 pyrosequencing machine (model FLX) has recently been released. Operating under the same sequencing principles as the GS20, this machine produces 250 bp reads. We tested *Pyrobayes*'s performance on two sequencing runs from *E. coli* K12. As shown in **Figure 3.5**, the overall accuracy of the FLX machine is higher than that of the GS20. Therefore, since *Pyrobayes* was calibrated using the GS20 machine, the *Pyrobayes* base quality values underestimate the actual FLX accuracy. Nevertheless, they are clearly more accurate than the native 454 base quality values, even without specific retraining for the FLX model. This suggests that our training method is robust. Yet to get the best performance possible, we will have to repeat our calibration for the FLX and for all subsequent 454 models.

Discussion

We demonstrate that the dominant errors in 454 reads are insertions and deletions, as opposed to substitutions (**Figure 3.4a**). This is because pyrosequencing reads often require that multiple bases and quality values be derived from a single incorporation signal. As a result, the *Phred* base quality value assigned to a base reflects the likelihood that the base in question is, in fact, part of the true DNA sequence, as opposed to an insertion error. Therefore, the most certain base-calls are the first bases in a homopolymer run (**Figure 3.6**). The least certain base calls, on the other hand, are often the last bases in long runs, where the lower quality value reflects the uncertainty about the existence of the last bases. An alternate approach is to assign an average quality value for each base in the homopolymer run (Brockman et al. 2008). The consequence of this method is that all the bases in a homopolymer run are penalized with a lower quality score. Consequently, such approaches are likely to be less sensitive for SNP calling while the *Pyrobayes* approach is more likely to be less specific, especially in shallow sequence coverage.

A decision any base-caller must make is the minimum evidence required before it calls a base. A higher threshold increases the deletion rate and a lower threshold increases the insertion rate. In our experience, the primary source of substitution errors is misalignment. This is most often the result of nucleotide under-calls in the 454 sequence (for an example see the upper alignment in **Figure 3.4b**). Erring towards calling more bases in homopolymer runs often allows us to correct the alignment (lower alignment, **Figure 3.4b**) and substantially reduce substitution errors at the cost of an increased insertion error rate (**Figure 3.4a**). For SNP calling applications this is clearly the logical choice.

We were able to call SNP candidates in our low-coverage 454 *Drosophila* sequences with a high true SNP rate, while missing only 15% of the SNPs in the regions covered by 454 reads. This is half the missed SNP rate that would have been possible with the native 454 quality values, and is attributable to the higher base quality values that *Pyrobayes* assigned to the true polymorphic alleles. On the other hand, higher base quality values nearly always result in a confident SNP call for spurious mismatches caused by misalignments. That is to say, a misaligned base with a lower native 454 base quality value would not be assigned as confident a SNP probability as the same base with the higher *Pyrobayes* base quality. This may explain why we observed only a 21% drop in the rate of spurious SNP calls in our non-polymorphic *iso-1* reads, despite a 60% reduction of the substitution rate (which was calculated without regard to quality value). Although we show that single-read 454 data is useful for SNP discovery, it is unclear if such low-coverage sequences are similarly suitable for short-INDEL discovery.

Given that the native 454 quality values consistently underestimate actual base accuracy (**Figure 3.1a**) one might argue that an alternative to our approach would be to simply re-assign the base quality values according to the higher, measured base accuracy. We tested this and found that although the re-assigned base quality values improve upon the original values, the fraction of high-quality bases remains substantially below those called by *Pyrobayes*, especially in the 20-40 quality range (**Figure 3.1d**). Furthermore, simply reassigning base quality values does not affect how many bases are called, and therefore does not allow one to fine-tune the base caller to minimize misalignments.

The increased accuracy of our base calls and base quality values will likely permit more sensitive biological studies using the 454 machines. We illustrate this for low-coverage, survey-type applications. Even in deeper overall coverage, statistical fluctuations (Lander and Waterman 1988) will result in regions of shallow read depth. The ability to analyze such regions without a loss in accuracy will permit more complete analyses of whole-genome alignments. *Pyrobayes* processes a single run of GS20 reads in less than 40 seconds, and a run of FLX reads in under 120 seconds using minimal computer resources.

Methods

454 sequencing.

Genomic DNA from mouse and chimp BAC clones as well as the *iso*-1 strain of *Drosophila melanogaster* was sequenced by Elaine Mardis and Vincent Magrini at The Washington University Genome Sequencing Center. The *Drosophila melanogaster* DNA was obtained by Chuck Langley (UC Davis) and Andy Clark (Cornell University). Standard 454 sequencing protocols were used.

Determination of base number probabilities (data likelihoods).

We determined the frequency that, given an observed signal from a nucleotide test, the actual number of incorporated bases was 1,2,3..,etc. by aligning 454 reads from a mouse BAC to the known BAC reference sequence. Since we are able to conclude that any observed mismatch between a 454 read and the BAC reference was a sequencing error, the observed frequencies serve as estimates for the data probabilities, in our Bayesian framework--where *s* is the observed nucleotide incorporation signal and *n* is the homopolymer length. The 454 reads were aligned to the BAC reference sequence with our novel, reference-sequence guided alignment software, *Mosaik* (Michael Stromberg, manuscript in preparation).

Estimating the prior homopolymer probabilities.

According to a model of random aggregation of consecutive bases in a DNA sequence, the expectation for the frequency of homopolymers of length *n* is proportional to $1/4^n$. To check the validity of this expectation we computed the frequency of observed homoploymers in the genomes of seven species (*Influenza* Type B, *Escherichia coli* K12, *Pichia stipitis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens*). As **Figure 3.2b** shows this random expectation grossly underestimates the actual frequency of longer homopolymers in the genomes we analyzed. For the prior probability values, we used the average frequency of the eukaryote homopolymer frequencies. The prior probability for a homopolymer of length zero (i.e. the prior likelihood that no base is incorporated in a flow) was tabulated by counting what fraction of the nucleotide tests do not correspond to an actual base in a single run of 454 reads from a mouse BAC shotgun library.

Determination of the most likely number of bases.

Using the data likelihoods and the prior probabilities, we determined , i.e. the base number probabilities, for every possible (up to a rational limit of n=100) base number, according to the following formula: $\Pr(n \mid s) = \frac{\Pr(s \mid n) \cdot \mathrm{Prior}(n)}{\sum_{\text{every possible } n_i} \Pr(s \mid n_i) \cdot \mathrm{Prior}(n_i)}$. The number n for which this posterior probability is highest is the most likely number of bases.

Parent Distribution Fitting.

Longer homopolymeric runs are inherently less frequent than short runs. Regardless of the amount of available testing data, the number of examples for long homopolymeric runs will always be too small for reliable frequency estimation. In the case of our own test data set, there was not sufficient data to estimate the frequency of homopolymers longer than seven-nucleotides. In order to both extend base calling to longer homopolymeric runs and to improve runtime, we replaced the observed data likelihoods with appropriate parent distributions. For this purpose we used non-central Student's *t* distributions. The non-central *t* fit parameters for shorter homopolymers were extrapolated to longer homopolymer distributions.

Base quality assignment.

The base quality value assigned to each base represents the probability that the base is question was, in fact, incorporated within the test. For example, if the most likely number of nucleotides for an observed signal is three, then the base quality value for the first nucleotide in the run of three reflects the probability that, based on the signal, at least one nucleotide was incorporated. Similarly, the second base reflects the probability that there were at least two nucleotides incorporated, and so forth. Consequently, the bases with the highest assigned base qualities come from the first bases in longer homopolymer runs (**Figure 3.2d, Figure 3.4**). The highest base quality value that *Pyrobayes* assigns is 50, which represents an expected error rate of 1 in 100,000. Nominal base quality values above this value are truncated back to 50.

Sequence alignment.

To align next-generation sequencer reads to entire reference genome sequences efficiently and accurately, we developed a new re-alignment and assembly algorithm, *Mosaik*. *Mosaik* uses a hash-based approach (Altschul et al. 1990; Altschul et al. 1997; Kent 2002) for a fast initial read placement, followed by an exhaustive local Smith-Waterman-Gotoh (Smith and Waterman 1981) pair-wise alignment. Each read was placed at the location where it best aligned, as long as minimum alignment criteria were met (> 95% aligned length; > 95% sequence identity). To avoid misalignment due to paralogy, however, if the read can be mapped to another location with a Smith-Waterman alignment score that is at least 80% of the best alignment score, the read is rejected from the alignment altogether.

Overall sequence error rates.

We aligned the *Drosophila melanogaster iso-1* reads base called by both *Pyrobayes* and the native 454 base-caller to the *Drosophila* reference genome sequence. The sequence differences between every aligned read and the reference sequence were tabulated and used to compute the overall, insertion, deletion and substitution error rates for each method.

Measured base quality value calculation.

Using the same *Drosophila melanogaster iso-1* reads aligned to the reference genome, the measured (i.e. actual) base quality for each assigned base quality was derived by tabulating the number of correct and incorrect base calls made with a given assigned base quality. We then computed an actual base quality for the assigned base quality, according to the following formula:

$$Q = -10 \bullet \log_{10}(1 - (\frac{\#correct}{total}))$$

SNP calling and validation.

SNPs were called among base-called sequences from an African isolate of *Drosophila melanogaster* by both *Pyrobayes* and the native 454 base caller. The resulting sequences from each method were aligned to the *Drosophila* genome reference sequence. We then used the *PolyBayes* SNP discovery program to call SNP candidates among the aligned sequences from each base caller. *PolyBayes* used the base quality values from the respective base caller to assign a SNP likelihood to all observed mismatches between the aligned read and the reference sequence.

We submitted the SNPs identified from the sequences called by *Pyrobayes* to experimental validation by PCR-based capillary sequencing. We computed the validation rate from the fraction of SNPs that we identified by *Pyrobayes* but not confirmed by the capillary validation sequences. Ginger Fewell at The Washington University Genome Sequencing Center performed the validation experiments for polymorphism candidates.

Missed SNP rate calculation.

We estimated the missed SNP rate using the capillary sequences produced for candidate SNP validation. First, we used the *PolyPhred* program (Stephens et al. 2006) to call SNPs in the capillary validation traces from the 46-2 isolate. Second, we manually inspected these SNPs and excluded obvious false positives. We treated the remaining *PolyPhred* calls as "true" SNPs. Third, for each "true" SNP, we determined if it was also covered by a base-called read from the same isolate. We counted cases where there was coverage by a base-called read but the SNP was not discovered by *PolyBayes* as missed SNPs. We calculated the missed SNP rate as the number of missed SNPs divided by the number of "true" SNPs for which there was coverage by a base-called read.

Chapter 3 Figures

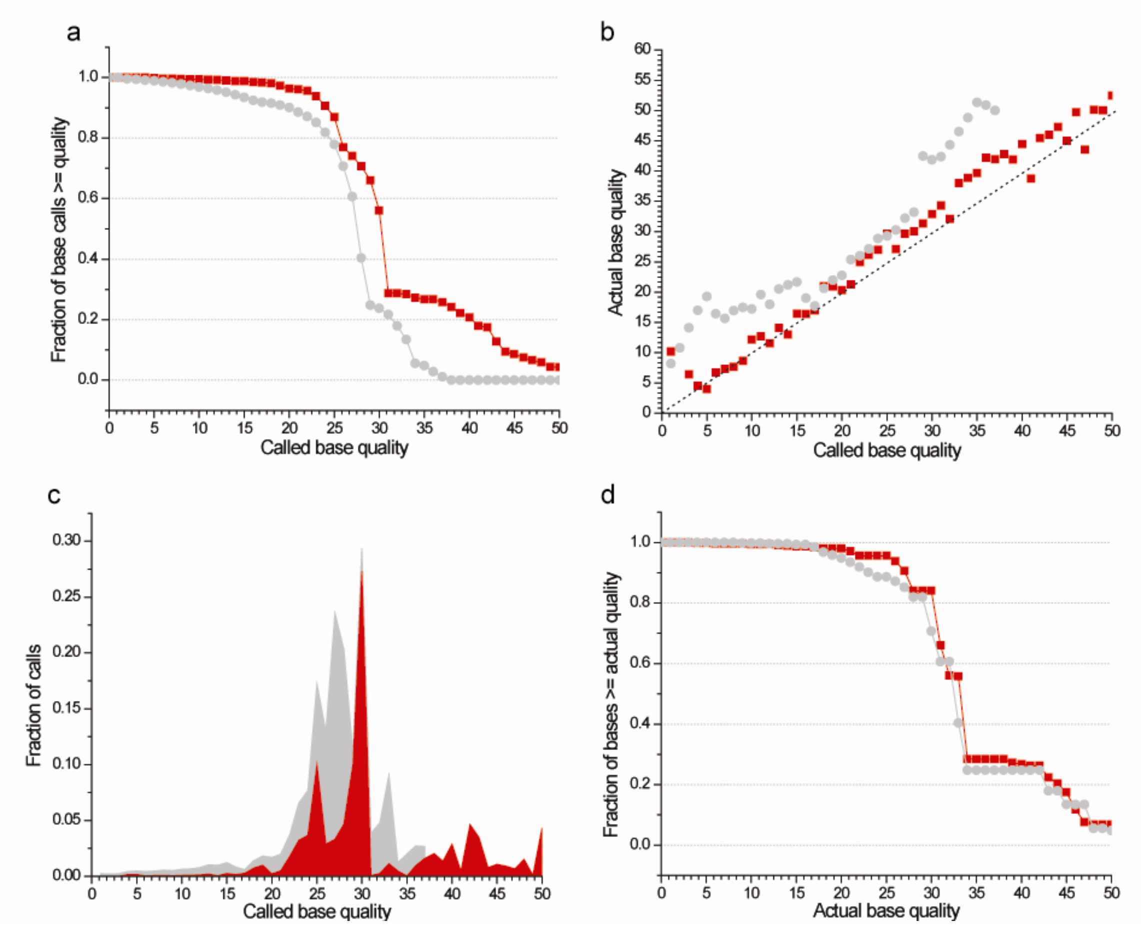


Figure 3.1. Comparison of the base qualities assigned by *Pyrobayes* and the native 454 base caller. a, The cumulative distribution of base quality values assigned by *Pyrobayes* (red) and the native 454 base caller (gray). b, Comparison between assigned base quality value and the base quality calculated from the actual base accuracy for *Pyrobayes* (red) and the native 454 base caller (gray). Actual base quality (Q) is calculated as: $Q = -10 \bullet \log_{10}(1 - (\frac{\#correct}{total}))$: a value of 50 was assigned when no errors were found. c, The distribution of base calls that are assigned a given base quality value by *Pyrobayes* (red) and the native base caller (gray). d, The cumulative distribution of the actual quality of the base calls made by *Pyrobayes* (red) and the native 454 base caller (gray). Actual quality is calculated as above.

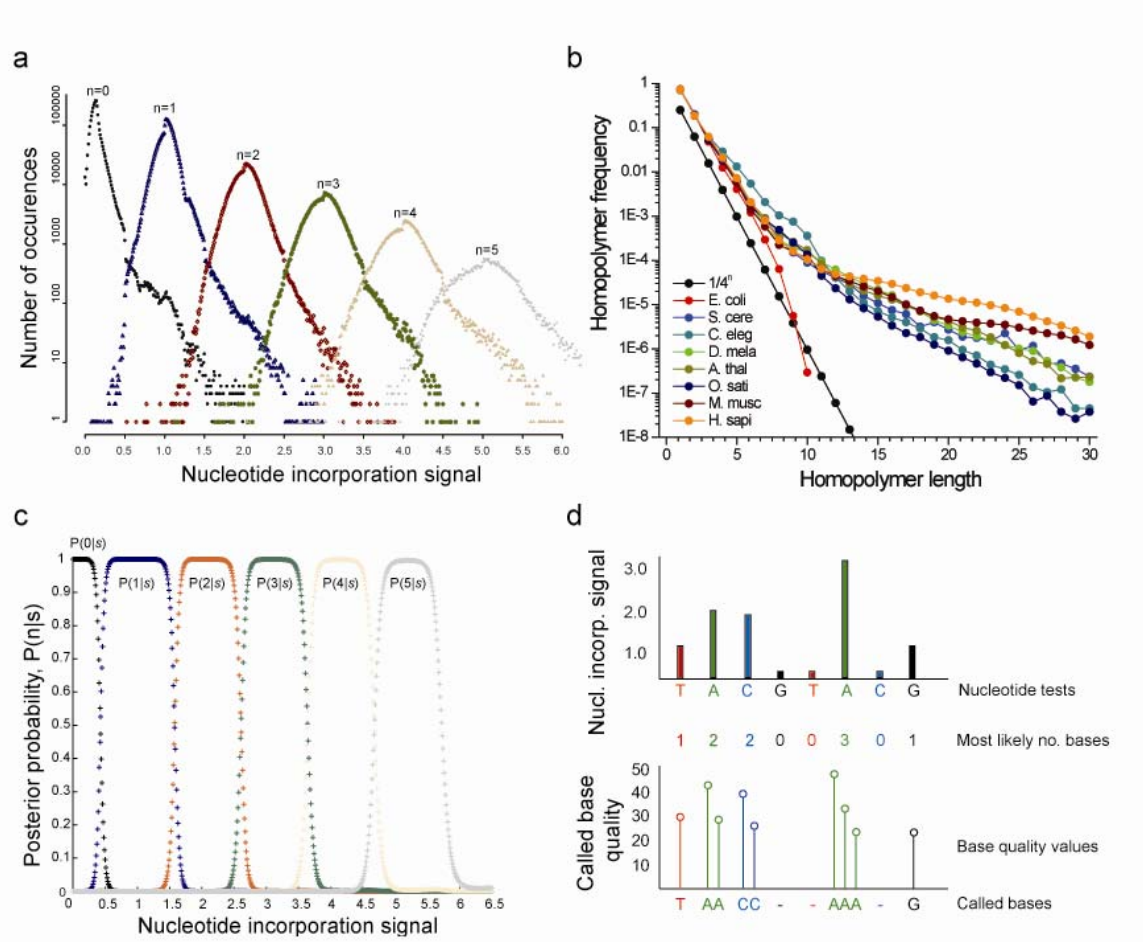


Figure 3.2. The *Pyrobayes* Bayesian base calling approach. **a**, The frequencies that the actual number of nucleotides is n given a nucleotide incorporation signal s are used as data likelihoods. These observed frequencies help to resolve the most likely number of nucleotides for ambiguous signals. **b**, The observed frequencies of homopolymers of varying lengths are show in seven different organisms. For comparison, the expectation of exponential decay is also included. **c**, Using the data likelihoods (panel a) and the prior probabilities (panel b), the posterior probabilities of homopolymers lengths 0 – 5 are shown. **d**, The most likely number of bases is shown for eight consecutive nucleotide tests. The base quality value assigned to each called base is the likelihood that the base in question was part of the DNA sequence.

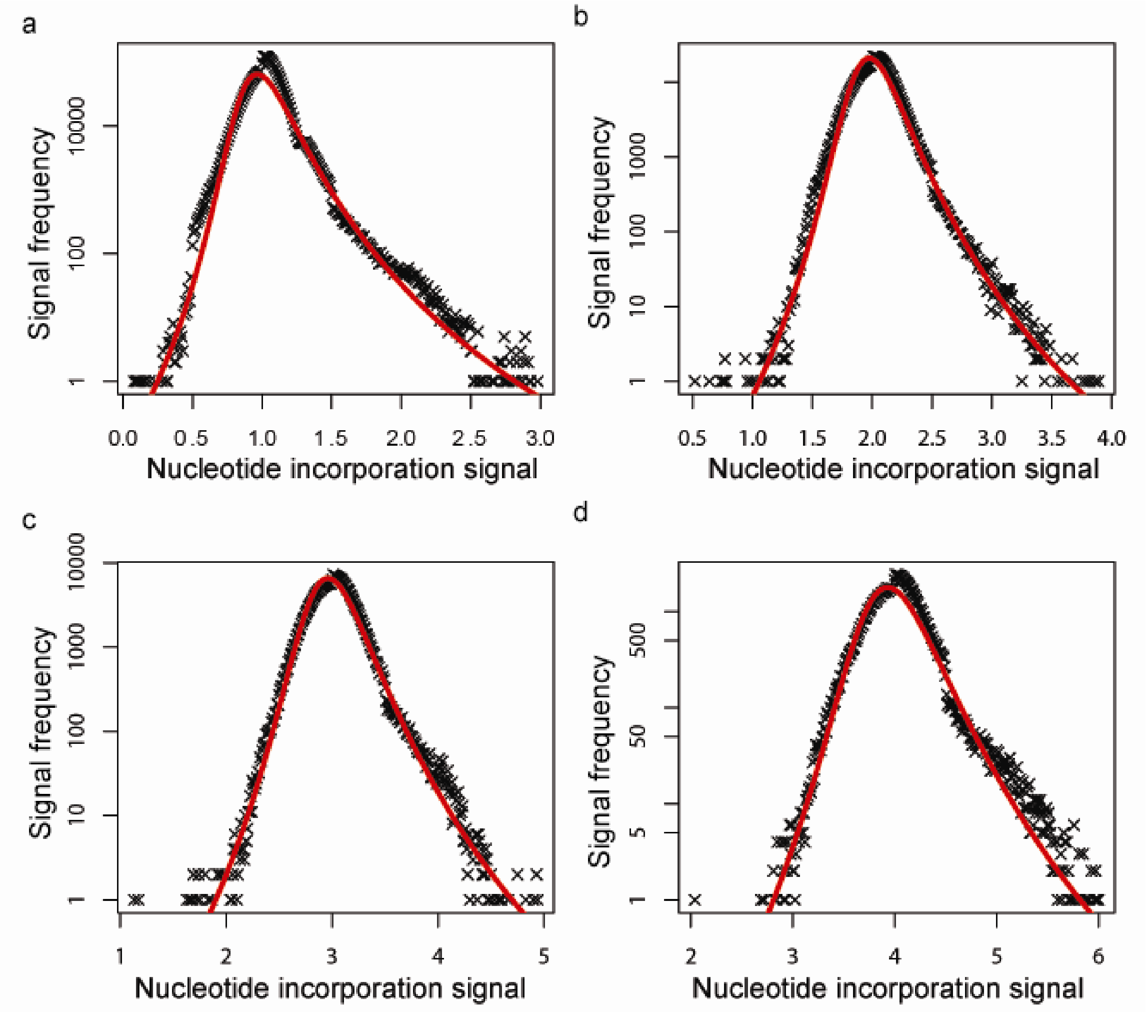


Figure 3.3. **Estimation of data likelihoods with parent distributions**. The observed distributions of homopolymers of length 1 (**a**), 2 (**b**), 3 (**c**) and 4 (**d**) are shown with black x-s. The approximations using non-central Student's *t* distributions are shown in red.

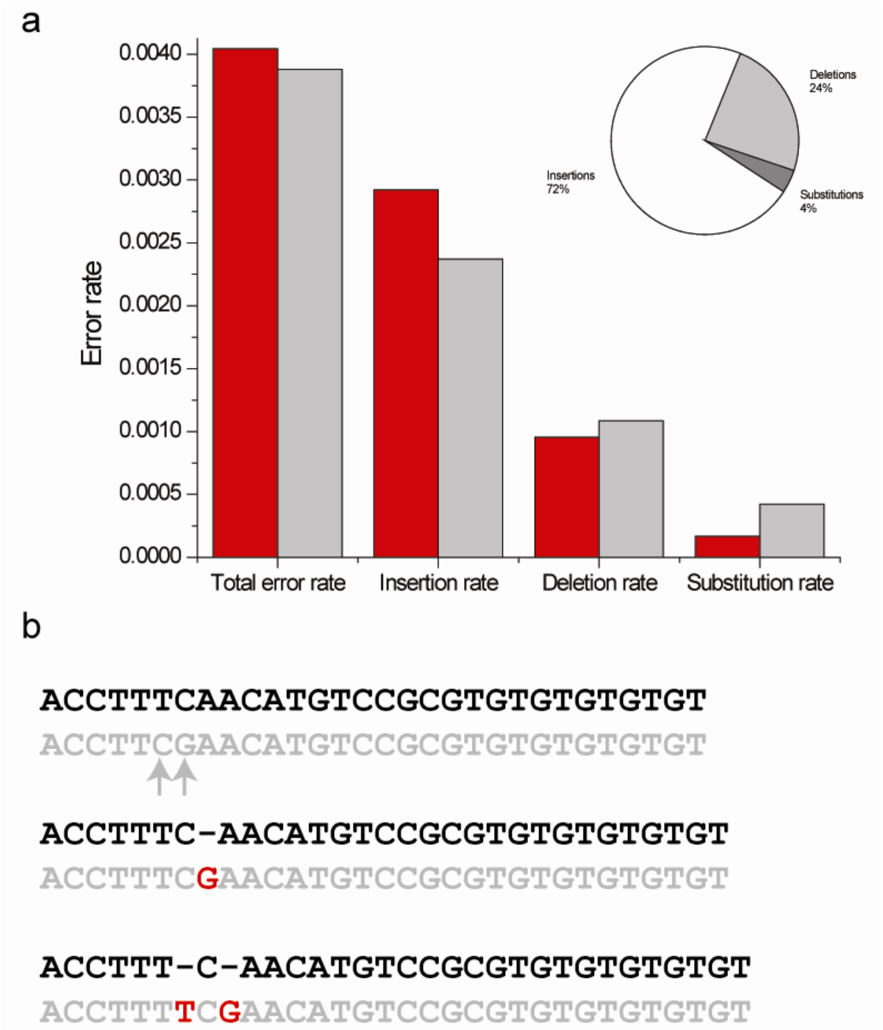


Figure 3.4. Comparison of the error profiles of *Pyrobayes* and the native base caller. a, The overall, the insertion, the deletion and the substitution error rates are shown for *Pyrobayes* (red) and for the native 454 base caller (gray). The relative contribution of each error type is also shown in a pie chart (for *Pyrobayes* calls). **b**, Illustration of the effects of calling too few or too many bases on the alignment of a read (gray) to the reference sequence (black). Top panel: too few thymines (Ts) were called, resulting in two spurious mismatches (arrows) by mis-aligning the correctly called C and the inserted G (red) in the 454 read. Middle panel: the correct number of Ts were called resulting in the correct read alignment of the single insertion error (red) in the 454 read. Bottom panel: too many Ts were called resulting in the correct read alignment of the two base insertion errors (red) in the 454 read.

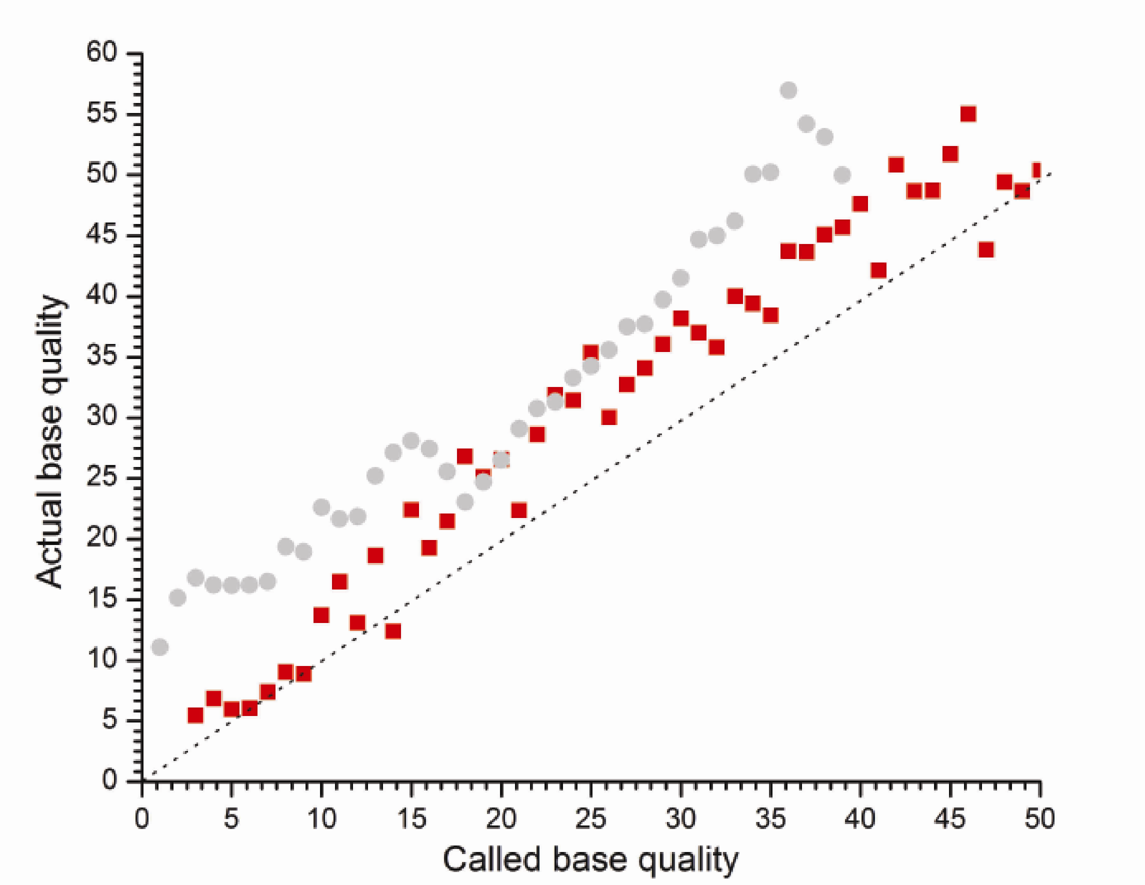


Figure 3.5. *Pyrobayes* base quality accuracy for the 454 Life Sciences FLX model. A comparison between assigned base quality value and the base quality calculated from the actual base accuracy for *Pyrobayes* (red) and the native 454 base caller (gray) is shown for the FLX model.

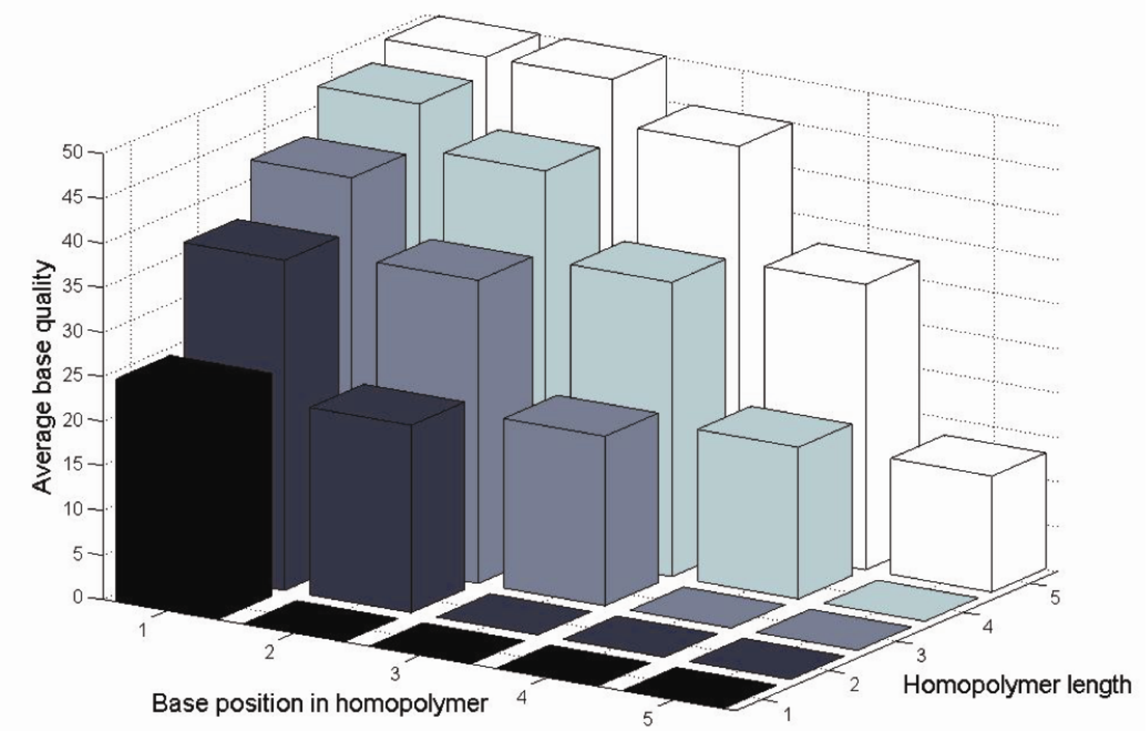


Figure 3.6. Distribution of base quality scores. The average quality score assigned by *Pyrobayes* is shown for each base position in homopolymers up to a length of five bases. For example, on average, the second base in a run of four identical nucleotides is assigned a quality value of 43.

My motto is: Contented with little, yet wishing for more.

-Charles Lamb

4. *Whole-genome polymorphism discovery in ten Drosophila melanogaster isolates using 454 pyrosequences*

Abstract

The recent development of novel, high-throughput sequencing technologies promises faster, more economical approaches to genomic studies (Bentley 2006; Margulies et al. 2005; Shendure et al. 2004). The pyrosequencing technology from 454 Life Sciences (Margulies et al. 2005) currently produces hundreds of thousands of 100-250 base pair sequences from whole-genome shotgun libraries. Although other, short-read sequencing machines produce more bases per run, the longer 454 sequence reads are easier to assemble and align to a reference genome. Individual re-sequencing studies seek to uncover polymorphisms in the most economical way (i.e. at the lowest possible sequence coverage). However, this has not been successful in low-coverage 454 sequences to date because of the difficulty in determining the number of actual bases in homopolymer runs (Ahmadian et al. 2006; Margulies et al. 2005;

Ronaghi et al. 1996). Here we report improved algorithms for basecalling, alignment and SNP calling using a refined error model, and apply these tools to a genome-wide SNP discovery project based on light-shotgun 454 pyrosequences from ten *Drosophila melanogaster* isolates. This study demonstrates that even single-coverage 454 reads are suitable for accurate polymorphism discovery, with an independent experimental validation rate of 93%, while missing only 2% of the existing variation. This approach represents an economical approach to genome-wide polymorphism surveys, significantly increases the rate and economy of SNP discovery, and facilitates rapid marker generation for genotyping chips. The accuracy achieved in single 454 reads suggests that, with increased overall sequence coverage, this approach will be suitable for complete mutational profiling of model organisms.

Introduction

The recently developed next-generation sequencing technologies produce hundreds of megabases to gigabases of short (less than 50 bp) and medium-length (100-250 bp) reads (Bentley 2006; Shendure et al. 2004). Although the short-read technologies (e.g. Illumina, AB/SOLiD) have much higher throughput, the sequencing instruments with medium-length reads (e.g. 454 Life Sciences) are more suitable for de novo sequencing and less susceptible to mis-alignment owing to paralogy. The 454 machines have been used successfully for genome sequencing (Margulies et al. 2005), microRNA discovery (Girard et al. 2006), individual mutation detection (Thomas et al. 2006) and bacterial resequencing (Velicer et al. 2006). These applications required deep read coverage which, despite the higher throughput of this technology, is still costly for large genomes. Therefore it is imperative that informatics tools extract the most information from the lowest possible sequence coverage. A typical low-coverage sequencing application is a genome survey for the estimation of nucleotide diversity among individuals or strains. To assess whether or not the 454 technology is suitable for such low-coverage re-sequencing projects we undertook a genome-wide survey of ten inbred *Drosophila melanogaster* isolates, using the high-quality *Drosophila* reference genome sequence as a template for read alignment.

In low coverage resequencing applications, the vast majority of reads align to the reference as singletons, and thus polymorphism discovery is only possible if the base quality of the reads (and of the reference sequence) is high enough to distinguish true polymorphisms from sequencing errors (Marth et al. 1999). As described in Chapter 4, we see that the majority (76%) of bases called by the 454 base caller have base quality values lower than 30 (i.e. they are reported to have more than a 1 in 1,000 bp error rate), and none of the called bases have a quality value over 40. Given that this error rate is comparable to typical pair-wise polymorphism rates, SNP discovery with such quality values comes at the cost of frequent false positive SNP predictions.

Results

The sequenced Drosophila isolates.

We sequenced the ten inbred *Drosophila* isolates, four from Malawi and six from North Carolina, each with a single run of the 454 GS20 pyrosequencer (**Table 4.1**) from whole-genome shotgun libraries (**Methods**). For each isolate, an average of 337,989 reads were produced, providing a 0.195-fold coverage of the 180 Mb genome (Adams et al. 2000) (see **Table 4.1** for details). The sequenced isolates were derived from North America and Africa and there was ample prior information on these populations to expect several nucleotide differences per kilobase (Hoskins et al. 2001). In order to develop an empirical error model for the basecalling algorithm, we also collected one 454 GS20 run of the *iso-1* reference strain (Adams et al. 2000), providing nearly 30 Mbp of sequence calls for which the correct sequence was known.

Basecalling, alignment and SNP discovery.

We basecalled the reads with *Pyrobayes* and aligned them to the euchromatin of the reference genome sequence with our new reference sequence-guided assembly program, *Mosaik* (see **Methods**, manuscript in preparation). We aggressively discarded sequences that mapped to multiple locations in the reference genome to avoid misalignments that could lead to spurious SNP calls. Additionally, we required that 95% of the entire length of each sequence was aligned and we allowed very few mismatches relative to the genome (**Methods**). On average 57.4% of the reads were aligned from each isolate (see **Table 4.1**). This seemingly low fraction is largely a consequence of the fact that we only aligned sequence reads to the *Drosophila* euchromatin which is roughly two-thirds of the total genomic sequence (120Mb of 180Mb). In other words, given that the sequenced DNA is from the entire genome, we would expect at most 67% of the reads to align. 70.2% of the euchromatin was covered by a read from at least one isolate (see **Figure 4.2**) and 36.9% was covered by at least two reads. We scanned the alignments for SNPs with an improved version of our SNP discovery program, *PolyBayes*, allowing us to efficiently process the 1.9 million aligned sequences. *PolyBayes* assigned a SNP probability to each candidate SNP (see **Methods** and **Figure 4.1**). We identified every candidate with a SNP probability greater than 0.01, yielding 593,315 candidate SNPs. The vast majority of these candidates (92.7%) had a posterior probability above 0.5 and over half (54.3%) were above 0.9.

To assess if such high SNP probability values are justified, we subjected 1,338 randomly chosen candidates from one of the ten isolates to experimental verification with PCR-based capillary sequencing (see **Methods**, **Figure 4.1**). Assays were successful for 1,317 candidates, and we confirmed 1,220 SNPs. This represents a 92.6% validation rate (i.e. the number of true SNPs divided by the number of candidates that could be conclusively assayed) and a 91.2% conversion rate (i.e. the number of true SNPs divided by all candidates submitted for validation). We also quantified the rate of missed SNPs by tabulating all SNPs in the capillary validation traces for which there was also coverage by at least one 454 read (**Methods**). Of 1,183 such SNPs only 92 were missed; a 7.8% missed SNP rate. Many of the SNPs that were missed were heterozygotes where the 454 read(s) only contributed the reference allele.

Clearly, without a 454 read with the alternate allele, it is impossible to identify these heterozygous polymorphisms. Excluding such heterozygotes, only 26 (2.3%) of 1,117 SNPs were missed. As expected, raising the SNP probability cutoff increases specificity (the validation rate is higher) but decreases sensitivity (more SNPs are missed), and vice versa (**Figure 4.3a, b**).

Pair-wise nucleotide diversity estimates.

Although a single sequencing run was insufficient to resequence each isolate's entire genome, there was sufficient overlap between isolates to estimate the inter-isolate pair-wise nucleotide diversity values (**Table 4.2**). The average value, 5.5 x 10-3, or 1 SNP per 181 bp, agrees with previous estimates (Hoskins et al. 2001). The nucleotide diversity was higher among the African isolates (1 SNP per 183 bp) than among North Carolina isolates (1 SNP per 214 bp), and highest across the two geographic cohorts (1 in 165 bp). Because of the large amount of read overlap between isolates, the diversity estimates are accurate within 1% (maximum standard error is 6.6 x 10-5).

Given the low sequence coverage in this study we expected that many of the SNPs would be singletons (i.e. the alternate allele is only present in one isolate).

Nevertheless, 18.4% (108,872) were polymorphic in two or more of the ten sequenced isolates (**Figure 4.4**).

Discussion

The high validation and low missed SNP rates we report in this study suggest that 454 pyrosequences are suitable for accurate and exhaustive SNP discovery even in single-read coverage, and therefore this technology is an economical yet informative alternative to traditional genome survey sequencing. The validation rates we observed were higher than would be predicted by the SNP probabilities calculated from the base quality values in the 454 reads. This is the result of the fact that these base quality values, by necessity, agglomerate sequencing error rates from insertions and substitutions, whereas insertion errors are the dominant error type in 454 reads (Huse et al. 2007; Margulies et al. 2005) (**Figure 4.5a**). However, once a base is aligned to the reference sequence as a mismatch it is *de facto* confirmed not to be an insertion, and the error rate calculated from the quality value is now an overestimation of the substitution rate. This illustrates the difficulty of representing the 454 sequencing error rate with a single *Phred* quality value, which was developed to describe substitution-type error in Sanger sequences. To remedy the overestimation of the substitution error rate, we are extending *Pyrobayes* to produce separate quality values for insertion and substitution errors. However, such effort is only useful if downstream software is able to interpret and use such separate quality values sequences.

In **Figure 4.3a**, the validation rate for SNP candidates with a probability of 0.9 or greater declined relative to the validation rate of slightly less probable candidates. In the majority of these cases, the alternative allele is an extra base in a homopolymer in the 454 read, aligned to a nucleotide of another kind in the reference sequence. Depending on whether the read is from the same or the opposite strand compared to the reference sequence, the extra base is the first base in the run (and assigned high base quality value) or the last base (and assigned low base quality value); see **Figure 4.5b**. The validation rate of such candidates represents an aggregate of these two situations. This phenomenon explains the decrease in the validation rate for SNP candidates with a probability of 0.9 (**Figure 4.3a**). We are currently extending our methods to make the SNP calling software aware of the alignment orientation which, in turn, would allow us to produce more accurate SNP probabilities for this class of candidates.

Additionally, we often find that spurious SNP calls arise in situations where the alternate allele is found in multiple reads having identical start and end sites (**Figure 4.5c**). It is highly unlikely that such reads represent two different DNA fragments, especially in low shotgun read coverage. It is much more likely that such reads are the result of emulsion PCR amplification errors incurred on a single template that was subsequently sequenced in multiple sequencing wells. One potential remedy is to exclude sequences with identical alignment start and end positions from the analysis, although this approach may be wasteful.

Genetic and comparative genomic studies of organisms whose genome has been sequenced are further empowered by dense marker maps and SNP genotyping chips requiring markers that segregate in the population. Nearly 20% of our SNP candidates were polymorphic in two or more isolates, representing an average density of 1 SNP per 1,100 base pairs of the melanogaster euchromatin. Clearly, these SNPs comprise a dense *Drosophila melanogaster* genetic marker map, and a useful candidate pool for a fruit-fly genotyping chip.

Based on the low missed SNP rate in this low-coverage genome survey, we anticipate that our tools will be able to find every SNP if complete genome coverage is available, as would be required for complete mutational profiling or individual resequencing. For applications where false negatives are easily tolerated, shallow coverage 454 resequencing and the algorithms described here provide a much cheaper solution for genome-wide SNP surveys.

Methods

454 sequencing.

Elaine Mardis and Vincent Magrini at The Washington University Genome Sequencing Center sequenced genomic DNA from the iso-1 strain as well as the 10 geographic isolates of *Drosophila melanogaster*. The *Drosophila melanogaster* DNA was obtained by Chuck Langley (UC Davis) and Andy Clark (Cornell University). Standard 454 sequencing protocols were used.

Whole-genome DNA library preparation for each isolate

A whole-genome shotgun library was created for each isolate from genomic DNA obtained from Charles Langley at the University of California, Davis. We fragmented the genomic DNA by nebulization according to standard 454 protocols. Nebulized DNA was analyzed by agarose gel electrophoresis, and we collected fragments within a size range of 500 bp. The collected fragments were linker ligated with a mixture of the two 454-specific linkers, one species of which is biotinylated. We then performed an enrichment step to remove fragments with the same species of un-biotinylated linker at both ends, by capturing those with biotinylated linkers on streptavidin magnetic beads.

Next, the fragments on the beads were denatured, allowing us to reclaim the non-biotinylated strand from a supernatant that is further utilized. First, the released single-stranded DNA fragments were run on the Agilent Bioanalyzer to calculate yield, then coupled to Sepharose beads that carried covalently linked oligonucleotides complementary to the linkers ligated onto the nebulized DNA fragments. Here, we adjusted the input concentration of DNA fragments to give, on average, a 1:1 association between beads and DNA fragments. The mixture was then emulsified in an oil suspension containing aqueous PCR reactants, and emulsion PCR enabled the amplification of millions of unique fragment-bead combinations in a large-batch PCR format. After combining the emulsion PCR (emPCR) reactions for the library, we enriched for Sepharose beads that contained amplified DNA, by the use of streptavidin magnetic beads to capture the biotinylated ends of amplified fragments complexed to Sepharose beads.

Following enrichment, the biotinylated strand is melted away by the addition of NaOH, and sequencing primers were annealed to the bead-bound amplicons.

Primer- and polymerase-bound Sepharose beads were loaded into a PicoTiterPlate (PTP) device that is essentially composed of hundreds of thousands of fused fiber optic strands, the ends of which are hollowed out to a diameter sufficient to contain a single Sepharose bead. Smaller magnetic beads, to which pyrosequencing (sulfurylase and luciferase) enzymes are covalently attached, were pipetted into the PTP subsequently, and a centrifugation step packed them in around each Sepharose bead. The PTP fits into a flow-cell device that positions it against a high-sensitivity CCD camera in the 454 GS-20 sequencing instrument. Pyrosequencing follows, whereby sequential flows of each dNTP, separated by an imaging step and a wash step take place. At each well address in the PTP, the incorporation of one or more nucleotides into the synthesized strand on each bead was captured by the CCD camera, which records positional information about each well address that reports a signal during the initial flow cycles and then monitors all addresses throughout the sequencing process.

Sequence alignment and assembly

Using *Mosaik, e*ach sequencing read was placed at the location where it best aligned, as long as minimum alignment criteria were met (>=95% aligned length; >=95% sequence identity). To avoid misalignment due to paralogy, however, if the read can be mapped to another location with a Smith-Waterman alignment score that is at least 80% of the best alignment score, the read is rejected from the alignment. 744,955 (22%) of the reads from the ten isolates were rejected in this manner. We aligned all 10 runs of 454 reads simultaneously. We used the reference-guided assembly functionality of *Mosaik* to create multiple alignments of reads from each isolate.

SNP calling

Using a new version of the *PolyBayes* SNP discovery algorithm that was rewritten and optimized for millions of short and medium-length sequences, we scanned the 454 read alignments for SNPs between the ten isolates and the *iso-1* reference genome. *PolyBayes* screens for single-nucleotide mismatches between aligned 454 reads and the reference sequence. The base quality values for the reference sequence and for all aligned alleles are used to calculate a SNP probability score (**Figure 4.1**). **Figure 4.6** illustrates that when an alternate allele from a single 454 read is aligned to the reference sequence, the SNP probability score increases as a function of the base quality value of the alternate allele. When reads from multiple isolates were aligned at the same genomic position, *PolyBayes* uses the alleles and quality values from each aligned read to calculate the likelihood that the locus is polymorphic among the aligned isolates.

Given that it is a highly accurate, finished genome, we assigned a quality value of 40 (1 error in 10,000 bp) to each base in the *Drosophila* reference sequence. Based on previous diversity estimates, we used a prior pair-wise polymorphism rate of .005 and we identified all SNP candidates for which the *PolyBayes* SNP probability score exceeded 0.01.

SNP validation

We chose 10,000 sequences at random from the African isolate 46-2, and submitted the 1,483 SNP candidates found in these reads to validation by PCR-based capillary sequencing. We were able to design PCR amplicons containing the candidate site for 1,466 of the candidates. These were amplified with universal primer-tailed oligo-nucleotides, and sequenced with two reads on each strand using ABI capillary machines. We then aligned the capillary traces to the candidate SNP site and manually examined whether the alternate allele found in the 454 read was confirmed in the corresponding capillary traces. Ginger Fewell at The Washington University Genome Sequencing Center performed the validation experiments for polymorphism candidates.

Missed SNP rate calculation

We estimated the missed SNP rate using the capillary sequences produced for candidate SNP validation. First, we used the *PolyPhred* program to call SNPs in the capillary validation traces from the 46-2 isolate. Second, we manually inspected these SNPs and excluded obvious false positives. We treated the remaining *PolyPhred* calls as "true" SNPs. Third, for each "true" SNP, we determined if it was also covered by a 454 read from the same isolate. We counted cases where there was 454 coverage but the SNP was not discovered by *PolyBayes* as missed SNPs. We calculated the missed SNP rate as the number of missed SNPs divided by the number of "true" SNPs for which there was 454 coverage.

Chapter 4 Tables

Isolate	**Origin**	**Num. of sequence reads**	**Total sequence**	**Avg. sequence read length**	**Nominal genome coverage**	**Num. sequence reads aligned passing alignment criteria**	**Fraction of sequence reads aligned**
28-5	Malawi	264,107	27,259,496	103.2	0.151	145,344	0.550
46-2	Malawi	341,600	35,051,630	102.6	0.195	197,055	0.577
56-4	Malawi	383,430	40,456,875	105.5	0.225	207,620	0.541
63-5	Malawi	325,694	33,583,151	103.1	0.187	174,748	0.537
301	N.C.	436,406	45,555,185	104.4	0.253	240,344	0.551
303	N.C.	333,418	34,686,165	104.0	0.193	194,066	0.582
306	N.C.	311,569	32,497,974	104.3	0.181	177,709	0.570
358	N.C.	360,650	37,531,671	104.1	0.209	231,878	0.643
375	N.C.	346,835	35,418,959	102.1	0.197	203,587	0.587
732	N.C.	276,176	28,711,855	104.0	0.160	167,185	0.605
	Total	***3,379,885***	***350,752,961***			***1,939,536***	
	Avg. / isolate	***337,989***	***35,075,296***	***103.8***	***0.195***	***193,954***	***0.574***

Table 4.1. Summary statistics for the sequenced D. melanogaster isolates.

	28-5	**46-2**	**56-4**	**63-5**	**301**	**303**	**306**	**358**	**375**	**732**
28-5	--	0.00575	0.00532	0.00476	0.00582	0.00617	0.00595	0.00599	0.00606	0.00601
46-2		--	0.00583	0.00568	0.00620	0.00630	0.00614	0.00624	0.00635	0.00624
56-4			--	0.00537	0.00596	0.00602	0.00618	0.00621	0.00616	0.00611
63-5				--	0.00580	0.00600	0.00590	0.00580	0.00611	0.00603
301					--	0.00479	0.00466	0.00453	0.00471	0.00487
303						--	0.00344	0.00484	0.00480	0.00479
306							--	0.00480	0.00462	0.00449
358								--	0.00477	0.00503
375									--	0.00488
732										--

Table 4.2. Pair-wise polymorphism rates between each D. melanogaster isolate.

Chapter 4 Figures

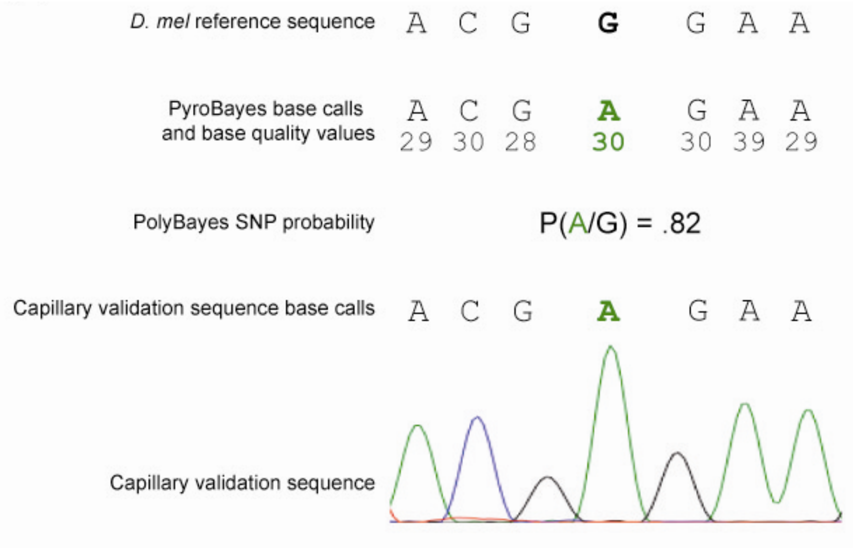

D. mel reference sequence
A C G G G A A
PyroBayes base calls
and base quality values
A C G A G A A
29 30 28 30 30 39 29
PolyBayes SNP probability
P(A/G) = .82
Capillary validation sequence base calls
A C G A G A A
Capillary validation sequence

Figure 4.1. Using base qualities in SNP discovery. A SNP candidate and a corresponding validation trace are shown. The SNP probability reflects the quality of the alternate allele. The SNP is confirmed in the capillary validation trace.

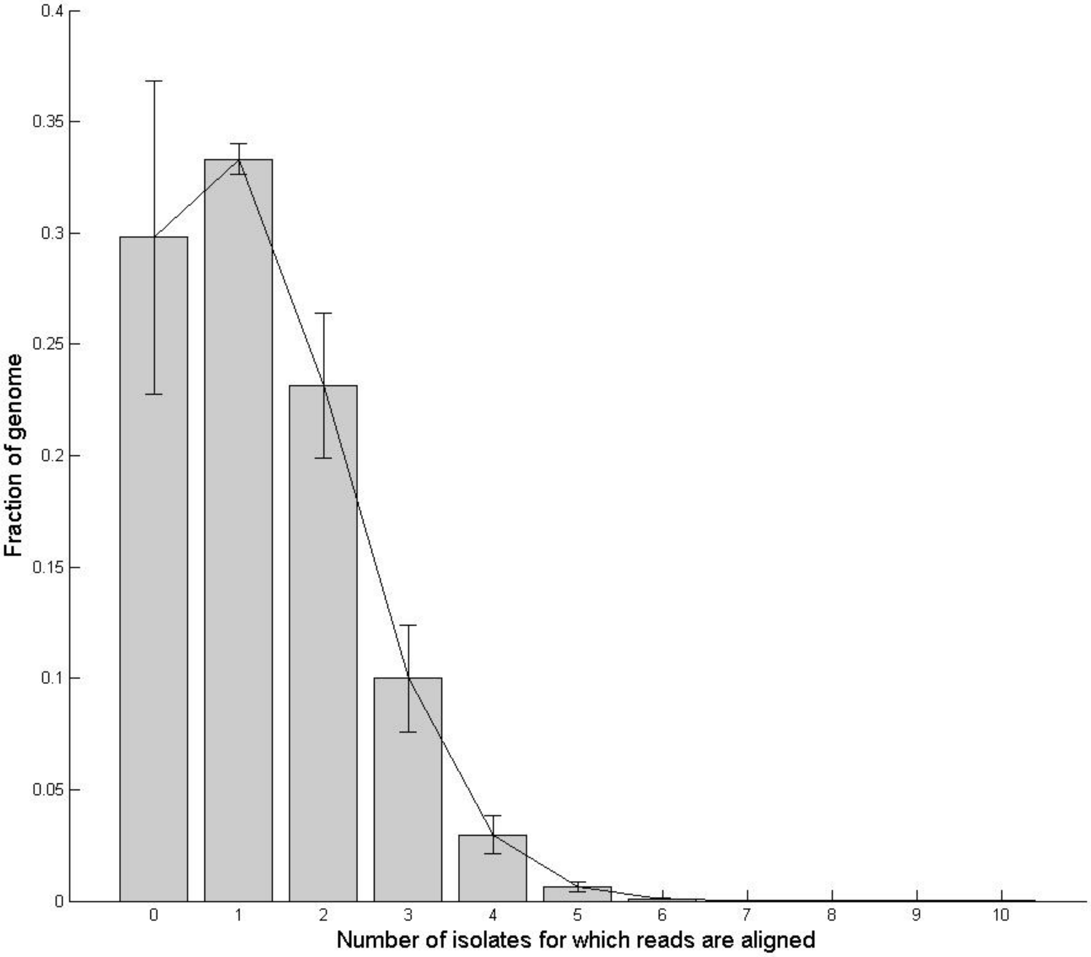


Figure 4.2. Depth of isolate-specific read coverage. The average fraction of the genome with aligned reads from 0, 1, 2, …, etc. isolates are shown. Error bars indicate the standard deviation among the melanogaster chromosomes.

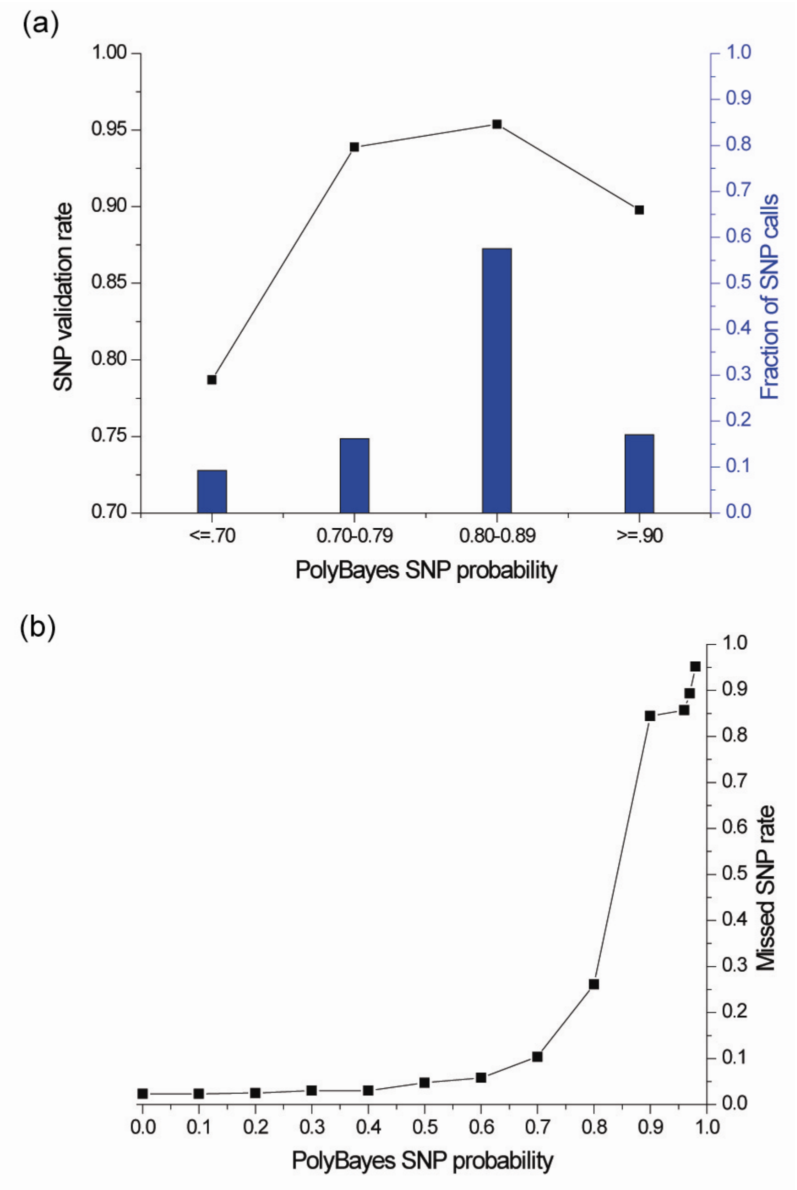


Figure 4.3. Validation and missed SNP rates in single sequence coverage. **a**, The SNP probability is compared to the experimental validation rate (black). The blue columns indicate the fraction of SNP candidates in each SNP probability bin. **b**, The missed SNP rate is shown as a function of the SNP probability cutoff.

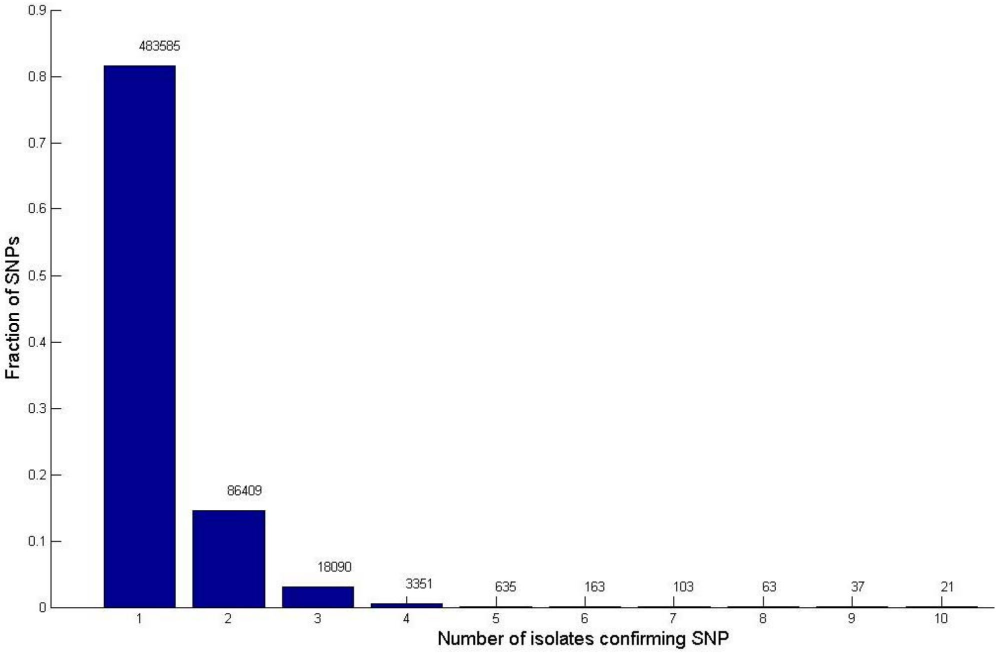


Figure 4.4. The fraction of SNP candidates identified in one or more Drosophila lines.

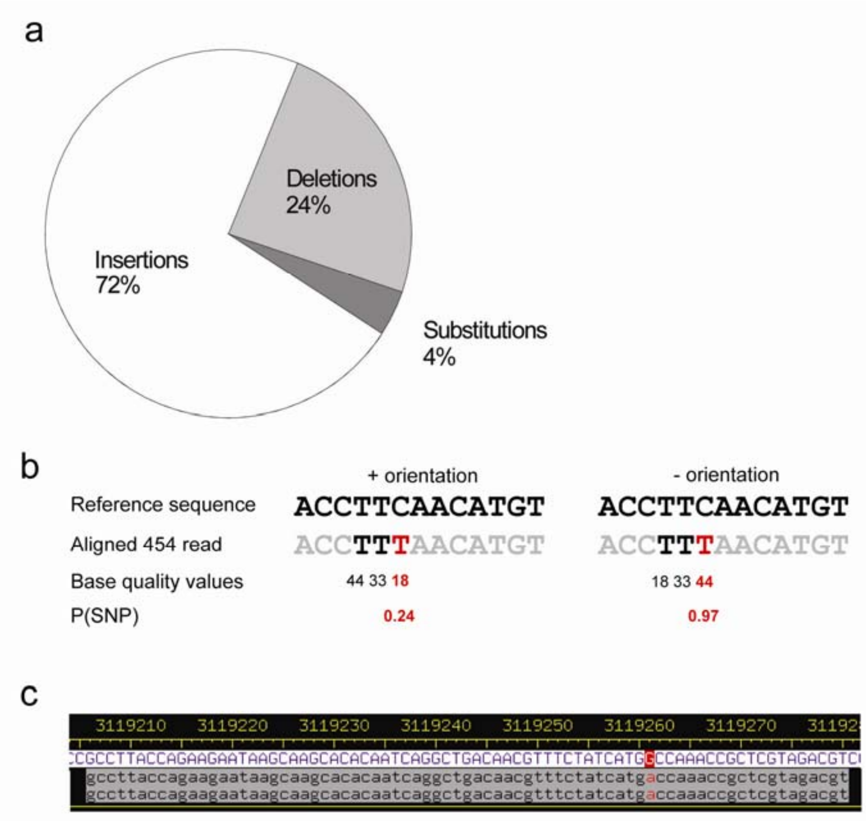


Figure 4.5. Common factors that affect the SNP validation rate. **a**, The contribution of each source of base calling error is shown for base calls made by *Pyrobayes*. **b,** Putative SNP candidates and their associated quality values and *PolyBayes* SNP probabilities (P(SNP)) are shown when a 454 read is aligned in the same orientation as the reference sequence (left pane) and in the opposite orientation as the reference sequence (right pane). **c**, An example of a spurious SNP candidate potentially arising from an error during emulsion PCR amplification.

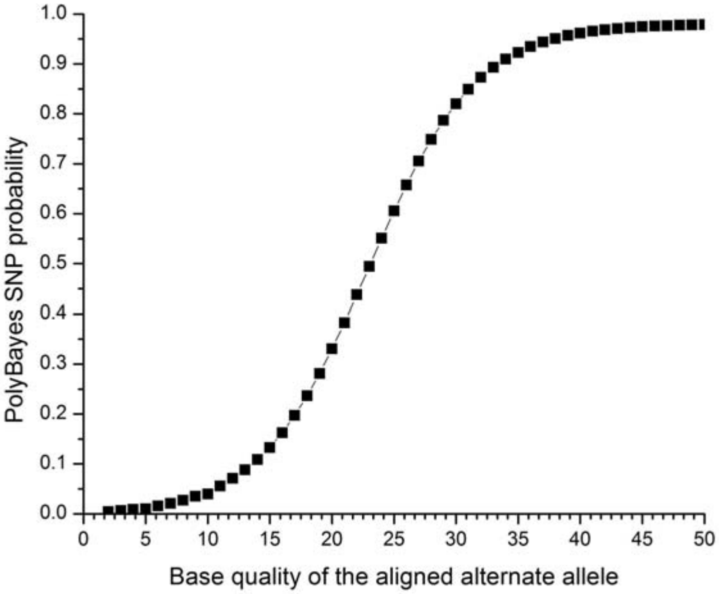


Figure 4.6. **The relationship between base quality and SNP probability.** For singly-aligned 454 reads, the *PolyBayes* SNP probabilities are shown as a function of the base quality of alternate allele.

Progress in science depends on new techniques, new discoveries, and new ideas, probably in that order.

-Sydney Brenner

5. Whole Genome Sequencing and SNP Discovery for C. elegans using massively parallel sequencing-by-synthesis.

Abstract

Next-generation sequencing instruments enable rapid and inexpensive DNA sequencing at unprecedented levels. Because these instruments are so new, their sequence data require characterization (read error profiles, base quality values, coverage models, and general utility). We resequenced the Bristol (N2) strain with the Illumina/Solexa sequencing technology in order to understand it's inherent error types and error rates. An immediate application of this technology is individual or strain resequencing in order to discover genome-wide sequence differences. Since this technology produces relatively short sequences, we developed a novel approach to assess the fraction of a genome that can be resequenced with short reads. We additionally compared Illumina/Solexa reads from the CB4858 strain of *C. elegans* to the N2 reference sequence using a novel sequence alignment program and screened for single

nucleotide polymorphisms (SNPs) and small INDELs with a vastly more efficient version of *PolyBayes*. This study is the first to broadly characterize the error profile of the Illumina/Solexa technology and to demonstrate the utility of massively parallel short read sequencing for whole genome resequencing and for accurate discovery of genome-wide polymorphisms.

Introduction

In 1998, a special issue of *Science* celebrated a landmark in biology; the decoding of the first animal genome sequence, that of the model organism *Caenorhabditis elegans* (Stein et al. 2003). First suggested as a model organism in the 1960's by Sydney Brenner, and due to the pioneering work of John Sulston, Alan Coulson and Robert Waterston to produce a physical map of its genome, the *C. elegans* genome sequencing project formed the cornerstone of efforts ultimately aimed at decoding the human genome (Lander et al. 2001). The entire *C. elegans* biology community has benefited enormously from the availability of the genome sequence and its ever-improving genome annotation (Chen et al. 2005; Harris et al. 2004; Harris et al. 2003), not to mention the comparative power of sequenced close relatives such as *C. briggsae* (Stein et al. 2003).

The emerging availability of massively parallel sequencing instrumentation is providing the capability to resequence genomes in a fraction of the time, effort and expense than ever before. Compared to capillary sequencing, these instruments produce relatively short read length sequences, using a combination of novel sequencing chemistry and fragment libraries that do not utilize a bacterial intermediate. Because of these differences, important aspects of resequencing remain uncharacterized, including read error profiles, base quality values, coverage models, approaches for read mapping to reference genomes, and the general utility of short read sequences. To address these, we revisited the *C. elegans* genome sequence, using the Illumina/Solexa 1G Sequence Analyzer to resequence a laboratory isolate of the *C. elegans* N2 Bristol strain. Our analyses of N2 Bristol sequences included a thorough description of the Solexa read error model and an evaluation of sequence differences between the resequenced N2 isolate and the N2 reference genome sequence. We revealed both possible sequencing errors in the original *C. elegans* reference genome, and putative sequence variants that had occurred in our passaged N2 Bristol strain since the reference genome was sequenced.

An immediate application of massively parallel sequencing is the comparison of an individual or strain to that species's reference genome sequence in order to reveal genome-wide sequence differences either for comparative and evolutionary studies or for discovering genetic variants. Given the short read length of the Illumina/Solexa technology (32 bp/read in this study), such studies require the assessment of what fraction of a genome can be resequenced with short read sequences. This is because short reads are more susceptible to multiple genome mapping locations than are sequences of 500+ bases from capillary instruments. Computational identification of these short repeats (we term them 'microrepeats') is therefore an important consideration for accurate short read characterization and analysis, and must include an allowance for mismatches due to sequencing errors or polymorphism. Here, Solexa sequence reads from the CB4858 strain of *C. elegans* (originally isolated in Pasadena, CA) were aligned to the microrepeat-masked N2 Bristol reference sequence using the *Mosaik* aligner/assembler. The aligned reads were then screened for single nucleotide polymorphisms (SNPs) and small INDELs with *PolyBayes*. Our results demonstrate the utility of massively parallel sequencing for whole genome resequencing and for accurate discovery of both single nucleotide and small insertion-deletion polymorphisms. This work establishes a framework for human genome resequencing toward similar purposes.

Results

Metrics of single end Solexa sequencing.

We completed a total of five Illumina/Solexa sequencing runs (**Methods**) yielding 3.5 runs of sequence for the resequenced N2 strain and 1.5 runs of sequence for the CB4858 strain from Pasadena. As described in **Table 5.1**, the five sequencing runs produced nearly 100 million sequence reads, which corresponds to over 3 Gb of raw sequence.

Illumina/Solexa Error profile.

At the time of this study, the error rates and error profile (that is, the patterns and systemic nature of sequencing errors) of the Illumina/Solexa sequencing technology had not been exhaustively quantified. Sequencing errors are a primary cause of spurious polymorphism calls and therefore we sought to understand the mechanisms of error in hopes of accounting for them as part of our variation discovery analysis.

Since the N2 Illumina reads were obtained from the same isolate as the reference genome sequence, we expected few true sequence variations between the two. Thus the N2 Illumina reads are an ideal dataset with which to assess true sequence error rates. By carefully aligning (**Methods**) the N2 Illumina reads to the *C. elegans* reference sequence, Derek Barnett and Weichun Huang were able to quantify the overall error rate and to assess the distribution of errors among the reads. Overall, the accuracy of the Illumina/Solexa reads is quite high. As shown in **Figure 5.1,** 57.2% of the reads contained zero mismatches, and 79.9% of the reads had 0 or 1 mismatch. These findings indicate that because nearly half of the sequencing reads have at least one error, one must account for sequence errors when determining the proper genome mapping location for a given read.

Not surprisingly, given the nature of the sequencing chemistry employed, we found that the error rate within the sequencing reads increases as the sequencing cycles proceed. **Figure 5.2** shows the decrease in accuracy depending upon a base's position within the sequence read (results collected by Derek Barnett and Weichun Huang).

Resequenceability of the C. elegans genome with the Illumina/Solexa technology. The short (32 bp) sequence reads used in this study are much more susceptible to non-unique genome mapping than the longer sequence reads produced by capillary sequencing machines. Traditionally, ambiguous capillary read mapping is prevented by aligning the reads to a reference genome where repetitive sequences have by "masked" by software such as *RepeatMasker* (Smit). However, *RepeatMasker* uses previously-known, organism-specific repetitive sequences that are typically much longer than potential repeats at the 32 bp level. Therefore even genome sequences that have been masked by *RepeatMasker* cannot prevent ambiguous read mapping. Proper read mapping is imperative in sequence-based polymorphism discovery projects as incorrect mapping will lead to spurious SNP and INDEL calls that arise merely from paralogy.

In order to assess the error profile of the Illumina/Solexa technology and to uncover sequence variations between the both the resequenced N2 and CB4858 strains relative to the N2 reference genome, we developed a novel method to identify genomic repeats at the 32 bp level. This 'resequenceability' analysis sought to identify regions of the reference genome with a significant potential for ambiguous read alignment. First, we identified all unique 32mers in the reference sequence, but since our error rate analysis (**Figure 5.1**) indicated a drop-off in the error rate beyond 2 errors per read, we defined a putative repeat region as a 32mer that appears in the genome more than once with 0-2 mismatched bases (either substitutions, insertions, or deletions). We called these repeat regions "microrepeats" to distinguish them from the repeat regions marked by the widely used *RepeatMasker* program. Based on this definition, we identified microrepeats with 1 or 2 mismatches using by *BLAT* (**Methods**) and determined that 19.8% of the genome is comprised of perfect and near-perfect microrepeats. *RepeatMasker* masks 14.5% of the bases in the genome. The relationship between *RepeatMasker*-masked bases and microrepeat bases identified by our methods is shown in **Figure 5.3** (results collected by Dr. Chip Stewart). Although there is a significant overlap (11.11% of the genome) between the regions masked by these two methods, 8.7% of the genome that we identify as microrepeats is not masked by *RepeatMasker*. On the other hand, 3.4% of the genome was masked by *RepeatMasker* only, indicating that some fraction of *C. elegans*-specific repeat elements can in principle be uniquely sequenced with 32 bp reads. Taken together, *RepeatMasker* and hash-based micro-repeats cover 23.2% the genome.

SNP and INDEL discovery in the CB4858 strain.

Of the total 37.9 million CB4858 Solexa reads, we were able to align 29.8 (78.6%) to the *C. elegans* reference genome with *Mosaik* (**Table 5.1**). Once aligned, we applied our combined microrepeat plus *RepeatMasker* masking to exclude potential paralogous alignments. We then used *PolyBayes* to identify high quality sequence variations (e.g. **Figure 5.4**) and finalized a set of 45,539 SNPs and 7,353 single base-pair indels. This yields a rate of 1 SNP per 1,629.81 bp. That is, the pair-wise nucleotide diversity (θ) between the CB4858 nd the N2 Bristol strain is 6.136 x 10^{-4}. This agrees with the ~1:1,500 rate posited in a previous description of CB4858 (Denver et al. 2003). The corresponding INDEL rate was 1 per 9894.99 bp. All discovered CB4858 sequence variants have been submitted to Wormbase.

Roughly 1,000 candidate SNPs and INDELs were selected for PCR-based capillary sequence validation. Following sequencing and evaluation, we determined a SNP validation rate of 96.3% (438/455) and an 89.0% conversion rate (438/492) for the candidates identified by *PolyBayes*. We also sequenced 239 of our putative single base INDELs, finding they validated (93.8%) and converted (87.7%) at practically the same rate as SNPs. This INDEL validation rate is much higher than has been achieved with capillary sequencing technologies and is indicative of the fidelity of Solexa sequencing terminator chemistry. Insertions and deletions relative to the reference genome sequence were nearly equally represented (insertions: 2,948 or 47.1%, and deletions: 3,316 or 52.9%). Many of the INDELs were variable numbers of bases in mononucleotide repeats e.g. 5 A's vs. 4 A's. These are traditionally very difficult areas for INDEL detection. Our high validation rate indicates that Solexa reads resolve base numbers in mononucleotide runs very well. **Table 5.2** summarizes the SNPs and INDEL candidates in the CB4858 strain, and illustrates the significant impact of microrepeat masking on accurate SNP discovery to eliminate spurious SNPs and INDELs due to paralogous read mapping.

We estimated false negative rates of our *PolyBayes* SNP calling pipeline by assessing what fraction of additional SNPs found in the capillary validation traces were missed. We ran *PolyPhred* (version 5.0) on these validation traces and found that 26 of 693 SNPs were missed by *PolyBayes* when there was sufficient Solexa read coverage to find the SNP. This equates to a false negative (that is, missed SNP) rate of 4.4%.

In 2000, Koch illustrated in *C. elegans* that, as expected, the non-synonymous substitution rate was much higher in the first and second codon positions than in the third (Koch et al. 2000). As illustrated in **Figure 5.5**, our study confirms these earlier results and provides a detailed, genome-wide estimate of coding polymorphisms in the CB4858 strain. In total, we found 6,255 SNPs positioned within an exon, of which 3,275 putatively introduce an amino acid change. Through our experimental validation, 100 of 119 (84%) non-synonymous mutations were confirmed, indicating that our methods provide an important first step in describing the complete mutational profile in a strain-to-reference paradigm. Furthermore, we evaluated SNP positioning on a chromosome-by-chromosome basis, as shown in **Figure 5.6**. We found that the polymorphism density is much higher on Chromosomes II, III and X. The densities on Chromosomes IV and V suggest a very low variation rate for much of the chromosome, yet a comparatively high mutation rate on the right half of each chromosome.

Sequence differences between the resequenced N2 strain and the reference N2 genome.

We unambiguously aligned 79.4% of nearly 62 million N2 sequencing reads to the reference genome with *Mosaik.* Using these alignments combined with the CB4858 alignments, we scanned for sequence differences in order to uncover possible errors in the reference genome sequence. We ignored sequence differences that solely existed in the N2 sequence reads, as these likely reflect genetic differences in the passaged N2 strain relative to the N2 strain that was sequenced for the *C. elegans* genome. Instead, we focused on sequence differences that were found in both the N2 and CB4858 strains relative to the reference genome. Because they exist in both strains, they are more likely to be true errors in the reference genome. In total, we found 617 such differences. This is indicative of the high quality of the *C. elegans* reference genome as this reflects an overall error rate of approximately 6 errors per Mb of genome sequence.

Discussion

The application of the Illumina/Solexa massively parallel sequencing technology to investigate and characterize genome-wide variation is a compelling paradigm because of the apparent increases in throughput and economy. Yet because this is a nascent technology, it poses significant bioinformatics-based hurdles for proper use and interpretation. In our study, the Illumina/Solexa platform was used to successfully re-sequence the genomes of the nematode *C. elegans* N2 and CB4858 strains. This study required 3 weeks to produce a library and generate

>20X genome coverage. Subsequent analytic efforts provided novel insights into proper methods for reference genome masking and elucidated aspects of Illumina/Solexa sequencing error profile. Our read mapping results indicated that large stretches of the unique genome could be successfully covered with short reads, and that both isolate-specific variants and reference genome errors could be identified. We find that short sequence reads also provide a powerful capability for genome-wide SNP and small INDEL discovery. This is a promising result as capillary sequencing technologies are typically not an accurate substrate for small INDEL discovery. We find that proper masking of "microrepeat" sequences is required to yield high confidence alignments and to eliminate non-paralogous alignments that can falsely indicate sequence variants. Aside from SNP/INDEL discovery, whole genome resequencing also could be utilized following a random mutagenesis approach, to identify and characterize each mutagenized location in the genome. Our analytical approaches provide a valuable baseline toward using Illumina/Solexa technology for resequencing human genomes.

Methods

Illumina/Solexa sequencing.

Elaine Mardis and Vincent Magrini at The Washington University Genome Sequencing Center sequenced genomic DNA from the N2 and CB4858 strains of *C. elegans*. Standard Illumina protocols were used and are described below.

Preparation of Solexa fragment libraries.

Genomic DNA (5 μg) was nebulized for 2 minutes at 45 psi of compressed air, to obtain an average fragment size of 500 bp, then further purified and concentrated with Qiaquick PCR purification spin columns (Qiagen Inc., Valencia CA). Treatment to remove 3′ overhangs and fill in 5′ overhangs resulted in blunt ended genomic fragments. An A residue was added by terminal transferase to the 3′ end and the resulting fragments were ligated with Solexa adapters. Adapter-modified DNA fragments were enriched by an 18 cycle PCR using 50 ng of the ligation reaction and Solexa universal adapter primers. The resulting PCR products were separated by agarose gel electrophoresis and the band between 150-200 bp was excised from the gel. The DNA fragments were extracted from the agarose slice using a Qiaquick Gel Extraction Kit (Qiagen Inc.), and further purified by drop dialysis using a 0.025um/25mm filter (Millipore Inc., Billerica MA) and tissue culture-grade water (Sigma Chemical, St. Louis MO). The DNA fragment library was quantitated, then diluted to a 10 nM working stock, appropriate for cluster generation.

Sequencing cluster generation.

Adapter-ligated fragments (2 nM)were denatured in 0.1N NaOH for 5 minutes, then were further diluted to a final 9 pM concentration in 1 ml of pre-chilled hybridization buffer, and introduced onto the Solexa flow cell using the Cluster Station, an automated device supplied by Solexa. On this apparatus, the oligo-derived flow cell surfaces hybridize to library fragments by adapter-to-oligo pairing. “Clusters”, representing discrete populations of unique single-stranded library fragments amplified in situ, are generated by isothermal amplification using a proprietary process. In practice, each cluster produces a single Solexa read. Our experiments aimed for an average cluster density of 30,000 (+/- 5,000) clusters per flow cell lane.

Illumina/Solexa sequencing process.

The Illumina/Solexa 1G Analyzer provides up to 32 sequential flows of fluorescently labeled, 3′-OH blocked nucleotides and polymerase to the surface of the flow cell. After each base incorporation step, the flow cell surface is washed to remove reactants and then imaged by microscope objective. Our experiments collected 200 tiled images (“tiles”) per flow cell lane, each containing on average 30,000 clusters. Solexa single end reads were generated for N2 Bristol and CB4858 strains using a 30 base read length for the titration run (used to determine the correct input library DNA amount), and a 32 base read length for standard runs

Sequence accuracy of Solexa N2 Bristol reads.

In order to isolate sequencing errors from simple alignment errors, we used a version of the Smith-Waterman-based global alignment algorithm that reports all optimal and sub-optimal alignments above a pre-specified alignment score. Although time-intensive, this algorithm identifies all alignable positions in the *C. elegans* genome for every read. Here, we generated three random sample sets of 20,000 Solexa N2 Bristol reads, and aligned each read set to the unmasked reference genome, allowing up to 4 mismatches (substitution, insertion or deletion). We kept only reads for further consideration of accuracy that aligned at a single locus in the genome.

Mosaik alignment of Illumina/Solexa reads.

Accurate mapping of short resequencing reads to their exact place of origin necessitates the identification and masking of microrepeated sequences that correspond to the short read lengths, prior to read mapping. Here, we identified both perfect microrepeats and microrepeats with up to two mismatches (substitutions, deletions or insertions) in order to encompass the possibility of sequencing errors in the reads (due either to nucleotide mis-incorporation or base calling error) or of polymorphism in the genomes being compared. Our approach used two fundamental methods: (1) hash-based, and (2) sequence alignment-based. Our hash-based method enumerated every 32-mer in the *C. elegans* reference genome sequence, and recorded its map location. If the same 32-mer occurred in multiple locations (either strand), it was marked as a microrepeat. Near-perfect microrepeats were identified by asking whether each specific 32-mer, or any of the other 32-mers obtained by introducing "mutations" up to a pre-specified number of mismatches, appeared at any other genomic location. The sequence alignment-based method used the BLAT algorithm with the following parameters: -stepSize=8 -tileSize=16 -minMatch=1 -minScore=28 -oneOff=1 to enumerate every 32-mer in the reference genome, and then to search the rest of the genome for a perfect or near-perfect match, up to a pre-specified number of mismatches. Custom scripts identified hits that had up to 2 mismatches (any combination of substitutions, insertions, or deletions), and combined the results of our hash-based and sequence alignment-based methods to produce a microrepeat-masked reference genome.

We next aligned the Illumina/Solexa reads to the microrepeat masked *C. elegans* reference genome with *Mosaik*. In this analysis, we allowed a maximum of two mismatches (including substitutions and indels), a decision motivated by the fact that over 79% of the Solexa reads had either one or zero error. Allowing a single polymorphic difference in such reads would bring the maximum number of mismatches to two. The assembler algorithm first registers every gap induced by any of the aligned reads on the reference sequence, then introduces alignment gaps into all aligned reads to preserve the positions of all pair-wise aligned bases in every pair-wise read alignment. The resulting multiple alignments are then reported either in ACE (Gordon et al. 1998a) or in binary formats used by other downstream analysis software.

SNP and INDEL discovery in strain CB4858.

Starting with the multiple read alignments produced by the *Mosaik* aligner, we performed an analysis with a version of *PolyBayes* that was completely re-engineered to enable efficient analyses of millions of short read sequences at once. The program evaluates each aligned base and its base quality value at each position, to indicate putative SNPs and small (1-3 bp) putative INDELs, and their corresponding SNP probability values. Base quality values are converted to base probabilities corresponding to every one of the four possible nucleotides (and to the probability that the nucleotide in question is, in fact, an insertion error in the sequence). Using a Bayesian formulation, a SNP (or INDEL, as appropriate) probability value is calculated as the likelihood that multiple different alleles are present between the reference genome sequence and the reads aligned at that position. If the SNP probability value exceeds a pre-specified threshold, the SNP candidate is reported in the output. For the collection of bases contributed by such reads, a single "consensus" base call and its base quality value are computed. The corresponding base probabilities are then used in the Bayesian SNP probability value calculation. In this study, we used a PSNP cutoff value of 0.7.

Validation of suspected sequence differences.

A subset of candidate insertion, deletion, and polymorphic sites from the above analyses were submitted for orthologous validation using PCR-based sequencing and variant analysis. For each type of variant, we validated 50% of the candidates identified in coding and 50% in non-coding sequences. We designed primers with 300 bp of sequence both to the left and to the right of the target site, using Primer3 (http://primer3.sourceforge.net/). A 5′ universal M13 forward or reverse sequence was added to each primer pair to allow processing of the resulting PCR products in our high-throughput sequencing pipeline. Ginger Fewell at The Washington University Genome Sequencing Center performed the validation experiments for polymorphism candidates.

Evaluation of validation reads.

We aligned validation reads to the reference sequence using *PolyPhred*, determining by manual inspection whether a Solexa variant was confirmed in the 3730 trace. Both a validation rate (defined as the number of confirmed variants divided by the number of successful sequencing reactions) and a conversion rate (defined as the number of confirmed variants divided by the number of variants submitted for validation) were calculated.

Evaluating variants for exonic disruption.

We investigated SNP candidates in CB4858 that introduced amino acid changes relative to Bristol. Conceivably, such differences might suggest subtle chemosensory or other adaptations specific to CB4858. Using the Wormbase gene annotations, we characterized all SNPs that lie within exons by their codon position and whether the variant caused a synonymous or non-synonymous amino acid change.

Chapter 5 Tables

Strain	Number of Illumina/Solexa sequencing runs	Number of sequence reads (in millions)	Number of sequence reads aligned (in millions)	Fraction aligned
N2	3.5	61.8	49.1	79.4%
CB4858	1.5	37.9	29.8	78.6%

Table 5.1. Illumina/Solexa sequencing statistics.

Variation	Mask type applied	No. Submitted to validation	No. Assay successful	No. candidate confirmed	Validation rate (%)
SNP	Known repeats	598	582	482	86.5
SNP	Exact microrepeats	579	559	475	91.7
SNP	Near-exact microrepeats	492	482	438	96.3
Indel	Known repeats	239	228	202	91
Indel	Exact microrepeats	232	223	201	92.6
Indel	Near-exact microrepeats	220	213	193	93.8

Table 5.2. Validation rates for *PolyBayes* SNP and single base INDEL candidates. The validation rates for both SNPs and INDELs increased as more specific masking methods were employed in order to detect microrepeats.

Chapter 5 Figures

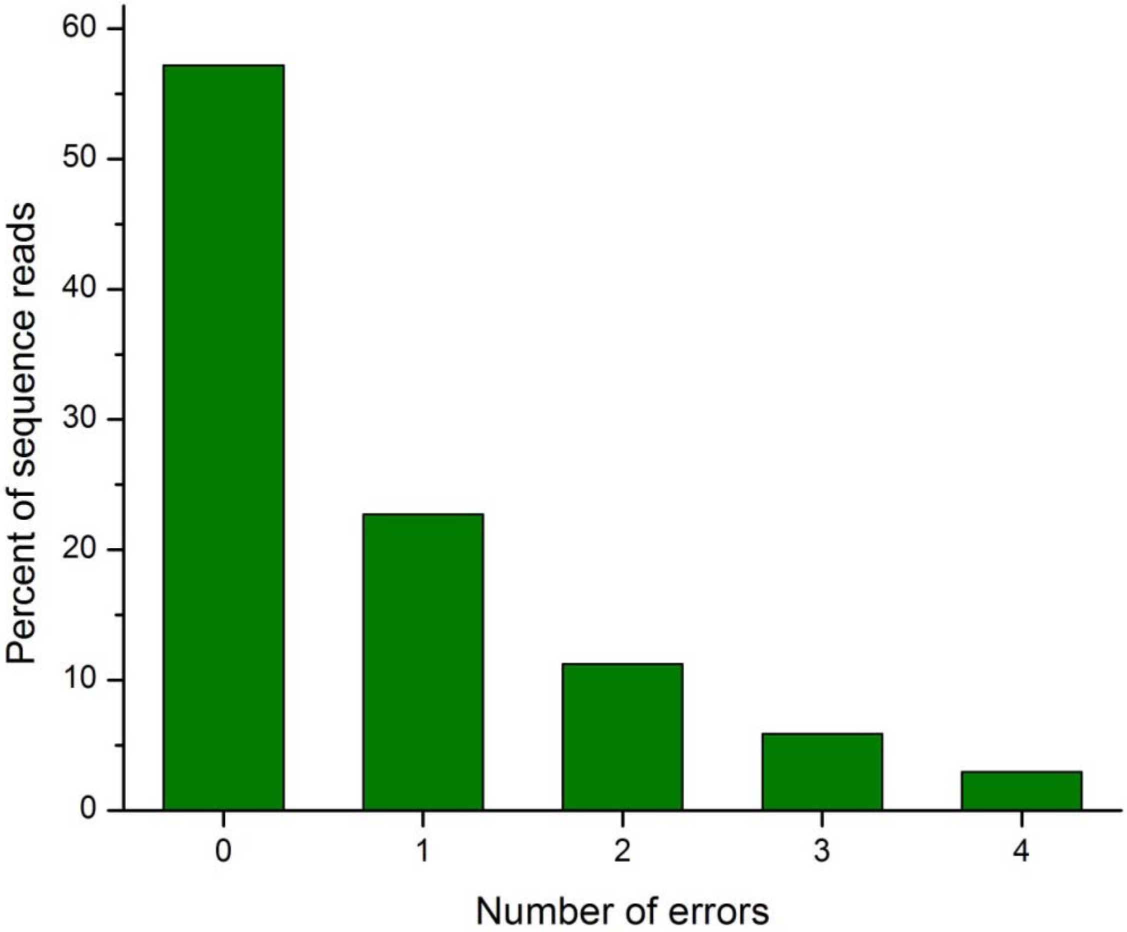


Figure 5.1. Distribution of errors among Illumina/Solexa reads. The fraction of reads (y-axis) with zero or more errors (x-axis) is shown.

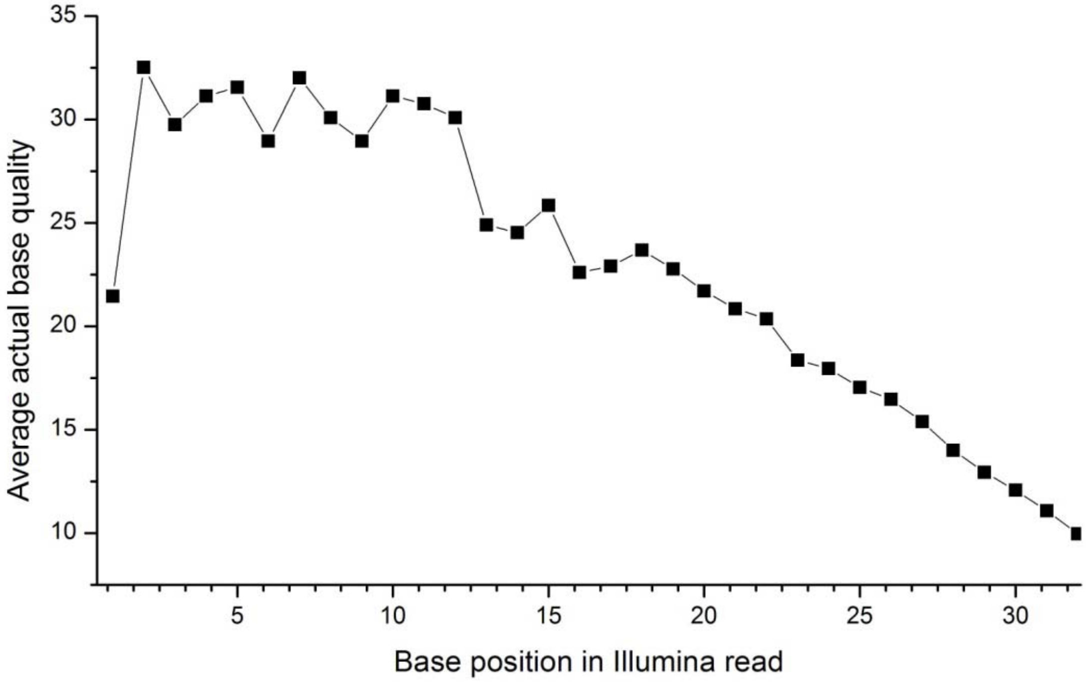


Figure 5.2. The position-dependent accuracy of Illumina/Solexa reads. For all 2 base pair Illumina reads, we examined the actual accuracy of all bases that were assigned a quality score of 30 (i.e. estimated to have a 1/1000 error rate). The plot shows the actual accuracy (y-axis) of these called bases as a function of the position in the Illumina read. The first 10-15 bases are two orders more accurate than the last 10 bases. This indicates that the native quality values assigned by the Illumina software show be re-calibrated to reflect their actual accuracy.

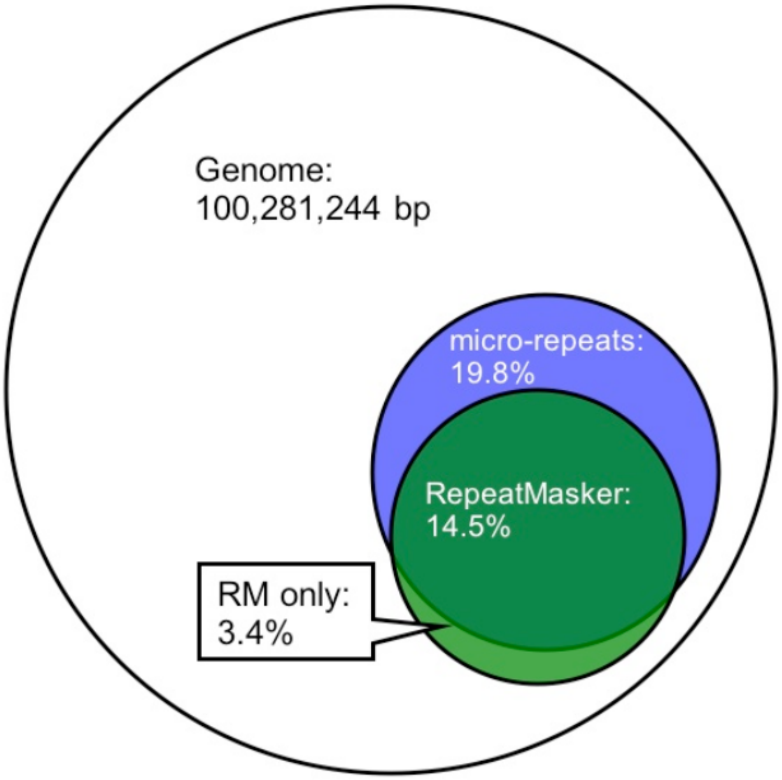
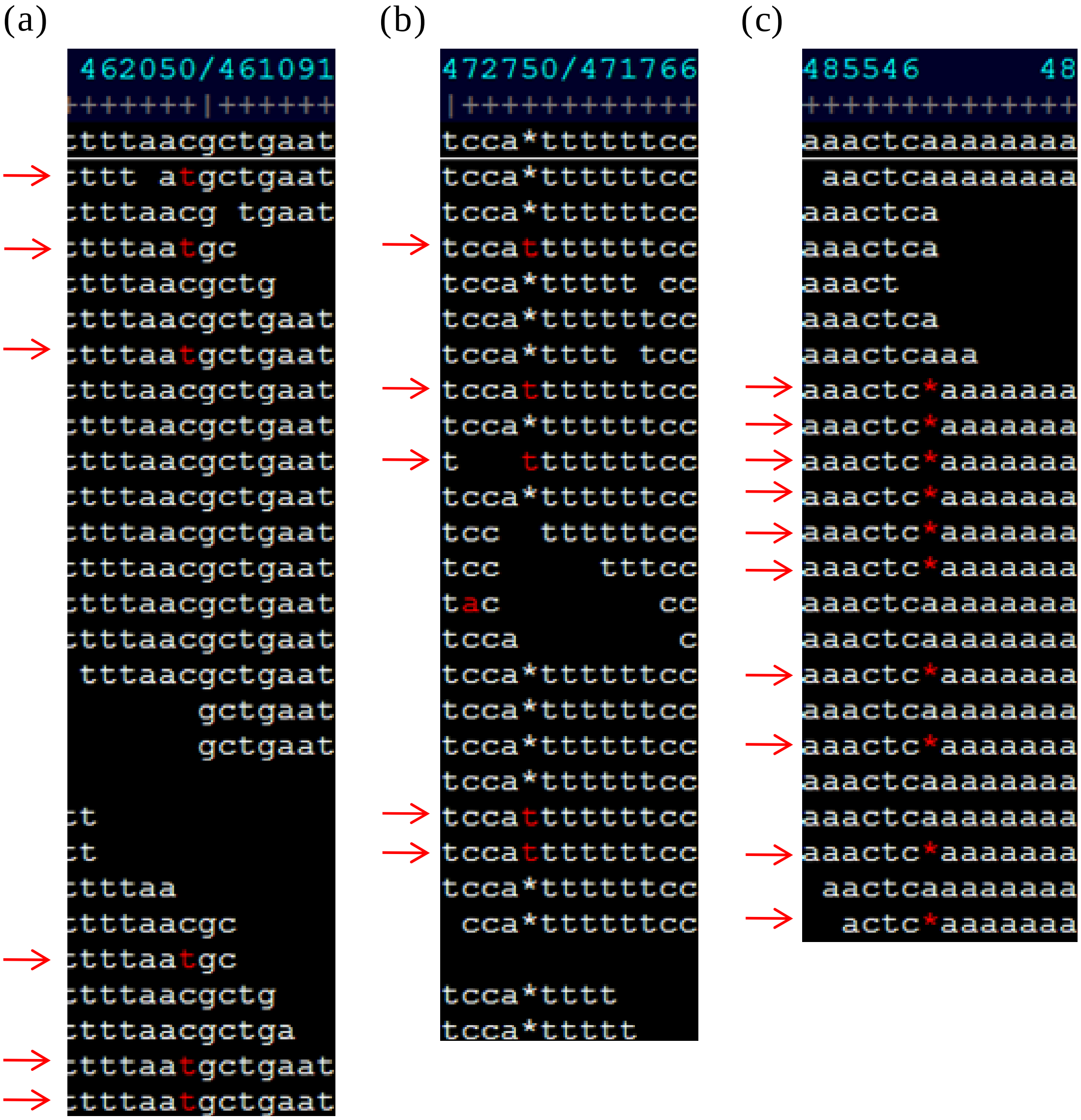


Figure 5.3. Concordance of microrepeats with the repeats identified by *Repeatmasker*. The repetitive fractions of the *C. elegans* genome (build ws170) are shown as identified by RepeatMasker (green) and by our custom method for identifying micro-repeats (blue). As expected, the micro-repeat method identifies short repetitive sequences that are missed by RepeatMasker.



(a)

```
462050/461091
+++++++|++++++
ctttaacgctgaat
cttt atgctgaat
ctttaacg tgaat
ctttaatgc
ctttaacgctg
ctttaacgctgaat
ctttaatgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
ctttaacgctgaat
 tttaacgctgaat
       gctgaat
       gctgaat

ct
ct
ctttaa
ctttaacgc
ctttaatgc
ctttaacgctg
ctttaacgctga
ctttaatgctgaat
ctttaatgctgaat
```

(b)

```
472750/471766
|++++++++++++
tcca*ttttttcc
tcca*ttttttcc
tcca*ttttttcc
tccatttttttcc
tcca*ttttt cc
tcca*ttttttcc
tcca*tttt tcc
tccatttttttcc
tcca*ttttttcc
t   tttttttcc
tcca*ttttttcc
tcc  ttttttcc
tcc     tttcc
tac        cc
tcca        c
tcca*ttttttcc
tcca*ttttttcc
tcca*ttttttcc
tcca*ttttttcc
tccatttttttcc
tccatttttttcc
tcca*ttttttcc
 cca*ttttttcc

tcca*tttt
tcca*ttttt
```

(c)

```
485546      48
++++++++++++++
aaactcaaaaaaaa
 aactcaaaaaaaa
aaactca
aaactca
aaact
aaactca
aaactcaaa
aaactc*aaaaaaa
aaactc*aaaaaaa
aaactc*aaaaaaa
aaactc*aaaaaaa
aaactc*aaaaaaa
aaactc*aaaaaaa
aaactcaaaaaaaa
aaactcaaaaaaaa
aaactc*aaaaaaa
aaactcaaaaaaaa
aaactc*aaaaaaa
aaactcaaaaaaaa
aaactcaaaaaaaa
aaactc*aaaaaaa
 aactcaaaaaaaa
  actc*aaaaaaa
```

Figure 5.4. Example SNP and INDEL candidates discovered by *PolyBayes*. We show an example of a SNP (a), insertion (b) and a deletion (c) in the Pasadena strain, relative to the N2 reference strain. Red arrows indicate the Pasadena alleles. Unmarked alleles are from the N2 reference strain.

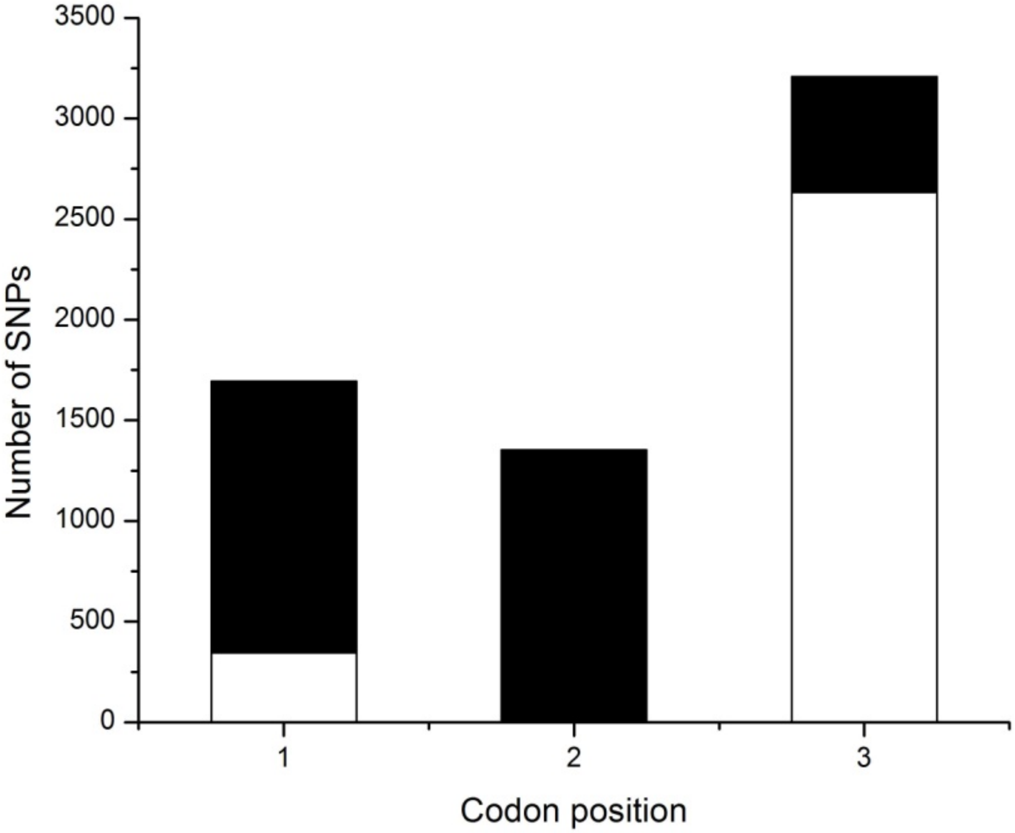


Figure 5.5. Distribution of SNPs according to codon position. The fraction of synonymous (white) and non-synonymous polymorphisms are shown according to their codon positions.

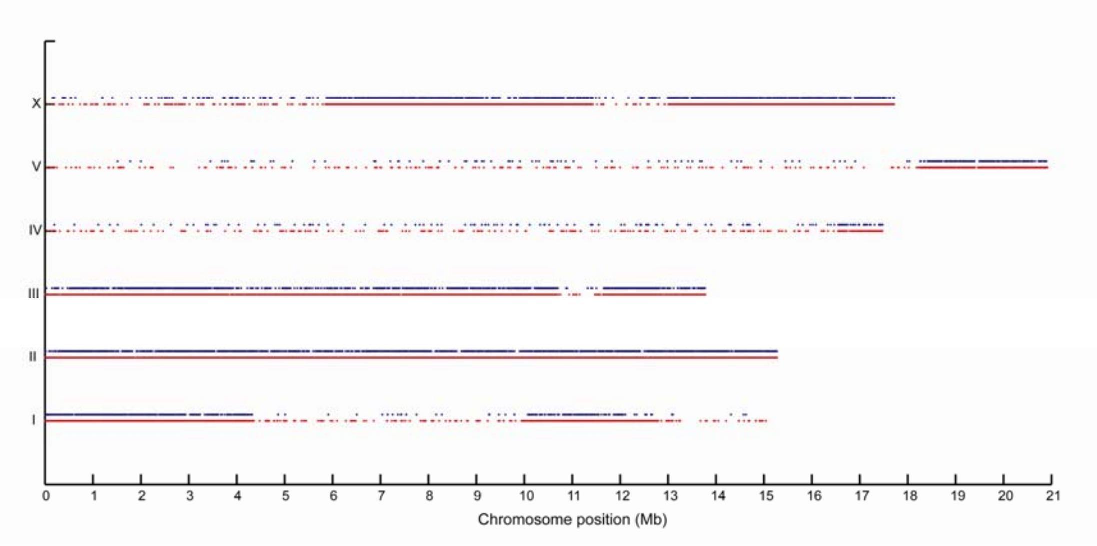


Figure 5.6. Polymorphism density in the CB4858 genome. The density of SNPs (red dots) and INDELs (blue dots) is shown for each of the six *C. elegans* chromosomes.

Efficiency is intelligent laziness.

-Anonymous

6. Rapid whole-genome mutational profiling of Pichia stipitis using next-generation sequencing technologies

Abstract

Adaptive evolution, forward genetic mutational studies and phenotypic screening are powerful tools for mutant strain development that complement molecular genetic approaches in metabolic engineering. However, mutations generated in the process cannot be easily identified with traditional genetic tools. We show that using new high-throughput, massively parallel sequencing technologies one can completely and accurately characterize the mutant genome relative to a previously sequenced parental strain. We studied a mutant strain of *Pichia stipitis*, a yeast capable of converting xylose to ethanol. This unusually efficient mutant strain was developed through repeated rounds of chemical mutagenesis, strain selection, transformation and genetic manipulation over a period of seven years. We resequenced this strain on three different sequencing platforms. Surprisingly, this study revealed

fewer than a dozen mutations in open reading frames. Only one lesion could have been predicted prior to the resequencing study. All three platforms we used were able to identify each single nucleotide mutation given at least 10-15 fold nominal sequence coverage. This represents a rapid and efficient alternative to traditional mutation mapping techniques used in reverse genetic screens.

Introduction

Pichia stipitis (Pignal) is a haploid yeast related to endosymbionts of beetles that degrade rotting wood (Suh et al. 2003). It is an important organism for bioenergy production from lignocellulosic materials because of its high capacity to ferment xylose and cellobiose to ethanol. Jeffries *et al* previously sequenced the reference strain, *Pichia stipitis* CBS 6054, resulting in a completely characterized 15.4 Mb genome of eight chromosomes (Jeffries et al. 2007). Prior to sequencing, this strain had been subjected to chemical mutagenesis, phenotypic selection, genetic engineering and adaptive evolution to improve ethanol production. Mutation and selection resulted in small advances attributable in part to carbon catabolite derepression (**Figure 6.1**). Disruption of CYC1 to create strain Shi21 increased the specific ethanol production rate by 50% and the ethanol yield by 10%, however the nature of mutational events preceding this phenotype were unknown. Traditional methods for identifying mutations are labor and time intensive, so we wanted to test the ability of next generation sequencing technologies to determine the differences in this improved strain's entire genome relative to the reference strain. We generated high-coverage, whole genome data sets using three next generation sequencing platforms (454/Roche; Solexa/Illumina; and SOLiD/ABI). This allowed us to determine the effect of coverage on the accuracy of mutation detection, the lowest levels of coverage required for effective and exhaustive detection of the mutations, and to compare the efficiency of the three platforms for this application.

Results

Sequencing and alignment strategy.

Genomic DNA from *P. stipitis* (Shi21) was sequenced using the Roche, Illumina, and Applied Biosystems advanced sequencing platforms. We processed the sequence reads from each technology with the manufacturer-supplied base calling software. We additionally re-called the 454 pyrosequences with *Pyrobayes* because, as described in Chapter 4, it produces a lower number of substitution errors and more accurate base quality values than the native base-calling program (**Methods**) (Quinlan et al. 2008). We first identified all repetitive elements within the *P. stipitis* genome that would interfere with unique read alignments, including short genomic repeats as well as NUMTs (Richly and Leister 2004), sequences of mitochondrial origin inserted into the nuclear genome (**Methods**, **Table 6.1**). Due to the nature of the unpaired short reads produced by these methods, this repeat masking resulted in 93.2%-94.7% of the reference genome accessible to unique placement of the reads (**Methods**). The total number of aligned reads passing alignment quality filters and the corresponding aligned read coverage are shown in **Table 6.2**. Alignment of reads from each technology to the repeat-masked reference sequence, resulted in 11-175X coverage of the genome depending on the type of platform and number of runs (**Table 6.2**).

Mutation discovery.

Multiple read alignments from the 454 and Illumina platforms were screened for mutations using a new version of the *PolyBayes* SNP discovery program (**Methods**). Color-space alignments of the SOLiD data were similarly screened using software supplied by Applied Biosystems. The 17 candidate mutations discovered among the three sequencing technologies were re-sequenced in CBS 6054 and in each of the four derived strains with a capillary sequencing machine and were all confirmed (**Table 6.3**). However, three of the changes appear to be the consequence of sequencing errors in the original reference sequence, as the alternate base is not only present in the validation traces from all sequenced mutants but also in the parent strain. This is to be expected in a typical sequencing project, and 3 errors in over 15 Mb of finished sequence is, in fact, a very low error rate. Given that the mutations were discovered in very deep datasets and independently confirmed by four different sequencing methods, it is unlikely that we missed any additional mutations in the *Pichia* strains. We therefore believe that the remaining 14 mutations comprise the complete set of single nucleotide polymorphisms between the mutant and the parent (i.e. reference) *Pichia* strains. Since the *Pichia* genome is haploid during vegetative growth, all mutations are expected to be homozygous. An apparent heterozygous change at position 358,358 on chromosome 8 is a result of the intentional gene disruption of CYC1 with a URA3 selection cassette, which resulted in an URA3 duplication. This variation represents a paralogous difference in a duplicated gene and thus cannot be considered a true point mutation. Given the nature of the short unpaired reads produced by these technologies, it was not possible to positively identify large insertion/deletion events. No small INDEL polymorphisms were found, and this is not surprising considering that the alkylating agents (**Methods**) used in mutagenesis principally induce base substitutions.

Comparing the accuracy of various sequencing technologies.

A primary focus of this study was to evaluate the utility of next-generation sequencing technologies for mutational profiling. Therefore, we compared the capabilities of the three platforms for the identification of the 14 confirmed point mutations in the *Pichia* mutant. I evaluated the accuracy of the Illumina and 454 technologies while Heather Peckham from Applied Biosystems evualted the performance of the SOLiD sequencing platform. Each of the three sequencing technologies was able to correctly identify all 14 variations with essentially no false positives when all available reads generated on the platform were used (**Table 6.2**, **Figure 6.2**). We observed a single false positive prediction in the complete 454 FLX data that produced lower overall coverage than the other platforms, and was most likely the result of a PCR error during sequence library construction (data not shown). The complete Illumina and AB alignments afforded perfect accuracy: all 14 mutations were found and no false positive predictions were made. The accuracy we observed is encouraging given that low false discovery (i.e. that is, the fraction of erroneously identified mutations) and false negative (i.e. the fraction of true mutations that were missed) rates are critical considerations for the application of these technologies to rapid forward genetic mutational profiles. These results show that all three technologies are suitable for highly accurate mutation screening.

What is the minimum sequence required for reasonable accuracy?

Another important consideration is the depth of sequence coverage required to achieve the sensitivity and specificity we observed. To determine how the error rate changes as fewer reads are used, we assembled subsets of reads of varying size from each of the three full datasets and subjected the resulting lower-coverage assemblies to our mutation discovery analysis. As shown in **Table 6.2**, if we limit the combined missed mutation (false negative) and erroneously called mutation (false positive) error rate to 50%, we were able to reduce to 1.5 454 FLX machine runs (8.15-fold aligned read coverage), yielding 6 FP and 1 FN errors. A single lane of Illumina reads (6.32-fold aligned read coverage), resulted in 2 FP and 2 FN errors. Similarly, 6-fold coverage of AB SOLiD reads yielded 0 FP and 6 FN errors. These results indicate that genome coverage of 10-15X is optimal and should be targeted given the constraints of plate configurations and run conditions on the different platforms.

Discussion

The distribution of mutations in open reading frames as opposed to non-coding regions (78%) was slightly higher than the average gene density (56%) (Jeffries et al. 2007). In the absence of selection, about two-thirds of the base changes should have resulted in silent mutations at the amino acid level, due to redundancy in the genetic code. Surprisingly, none of the induced or spontaneous mutations were silent. All mutations retained in open reading frames resulted in amino acid changes (**Table 6.3**). Selection identified strains growing faster under restrictive conditions. Selection of FPL-061 involved repeated mutagenesis and selection on poorly used carbon sources. Selection of FPL-DX26 involved serial selection of strains showing more rapid growth on D-xylose in the presence of 2-deoxyglucose, which normally suppresses xylose utilization in wild-type cells. In the case of the cyc1 disruption, the mutant grows substantially slower than the parental strain, and it is possible that a number of compensating mutations conferring faster growth might have arisen. Relatively little is known about the physiological effects of the various genes or mutational events. Only URA3 and ALD7 have been characterized (**Table 6.3**). The former is widely used as an auxotrophic selectable marker (Boeke et al. 1984). The latter is an NADP-specific secondary alcohol dehydrogenase. One spontaneous mutational event apparently occurred following isolation of the Shi21 cyc1D in MDM34, which is involved in determining the shape of the mitochondrial outer membrane. Mutational events obtained in the screening and selection process were possibly lost if the original forward mutation conferred disadvantages to the cell once the selective pressure was released. This could have been true of the selection for 2-deoxyglucose resistance or resistance to respiration inhibitors, however the sequencing data do not show signs of this occurring. Further characterization of the identified mutational events through physiological and genetic studies will be necessary to determine how they affect cell phenotype.

Overall, our results demonstrate that the new sequencing technologies tested are well suited for mutational analysis of novel yeast strains derived from multistep mutagenesis procedures. The approach is expected to be equally suitable for the analysis of bacterial, fungal and larger organisms derived by directed evolution and natural variation, which could help accelerate the development of novel organisms for bioenergy and biotechnology applications.

Methods

Derivation of the mutagenized SHI-21 strain.

In an attempt to produce a *P. stipitis* strain that was more efficient at ethanol production, a series of directed and random mutagenic steps were undertaken. Starting from the parental wild-type strain, *P. stipitis* CBS 6054, four generations of mutants were created. The first, *P. stipitis* FPL-061, was derived from CBS 6054 by mutagenesis with N-methyl N′-nitro-N-nitrosoguanidine (MNNG) followed by selection for rapid growth on L-xylose or D-arabinose in the presence of the respiration inhibitors salicylhydroxamic acid and antimycin A. The second, *P. stipitis* FPL-DX26 (NRRL Y-21304) was derived from FPL-061 by mutagenesis with ethyl methanesulfonate (EMS) and selection for growth on D-xylose in the presence of 1.0 g/l 2′-deoxyglucose6. The third, *P. stipitis* FPL-UC7 (ura3-3), a spontaneous URA3 mutant was derived from FPL-DX26 by selection for resistance to 5′-fluoroorotic acid7. Finally, targeted disruption of CYC1 with URA3 then created *P. stipitis* SHI21 (cyc1::URA3)(NRRL Y-21971). Both alkylating agents used (MNNG and EMS) act principally on guanine resulting in mismatch mutations (Shi et al. 1999). Following isolation, each strain was suspended in 15% glycerol and stored at -80°C or lyophilized until shortly before genomic analysis.

Sequencing.

Genomic DNA from *P. stipitis* SHI21 was sequenced on the Illumina platform by Paul Richardson's group at The Joint Genome Institute. Doug Smith at Agencourt sequenced this strain on the 454 technology and Heather Peckham and Joel Malek sequenced SHI21 on the AB SOLiD platform. The specifics of the 454 and SOLiD protocols are described below.

Chromosomal DNA from *P. stipitis* SHI21 was prepared, sheared to the recommended size range and ligated to adapters according to the manufacturer's protocols for each of the sequencing platforms. The DNA fragments were then clonally amplified onto microbeads (454 and SOLiD) or onto the surface of a flow cell (Illumina), sequenced, and the resulting data were processed according to the manufacturer's protocols. For confirmation sequencing, PCR products were generated from genomic DNA of each strain using M13-tailed primer pairs, the products were sequenced on ABI3730xl instruments, variants were identified using *PolyPhred* and confirmed using *Consed*.

Repeat identification and masking in the Pichia genome.

Prior to sequence alignment, we identified and masked short repeats in the *Pichia stipitis* reference sequence in order to prevent spurious alignments. Given the dramatically different read lengths produced by the Illumina (32 base pair) and 454 Life Sciences (avg. of 225 base pair) technologies, we generated two repeat-masked reference sequences based on the expected read lengths of each technology.

Micro-repeat masking was performed using the *Mosaik* resequenceability analysis tool (*Mosaik-RA*). *Mosaik*-RA extracts all possible k-mers from a target genome and then aligns them to the genome. All k-mers that align to multiple regions in the genome within a specified edit distance are masked as repetitive regions. The chosen *Mosaik*-RA parameters guarantee that all repetitive regions within the edit distance are found. Masked genomes were generated for the fixed-length Illumina datasets (32 bp reads) and for the variable-length 454 FLX reads, where the lower end of the 95 % confidence interval of read lengths was used. The 454 FLX masked genome was generated using 134 bp k-mers (μ=224 bp, σ=44.7 bp). 6.8% and 5.3% of the *Pichia* genome was masked based on the 32 bp Illumina (allowing an edit distance of 2 or less) and ~225 bp 454 FLX reads (allowing an edit distance of 5 or less), respectively.

BLAT (Kent 2002) and BLAST (Altschul et al. 1990; Altschul et al. 1997) was used to identify NUMTs in the *Pichia stipitis* reference genome. Using the *Pichia stipitis* mitochondrial genome as the query all BLAST high-scoring segment pairs with an expectation value lower than 1e-4 were recorded. All BLAT hits with a blastz score over 2200 were recorded. In total, six genomic regions were masked as NUMT candidates and were added to the Illumina and 454 masked genomes described above.

Illumina and 454 sequence alignment.

We used our general reference sequence-guided alignment and assembly tool, *Mosaik*, to process the Illumina and 454 datasets. *Mosaik* uses a hashing scheme to seed full Smith-Waterman gapped alignments against the concatenated *Pichia stipitis* genome. The resulting pairwise alignments are then consolidated into a multiple sequence alignment (assembly) and saved as an ACE assembly file. These assemblies can be viewed by programs such as *Consed* (Gordon et al. 1998b) or *EagleView* (Weichun Huang *et* al, manuscript in preparation). To correct for 454 INDEL alignment errors, the Smith-Waterman scoring algorithm has been augmented to use an alternate gap open penalty when a homopolymer region is detected. For both the Illumina and the 454 reads, we required that at least 95% of each read align to the reference sequence. In order to ensure that we only aligned high quality reads from each technology, we also required that the reads from each technology had few sequence differences (i.e. mismatches, insertions or deletions) relative to the reference genome sequence. Thus, we allowed Illumina reads to have at most one sequence difference. Since the 454 reads were much longer than the Illumina reads, we allowed a maximum of two mismatches for each 454 read.

SOLiD sequence alignment.

The AB SOLiD alignment tool translates the reference sequence to 2-base encoding and aligns the reads in color space. The program guarantees finding all alignments between a read and the reference sequence with up to M mismatches for a user specified parameter M. The alignment tool uses multiple spaced seeds (discontinuous word patterns) to achieve a rapid running time. Position specific matrices are implemented in the aligning stage to allow flexibility so that quality values, masking of positions in the reads and user specified reference background SNP rates may be input.

Illumina and 454 mutation discovery.

We scanned the reference-guided 454 and Illumina sequence alignments produced by the *Mosaik* program using a new version of the *PolyBayes* SNP discovery program completely re-engineered for the efficient analysis of millions of next-generation sequence reads. This program sequentially evaluates aligned reads at every position of the reference genome sequence by considering the aligned base and the corresponding base quality value contributed by each aligned read. Given that *Pichia stipitis* has a haploid genome mutations were expected only between the mutant and the parent strain. In other words, every aligned read from the mutant genome is expected to carry an identical allele which (at mutant positions) may differ from the reference allele at that position. The Bayesian mathematical formulation implemented in the *PolyBayes* program is capable of dealing with situations where there is disagreement among the aligned reads from the mutant strain. Based on the aligned reads a "strain consensus" base is determined, and an associated "consensus" base quality value is calculated. The strain consensus allele and corresponding base quality value determined for the mutant strain is used in the comparison to the parent strain. *PolyBayes* then calculates the probability that apparent sequence differences between the mutant and the parent strain represent true mutations as opposed to sequencing errors. We reported every genome position where the probability of such a polymorphism event was above a pre-specified threshold (in this study 0.5 i.e. a 50% likelihood). We used identical parameters for SNP discovery among the Illumina and the 454 alignments. In each case, we assigned a quality value of 40 (i.e. an assumed 1 in 10,000 error rate) to each base in the reference genome sequence.

SOLiD mutation discovery.

All SOLiD beads are basecalled and evaluated. Currently there is no keypass filter or intensity filter to remove empty beads or beads with 2 templates. As a result this presents an artificially deflated percentage of beads matching the genome. Consensus calling on the AB 2-base encoded data was performed with AB developed software. 2-Base encoding is uniquely enabled by the ligation-based sequencing protocol used in the SOLiD system (a massively parallel sequencing technology based on ligation of oligonucleotides). Sequencing is carried out via sequential rounds of ligation with high fidelity and high read quality. In this system there are 16 dinucleotide combinations with 4 fluorescent dyes, each dye corresponding to a probe pool of 4 dinucleotides per pool. Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions. The double interrogation of each base causes a SNP to result in a two-color change while a measurement error results in a single color change. In addition, only one-third of all possible two-color combinations are considered valid and result in a base change. Single nucleotide polymorphisms were identified by a consensus of valid adjacent 2-base encoded mismatches. The confidence of each base call was determined by the position in the read as well as the 6-mer base space context in which the base call occurred and this confidence was used to weight the contribution of each set of adjacent base calls to the consensus call.

Chapter 6 Tables

Chromosome	Begin Pos.	End Pos.
3	1458286	1458416
5	702694	702738
6	1034827	1034863
7	31488	31530
7	539614	539641
8	196567	196664

Table 6.1. Locations of identified NUMT repeats in the Pichia stipitis genome. The beginning and ending chromosome coordinates for each identified NUMT repeat are reported.

Sequencing Technology	Total number of reads	Total sequence (bp, in millions)	Average sequence coverage from aligned reads	False positive (spurious) mutations	False negative (missed) mutations
454 FLX (2 runs)	887,123	199.35	10.78X	1	0
454 FLX (1.5 runs)	669,783	150.64	8.15X	6	1
454 FLX (1 run)	459,563	103.38	5.62X	17	1
Illumina (7 lanes)	25,818,266	826.18	44.24X	0	0
Illumina (3 lanes)	11,281,705	361.01	19.40X	0	0
Illumina (2 lanes)	7,548,407	241.55	13.00X	2	0
Illumina (1 lane)	3,674,253	117.58	6.32X	2	2
AB (2 flow cells)	228,191,758	7,986.71	175.09X	0	0
AB (30x)	39,111,512**	1,368.90	30.01X	0	0
AB (20x)	26,065,653**	912.30	20.00X	0	0
AB (10x)	13,045,859**	456.61	10.01X	0	0
AB (8x)	10,426,261**	364.92	8.00X	0	4
AB (6x)	7,819,696**	273.69	6.00X	0	5

Table 6.2. Sequencing and mutation discovery statistics. The overall and aligned sequence throughput is shown for each sequencing technology used in the study. We also report the number of spurious and missed mutations observed from each experiment. * % matching is a result of the Poisson nature of emulsion PCR. No filtering was performed to remove empty beads, beads with two templates and beads with dim templates. ** Estimated number of reads based on *in silico* degradation of coverage.

Chrom.	Location	Nucleotide change*	Amino acid change	Functional description of mutation
2	1,339,463	T>C	V>A	Error in reference sequence
2	2,598,869	C>A	-	Error in reference sequence
3	1,769,576	C>T	G>S	Error in reference sequence
1	1,143,120	C>T	G>S	*YHN8* (predicted GPCR)
2	746,465	C>T	D>N	*IFI3* (hypothetical protein; ID 29635)
2	1,102,664	G>T	-	Upstream of *RAD15*
3	104,338	T>G	-	Non-coding interval
4	1,499,156	T>A	K>N	*VSP36* (vaculoar sorting protein)
7	930,181	A>T	W>R	*FBX1* (Leucine rich repeat protein, contains F-box)
8	36,439	A>G	D>G	*POT11* (3-ketoacyl-CoA thiolase B)
1	839,170	C>T	V>I	*SEC31* (component of the COPII coat of ER-golgi vesicles)
2	617,666	G>A	S>F	*SLX8* (Zn finger RING domain protein; ID 54919)
1	670,317	G>A	R>K	*ALD7* (aldehyde dehydrogenase)
8	358,358	T>A	D>V	*URA3* (orotidine-5'-phosphate decarboxylase)
1	947,086	C>G	L>V	*MDM34* (mito. outer membrane protein involved in mitochondrial shape
3	885,477	G>C	-	Intergenic region between *LEU3* and *YXE1*
6	1,088,427	G>C	-	Upstream of *TSC11* (TOR binding protein; ID 84674)

Table 6.3. Summary of discovered point mutations relative to the *Pichia* reference genome.*Color coding indicates which strain each mutation first appeared in relative to the parent, CBS-6054. Orange: FPL-061 (rapid growth on L-xylose in the presence of the respiration inhibitors); Yellow: FPL-DX26 (2-deoxyglucose resistance); Green: FPL-UC7 (FOA resistance); Blue: Shi21 (CYC1:*ura3* targeted gene disruption).

Chapter 6 Figures

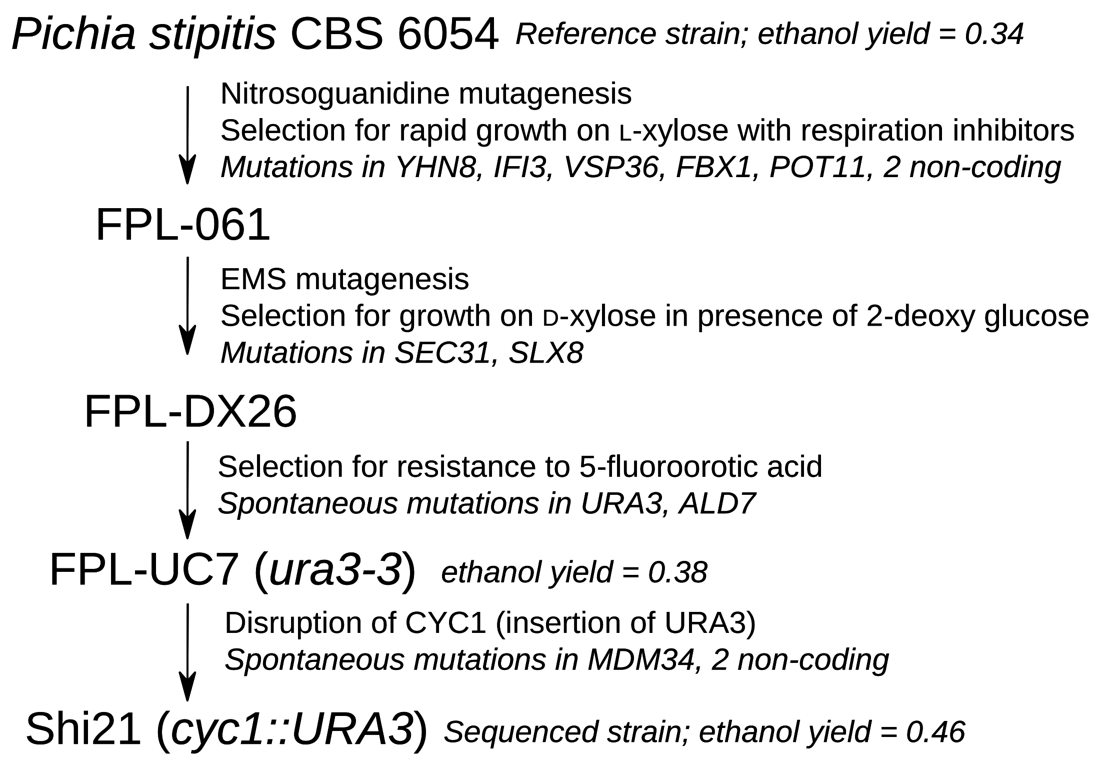

Pichia stipitis CBS 6054 *Reference strain; ethanol yield = 0.34*
Nitrosoguanidine mutagenesis
Selection for rapid growth on L-xylose with respiration inhibitors
Mutations in YHN8, IFI3, VSP36, FBX1, POT11, 2 non-coding
FPL-061
EMS mutagenesis
Selection for growth on D-xylose in presence of 2-deoxy glucose
Mutations in SEC31, SLX8
FPL-DX26
Selection for resistance to 5-fluoroorotic acid
Spontaneous mutations in URA3, ALD7
FPL-UC7 (*ura3-3*) *ethanol yield = 0.38*
Disruption of CYC1 (insertion of URA3)
Spontaneous mutations in MDM34, 2 non-coding
Shi21 (*cyc1::URA3*) *Sequenced strain; ethanol yield = 0.46*

Figure 6.1. Diagram of strain evolution. CBS 6054 is the reference strain originally sequenced by the JGI. Shi21 is the mutant strain that was sequenced using the SOLiD, 454 and Illumina technologies. The intermediate strains, mutagenesis and selection conditions are indicated along with the new mutations that were observed relative to the previous strains at each step. The ethanol yield is also provided in units of g ethanol / g xylose.

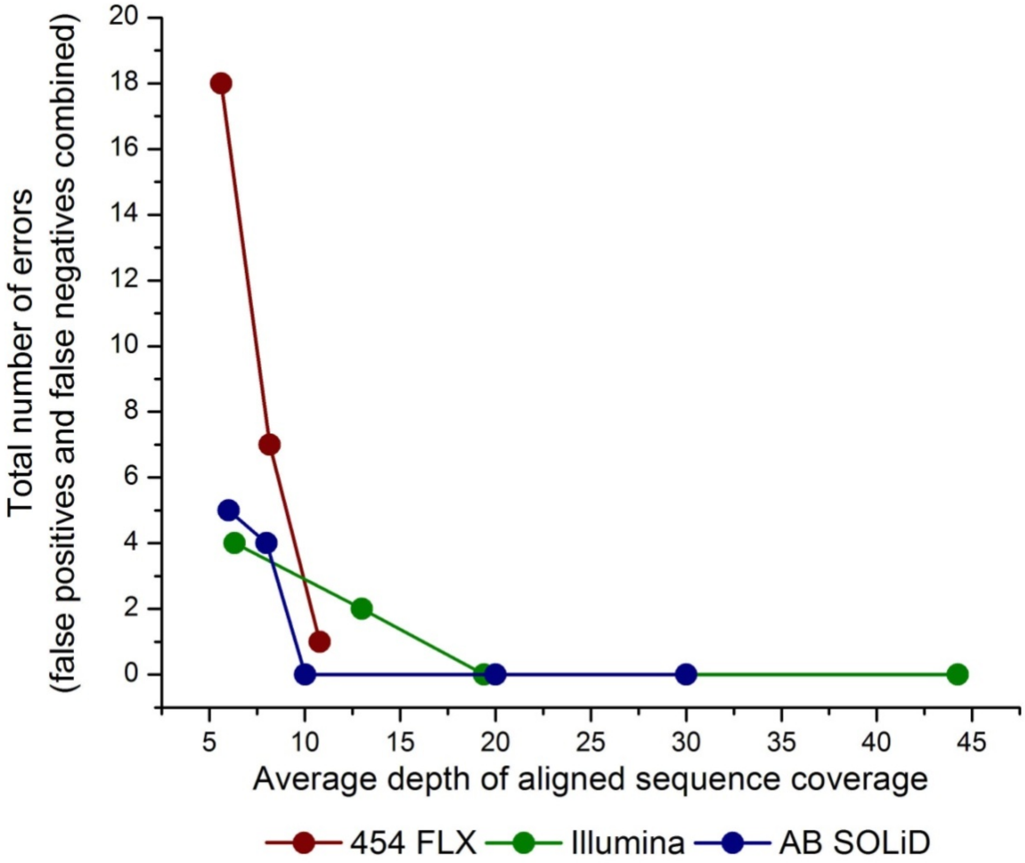


Figure 6.2. The effect of sequence coverage on mutation discovery accuracy. The total number of mutation discovery errors are shown for the three sequencing technologies at various levels of aligned sequence coverage.

Prediction is very difficult, especially about the future.

-Niels Bohr

7. Discussion

Research summary and context

Progress in the field of genomics has been and will continue to be driven by technological, computational, and methodological advances. This has never been more evident than in the last three years, during which time we have seen the development of several radically-different sequencing technologies. While these technologies have undoubtedly changed the way we think about solving open questions in genetic variation, they have also presented a myriad of computational difficulties. These challenges range from the mundane, such as storing and interpreting unprecedented amounts of sequence data to more interesting questions of how to align and quantify the accuracy of sequence reads, and how detect and compare the entire range of genetic variation among individuals.

My dissertation research has shown that these technologies are suitable for the accurate and sensitive discovery of genetic variation. When the first such technology, 454 Life Sciences, was released in 2005 there was skepticism in the genomics community regarding the accuracy and utility of these new data. The subsequent technologies from Illumina/Solexa and Applied Biosystems have faced similar doubts. Therefore, early studies have focused on characterizing the error rates and general utility of the technologies for studies in genetic variation and gene expression.

Pyrobayes, my basecalling algorithm for pyrosequences from 454 Life Sciences, more accurately estimates the error likelihoods associated with each sequence read than the native 454 software. This facilitates more sensitive and economical polymorphism discovery projects by supporting accurate polymorphism calls even at the lowest possible sequence coverage. As of this writing, *Pyrobayes* has been downloaded by over 120 laboratories worldwide. The increased sensitivity afforded by *Pyrobayes* allowed us to discover nearly 600,000 SNPs among 10 inbred isolates of *Drosophila melanogaster*. Despite very low sequence coverage per isolate, twenty percent of these SNPs were shared in at least two isolates. These implicitly-validated polymorphisms represent an average SNP density of roughly one SNP per 1000 base pairs of the *Drosophila* euchromatin and they can therefore be used as a dense marker map for population genetic comparisons in the fruitfly community. Furthermore, this study refutes previous estimates made even by 454 Life Sciences which predicted that at least 20X coverage was necessary for accurate polymorphic discovery with pyrosequences. Therefore, this study informs other researchers that accurate studies in genetic variation can be conducted with this technology at a fraction of previous cost estimates.

Similarly, our genome-wide polymorphism study in *C. elegans* was the first to illustrate that despite the very short read length, the sequence reads from the Illumina/Solexa technology are a sufficient substrate for accurate SNP and INDEL discovery on a whole-genome scale. Given the relatively high sequence coverage, the nearly 97% SNP validation rate we observed was expected. However, the similarly high (~94%) validation rate for short (1-2 bp) insertion-deletion polymorphism was extremely encouraging given the high false positive rate observed in previous studies. This is a welcome finding especially since a primary focus of the nascent 1000 Genomes Project is to develop an extensive catalogue of INDEL polymorphisms in the genomes of presumably healthy individuals. Furthermore, this study was the first to illustrate that sophisticated genome masking methods are necessary in order to prevent improper sequence mapping with short read technologies such as Illumina. We found that the accuracy of polymorphism discovery accuracy dramatically by masking all regions of the *C. elegans* genome that were predisposed to ambiguous sequence mapping. Similar approaches will undoubtedly be used in human resequencing studies with this technology since the human genome is much more repetitive than the worm genome.

We compared the accuracy and required sequence coverage of the three major next-generation sequencing technologies (454, Illumina, AB SOLiD) for discerning the complete mutational profile of a mutant strain of *Pichia stipitis*. In this study, we found that at roughly 10X overall genome sequence coverage, each of the technologies were accurate enough to discover each true point mutation in the mutant strain's genome with very few spurious discoveries. Because of the differences in throughput and cost, 10X sequence coverage is substantially more expensive with the 454 Life Sciences technology than with the Illumina or AB SOLiD technologies. Regardless, all of the technologies are inexpensive enough for smaller laboratories to achieve such coverage for prokaryotes and many model eukaryotes. Accordingly, this study illustrates that the new sequencing technologies can be used as a high-throughput means to discern the causal genotype in traditional reverse genetic screens.

Where is the field headed? What are the next challenges ?

Collectively, these studies illustrate to the larger genomics community that the new sequencing technologies can be used for large-scale studies in genetic variation with similar accuracy and greatly reduced cost relative to traditional sequencing technologies. Moreover, until a few years ago, there was a growing sentiment in the genomics field that "sequencing is dead" as more and more genome-wide association studies were based on chip-based genotyping platforms. Ironically, the current sentiment now seems to be that given the near-exponential throughput growth that the new technologies have seen, it is likely that chip-based technologies may soon be supplanted to similar techniques using sequencing technologies.

Whether or not sequencing technologies replace others, it is clear that as they mature and become ever more economical, they will likely be used to address increasingly sophisticated biological questions and to uncover more subtle genetic variation. As of the February 2008 Advances in Genome Biology and Technology conference, there are four next-generation sequencing technologies that manufactured and sold at least one sequencing machine. A fifth, Pacific Biosciences, appears to be on the cusp of releasing a new sequencing technology which they claim is capable of sequencing the human genome in less than a day for hundreds of dollars. While these claims may never come to fruition, it is quite likely that as more and more technologies compete for business, the costs will eventually reduce to the point that the sequencing of multiple human genomes is possible with funding on the scale of a typical NIH R01 grant. In fact, I believe that within the next ten years, the genomes of thousands of healthy and diseased individuals will have been sequenced. Were the cost of genome sequencing to drop substantially enough to allow this, it would be inevitable that next-generation sequencing machines would become as ubiquitous in molecular biology and genetics laboratories as thermal cyclers for PCR reactions.

However I believe that the potential boon of reduced sequencing costs will quickly be limited by the dearth of appropriate computational and statistical approaches for the interpretation of the vast amounts of genomic data that these technologies will inevitably produce. For example, imagine if a researcher studying genetic predisposition to Type I diabetes could sequence the genomes of thousands of cases and controls quickly and cheaply. What would this researcher do with these data? There are existing statistical techniques for detecting single-nucleotide variants that are associated with a disease phenotype yet there are no such methods for integrating all other types of genetic and epigenetic variation into such an analysis. Moreover, it is unclear whether healthy and diseased genome sequences should be compared to one another directly or whether they should be indirectly compared via comparisons to the human reference genome sequence. I believe the latter approach will bias such comparisons especially when comparing the patterns of structural variations in cohort genomes because segmental duplications and repetitive DNA are ironically the portions that are missing from the reference genome. Thus as more and more human genomes are sequenced to greater completion than the reference genome, it is likely that there will be several reference sequences based on population history and disease predisposition.

Open problems with the extant next-generation sequencing technologies

Longer sequence lengths will improve utility.

An obvious limitation of the existing sequencing technologies is their relatively short read length as compared to traditional capillary sequences. As mentioned, a direct consequence of a shorter read length is the frequent inability to unambiguously determine the origin of a given DNA fragment in the genome. This difficulty is especially troublesome when resequencing the human genome, given the high frequency of repetitive elements such as microsatellites, LINEs and SINEs (Lander et al. 2001). In fact, even assuming a technology were to be developed that made no sequencing errors, only about 91% of the human genome could be resequenced with 50 bp reads (**Figure 7.1**) (Whiteford et al. 2005). Therefore the current read lengths from the Illumina and Applied Biosystems technologies (30-40 bp/read) are insufficient for complete human resequencing studies even with the necessary raw sequence throughput. In contrast, the 400 bp or greater read lengths that are expected from the newest version of the 454 Life Sciences machines are theoretically sufficient to resequence over 99% of the human genome. Unfortunately, the overall throughput from this new machine is expected to be on the order of 500 Mb per machine run. Therefore, based on the current reagent costs, human resequencing with this technology will still likely cost over $100,000.

In order to mitigate the inherent limitations of shorter read lengths, all three current technologies (i.e. 454 Life Sciences, Illumina and AB SOLiD) are working to develop reliable protocols to produce 'paired-end' sequences. Such sequences represent the nucleotide sequences of the two extreme ends of a large fragment of DNA. Ideally, the DNA fragments that are sequenced are uniform in size so that the two sequenced ends (i.e. the 'paired-ends') can be inferred to be a consistent distance apart. This allows one to map the two paired-ends to the genome and assume that the physical distance between the two ends should fall within an expected length distribution. If the paired-end protocols produce reliable DNA fragments, then such methods can be used as a proxy for longer DNA sequences, as the sequence pairs represent the extremities of a much longer original DNA fragment. Unfortunately, the use of current protocols is perturbed by highly-variable and undesirably short DNA fragments.

Representational biases in the DNA sequence libraries.

Ideally, every fragment of DNA in a DNA library will have an equal chance of being sequenced by the next-generation sequencing machines. If this were the case, it would be relative simple to develop probabilistic models that detect when certain regions of the genome are statistically under- (deletions) or over-

(insertions) represented relative to normal statistical fluctuations in sequence coverage (Lander and Waterman 1988). Unfortunately, we have found that each DNA fragment does not have an equal chance of being sequenced by the new technologies. Therefore the resulting representational biases in the sequenced DNA make it difficult to determine whether unexpected sequence coverage is evidence of a true difference in the resequenced genome or whether the difference is merely the result of an inherent bias in the DNA library. Were there no biases, a Poisson distribution would accurately model the expected sequence coverage. As shown in **Figure 7.2**, we find that the observed sequence coverage distribution of the *C. elegans* genome is substantially different from the expected Poisson distribution (results collected by Dr. Chip Stewart). We believe that these biases are a consequence of the whole-genome amplification protocols that are used to increase the amount of total DNA in the library prior to sequencing on the respective sequence machines. The amplification protocols require that genomic DNA be sheared into fragments with either sonication or nebulization methods. We see that A-T rich DNA fragments are underrepresented in the resulting libraries: this is a logical consequence of the DNA fragmentation methods, as G-C rich fragments are more likely to remain as larger fragments because of the extra hydrogen bonds in such regions. Other biases likely exist during the amplification of the resulting fragments prior to sequencing.

Current technologies are still too expensive.

While the next-generation sequencing technologies produce vastly more raw sequence per machine run than traditional capillary sequencers, the cost per sequenced nucleotide is still prohibitively high for the routine sequencing of human genomes. This prevents the application of the new technologies to genome-wide disease association studies. The power of such studies would likely improve dramatically if one compared all the genetic variation among the sequenced cohorts. Yet based on current reagent costs, 10X diploid coverage of the genomes of 1000 human case samples and 1000 human control samples (the typical size of a sufficiently large association study) would cost over 360 million dollars.

Until costs reduce by at least an order of magnitude, such exorbitance will restrict sequence-based variation discovery and association studies to the realm of large international consortia. As of January 2008, the only such large scale study, the ambitious '1000 Genomes Project' is being funded by the NIH and the

Wellcome Trust. Given that the first new sequencing technology (454 Life Sciences) has been available for barely three years, it is likely that increased competition and technological improvements will make large-scale studies more economical.

Once such economy is realized and sequence-based association studies will be feasible with typical NIH funding, it is likely that smaller laboratories will pursue their own genetic association studies that integrate data produced by the larger consortia. This will clearly require standardized methods with which to collect, store and interpret results from these studies. As such, it is imperative that the associated computational methods stay in stride with the increased application of these new technologies.

Informatics and computational methods are immature.

The tools developed for these nascent sequencing technologies have not been subjected to the years of rigorous use placed on analogous tools for traditional capillary sequences. In addition, the sequencing technologies themselves continue to evolve as throughput and read lengths continue to increase, chemistries change and library preparation methods improve. Several open questions therefore remain. For instance, extant tools can align next-generation sequences to organismal reference genome sequences. Yet it is not clear, for any given sequencing technology, how much redundant coverage is needed for exhaustive variation discovery and mutational profiling. In addition, given that the genomes of individuals and strains are often greatly diverged from canonical reference sequences, standard resequencing approaches will often provide incomplete answers. The error profiles of the new sequencing technologies vary rather dramatically. Must we therefore develop computational methods that account for each specific error model? Lastly, methods are now being developed and applied to the discovery of genetic sequence variants; yet other, epigenetic variations have been demonstrated to be of medical importance. The discovery of such variations comes with special informatics challenges: for example, how do we align short-read sequences from bisulfite-treated (and therefore greatly mutated) DNA?

It is clear that the continued development of efficient and accurate computational methods will be necessary to support the increased use of the new sequencing technologies. Additionally, as the analytic methods improve, similar advancements will be required to standardize the storage and interpretation of results based on such methods.

Future studies

My research thus far has focused on the development of accurate methods for polymorphism discovery with the new sequencing technologies. As the accuracy, throughput, and economy of sequencing technologies improve, they will rapidly become amenable to many more research areas. I have expertise in the analysis of these data and wish to extend this knowledge to the improvement of traditional genetic and genomic methods such as *de novo* genome assembly and genetic screens. In addition, I would like to develop novel methods for the detection of DNA methylation and rare genetic variants. The primary focus of each future research goal outlined below is the development of reliable computational methods that extend the utility of the next-generation technologies and that can be easily used by other researchers in the field.

Develop efficient methods for the rapid mutational profiling of model organisms.

Once a desired phenotype has been observed in a mutant strain, traditional forward genetic screens are hindered by the time and labor required to map the specific mutation(s) responsible for the phenotype. However, since the mutant strains are typically generated through successive rounds of chemical mutagenesis, the mutant and the parental strain usually only differ by a small number of single-nucleotide, point mutations. The cost and throughput of the new sequencing technologies are sufficient to economically generate deep coverage of many model organisms (e.g. *E. coli, S. cerevisiae*). Therefore, as described in Chapter 6, it is possible to generate complete mutational profiles by comparing the resequenced mutant to the reference sequence for the same species. Thus, a laborious step in forward genetic screens can be substantially simplified.

I believe that this study is indicative of the power and economy of the new sequencing technologies. I would like to further develop the experimental and computational technologies involved in this study and expand them for use in reverse genetic studies where other genetic variations such as small INDELs, copy number changes and structural variations must be discovered. Further advances in the throughput and parallelization of the new sequencing technologies will undoubtedly facilitate the generation of rapid mutational profiles for many mutant strains at a time. For example, the 454 Life Sciences and

Illumina sequencing technologies can physically separate DNA from 16 and 8 samples, respectively. Yet for many pathogens such parallelization still produces too much redundant sequence. As “barcoding” (that is, using short, sample-specific DNA tags to identify sequence reads from individual samples) technology advances, further parallelization will allow the economical profiling of tens to hundreds of samples with a single machine run.

I plan to continue my collaboration with Gabor Marth in order to further refine the computational approaches to this research. I anticipate that single-end sequencing reads from either the Illumina or AB SOLiD platforms would provide a sufficient substrate for mutation discovery in microbial and yeast genomes. I initially expect this research to result in streamlined mutational profiling software for prokaryotes and lower eukaryotes and I expect to collaborate with others to define reliable protocols for discerning mutational profiles from similar studies in higher eukaryotes.

Develop reliable methods to discover rare variants.

As a result of the extensive research in human genetic variation (e.g. The International HapMap Consortium and The SNP Consortium) since the completion of the human genome sequence, it is widely believed that the vast majority of SNPs with a minor allele frequency (MAF) of at least 5% are known. However it is unclear to what degree these so-called "common" SNPs (**Figure 7.3**) represent the landscape of less common single-nucleotide polymorphism in the human genome. Therefore, a primary focus of the 1000 Genomes Project is to discover the majority of SNPs in the entire genome with a MAF of at least 1%. An additional goal is to uncover all SNPs in gene coding regions with a MAF of at least 0.1%.

The accurate discovery of such rare variants with next-generation technologies presents an intriguing statistical challenge. Since the sequencing reads produced by these technologies are clonally-amplified with various chemical approaches, each sequencing read inherently represents a fragment of DNA from a single chromosome of a chromosome pair. However, rare polymorphic alleles manifest in a diploid population as infrequent heterozygotes or even more infrequent homozygotes for the alternate allele. Thus, one must typically detect rare polymorphisms from infrequent heterozygous individuals among a much larger resequenced cohort (**Figure 7.3**). The detection of such a rare heterozygote with next-generation sequences therefore requires that DNA fragments from each allele are sequenced.

Assuming no biases for one chromosome in a chromosome pair for a given individual, there is a binomial likelihood that the sequence reads at a given locus were sampled from both chromosomes. Thus deeper sequence coverage at a heterozygous locus increases the likelihood that both heterozygous alleles were sampled. For example, if two reads are present at a heterozygous locus, there is a 50% chance that the two reads came from the two alleles (i.e. one read from each allele). While this is an unacceptably low likelihood, with ten reads, the likelihood jumps to 99.8%. Unfortunately, DNA library construction protocols cannot enforce that all loci are covered by ten sequence reads. Instead, there are always statistical fluctuations in coverage throughout the genome. With no representational bias in the sequence library, these fluctuations in depth of coverage follow a Poisson distribution. This means that if we intended to have an average depth of ten reads per loci (10X coverage), then according to a Poisson distribution, 15% of the loci would have exactly 10X coverage while nearly 100% of the loci would have at least 2X coverage (**Figure 7.4**).

Thus the detection of rare alleles in a large cohort of individuals will require rather sophisticated methods that account for the likelihood the full genotype was sampled given aligned reads at a given locus. Such methods become more difficult to develop when one accounts for the fact that depth of coverage has been observed to not follow a Poisson distribution owing to representational bias (see above). In addition, one must account for the fact that the new sequencing technologies make sequencing errors and therefore, the observed alleles in a given sequencing read will not always reflect the actual DNA that was sequenced. Clearly such problems are mitigated by deep sequence coverage, but since the new technologies are still relatively expensive, deep sequence coverage for each individual in a large study will be too costly. Accordingly, I intend to develop a new method for the discovery of common and rare alleles that incorporates a) the likelihood that both alleles were sampled for each individual, b) the likelihood that the reads were correctly aligned to the locus in question and c) local haplotype information that may corroborate the allele based on linkage disequilibrium with other local alleles.

Develop statistical and computational methods to integrate all types of genetic and epigenetic variation into disease association studies.

Nearly all genome-wide association studies seeking to uncover genetic variants that are statistically correlated to a given disease phenotype solely interrogate SNP markers to detect phenotype associations. There are several reasons that SNPs have remained the primary if not sole type of genetic variation studied in such research. First, SNPs are by far the most widely-characterized type of genetic variation in the human genome. Second, several inexpensive chip-based technologies (e.g. Illumina, Affymetrix, Nimblegen, etc.) have been developed to interrogate hundreds of thousands of SNPs in a single experiment. Finally, SNPs are predominantly di-allelic and as such are amenable to many standard statistical tests.

However, despite finding SNP markers that are correlated with the disease phenotype, most association studies thus far have failed to identify the causal functional allele(s). One reason for this may be that the phenotypes of the study cohorts have been poorly characterized and therefore weaken the power of the study. Yet another plausible reason is that other possibly causal types of genetic variation such as INDELs, copy number variations and epigenetic variation are ignored in current association study methods. While there is evidence that INDELs are often in linkage disequilibrium with regional SNPs, it is unclear to what degree copy number and structural variations (i.e. inversions and translocation) remain linked with nearby SNPs. If there is generally weak linkage disequilibrium between SNPs and nearby structural variation, then SNP markers are likely a weak proxy for detecting association between structural variations and disease. Moreover, it is unlikely that there is any linkage between SNP alleles and epigenetic variations such as CpG methylation.

Therefore it is conceivable that interrogating additional types of genetic variation in disease association studies will both increase the power of such studies and facilitate the identification of functional alleles. Yet the nature of such additional types of genetic variation will undoubtedly complicate the statistical methods used in association studies. For example, unlike di-allelic SNPs which nearly always have just three possible states (i.e. heterozygotes and the two possible homozygotes), the genotypes of INDEL polymorphisms are potentially more variable and difficult to conclusively assay. Copy number variations face the same difficulty as they inherently vary in degree (e.g. 1 copy, 2 copies, 3 copies, …, *n* copies), size (e.g. length of the deletion or insertion) and structure (e.g.

insertions and deletions occur in *cis*, while translocations occur in *trans*). Clearly, it is difficult to merely genotype such variations among a large number of individuals. In addition, correlating such plastic genetic variation with disease phenotypes will necessitate advanced statistical methods in order to integrate them into future genome-wide association studies.

Create novel methods for de novo genome assembly guided by known reference genomes.

De novo genome assembly traditionally involves the coalescence of whole-genome shotgun sequences into contigs by searching for substantial overlaps between pairs of sequence fragments. This process is computationally expensive and usually results in thousands of shorter contigs for repetitive genomes. Several new methods have been developed for the assembly of reads from next-generation sequencing technologies (e.g. SHARCGS and SSAKE) (Dohm et al. 2007; Warren et al. 2007). However, they require massive amounts of computer memory and as of this writing, they only produce reasonable results with the longer 454 Life Sciences reads. The rapid assembly of viral and microbial genomes is crucial in order to ascertain the genetic origins of particularly virulent strains and to develop therapies to combat them.

Finished reference sequences are available for hundreds of species. Since the primary bottleneck in traditional *de novo* genome assemblies is the pair-wise comparison of millions of sequence fragments, I intend to develop a hybrid approach that exploits known reference genome sequences to facilitate assembly and reduce computational expense. This approach, termed "guided de novo assembly" (**Figure 7.5**), is particularly attractive for assembling the genomes of pathogens, as there are currently over 2,000 finished viral and microbial genomes. Specifically, for a given un-sequenced species, one would need at least one reference genome sequence from a closely related species. Sequences from whole-genome libraries of the species in question would be aligned to the known reference genome of the similar species. In so doing, a substantial fraction of genome sequence of the unknown species should be ascertained. This fraction is clearly a function of the evolutionary distance between the two species. All unaligned reads could then be assembled into contigs using traditional methods or the novel approaches mentioned above. The resulting contigs could then be combined with the alignments to the reference genome to provide a reasonable draft genome sequence.

This approach is also attractive for de novo human genome assembly. Recent research indicates that structural variation accounts for anywhere between 4 and 25 Mb of genetic difference between any two individuals (Iafrate et al. 2004; Sebat et al. 2007; Sebat et al. 2004). Therefore, canonical resequencing approaches that merely align sequence to the human reference sequence are inherently limited in their ability to detect larger insertions, deletions, copy-number changes and structural rearrangements. Continued improvement of paired-end libraries for the new sequencing technologies will also increase the utility of guided *de novo* assembly methods for use in human assemblies.

I have developed prototypes for this methodology using a reference-guided alignment algorithm (*Mosaik*) developed by Michael Stromberg in our laboratory, as well as SHARGCS for the *de novo* assembly of the remaining unaligned sequences. In collaboration with a colleague at Washington University, I have begun to test this approach on three different Enterohaemorrhagic Escherichia coli (EHEC) strains, which have diverged rather substantially from the reference K12 strain. Previous studies have shown that reasonable assemblies of bacteria are possible with the longer reads produced by the 454 Life Sciences technologies. For this reason, I would advocate the use of either the single-end

454 libraries or paired-end libraries from the AB SOLiD technology, as they produce paired-end libraries with longer, less variable insert fragments. Important control experiments clearly must be performed on already finished microbial genomes where one could directly compare the accuracy and completeness of the newly-assembled genome with the finished genome. My goal is to produce a robust software package for guided *de novo* assembly that will be freely available for non-profit use.

Investigate the landscape of promoter CpG methylation in diverse, healthy tissues and among individuals.

Abnormal DNA methylation is a primary epigenetic hallmark of many types of cancer. Specifically, various cancers have been shown to exhibit both hypomethylation and hypermethylation in certain promoter regions. CpG cytosine methylation in promoter regions has been shown to inhibit the binding of transcription factors, which in turn, prevents the subsequent transcription of the associated gene (Eckhardt et al. 2004; Murrell et al. 2005). Excessive promoter hypermethylation (thus increased suppression) in various cancers has been shown in tumor suppressor genes, whereas hypomethylation (thus reduced suppression) has been shown in cell growth genes such as IGF (Cho et al. 2007;

Kim et al. 2008; Song et al. 2007; Yoo and Jones 2006). It is therefore crucial that we understand the landscape of normal DNA methylation in many healthy cell types so we can more precisely understand the mechanisms and extent of aberrant methylation in cancers. It is similarly important that we understand the variability of "normal" methylation profiles among cell types and individuals so that informative comparisons can be made between healthy and neoplastic cells.

Towards this goal, I would like to combine the throughput of the new sequencing platforms with probe-based, targeted genomic capture methods to investigate the methylation profile of all known promoters in the human genome. Several new probe-based methods have recently been developed which use custom designed probes tiling thousands of specific genomic regions and allow for the targeted capture of genomic DNA by hybridization. Such a chip could be designed to tile all of the known promoters in the human genome. The captured genomic DNA from the targeted promoters could subsequently be subjected to bisulfite treatment (which converts unmethylated cytosines to uracil while methylated cytosines remain unchanged) and then sequenced to deep coverage with any of the new sequencing technologies (**Figure 7.6**). Current capture methods are capable of interrogating ca. 30,000 distinct genomic regions with overlapping probes ranging from 60-90 nucleotides (Hodges et al. 2007).

While bisulfite sequencing remains the “gold standard” for detecting DNA methylation, a consequence of this approach is that after bisulfite treatment and PCR amplification, all unmethylated cytosines are converted to thymines. This conversion reduces the complexity of the sequence read, especially in DNA fragments that are largely unmethylated and cytosine-rich. Because of the reduced sequence complexity, it becomes more difficult to map the resulting sequence reads to their promoter of origin. A plausible approach to this dilemma is to perform an *in silico* bisulfite conversion of the human genome while assuming that all CpG cytosines are unmethylated. The bisulfite-converted sequence reads could then be aligned to the converted reference genome while employing a reduced C/T mismatch penalty to account for the proper alignment of unmethylated cytosines (which are now thymines). The 250-400 bp sequence reads produced by the 454 Life Sciences technologies are much more suitable for this experiment than the shorter-read (25-50 bp) technologies because the overall reduction in complexity caused by bisulfite conversion is mitigated by longer sequence fragments.

Once reliable alignment methods have been established, it becomes tractable to compare the methylation states of different cell types (see bottom alignment of **Figure 7.6**) and different individuals. Yet one must also account for the fact that variable “epigenetic haplotypes” may exist in different cell types and individuals. That is, a given CpG dinucleotide may have a different methylation state on each DNA strand and/or chromosome and there may be variable methylation states among cell types from the same individual. The development of sophisticated software for alignment and for the detection of epigenetic variation is therefore crucial to the success of this research.

The aim of this research is to discern whether consistent methylation patterns exist among the same cell type from different individuals. If so, further investigation into whether or not these patterns are correlated with consistent expression profiles would hopefully provide a baseline with which to compare cancerous cells of the same type. If such patterns exist, further studies would be required to investigate the mechanisms that drive the drastic change in methylation in the cell. This research would ideally define specific protocols and data analysis software for the accurate detection of differential CpG methylation among many cell types and individuals.

Summary

My dissertation work has been focused on the development of accurate methods for polymorphism discovery with the next-generation sequencing technologies. I have developed substantial expertise in the analysis of these data and hope to use this experience in a clinical research environment. I believe that as technologies continue to improve, the complexities of biological research will shift to computational and statistical interpretation of the vast amounts of data that these technologies will inevitably produce. As of this writing, computational biology is still considered somewhat of a niche research area in the larger field of biology. Yet many high-throughput sub-fields such as genomics and proteomics have already become inexorably reliant upon sophisticated computational methods.

It seems that at the current pace of technological advancement (the current pace is greater than Moore's Law), it will soon be possible for typical academic laboratories to cheaply and accurately sequence complex genomes. Additionally, these technologies will also enable sensitive studies in transcript expression, DNA methylation and chromatin modifications. When taken together, such experiments will undoubtedly provide a more detailed picture of the relationship between genotype (and/or "epigenotype") and phenotype. The hope is that such advancements will further our understanding of disease predisposition and etiology and translate into improved therapies for human diseases.

The computer software and ancillary methods that were written as part of this dissertation will be archived on physical media (e.g., DVDs) and will be placed in the care of the Marth laboratory for future reference.

Chapter 7 Figures

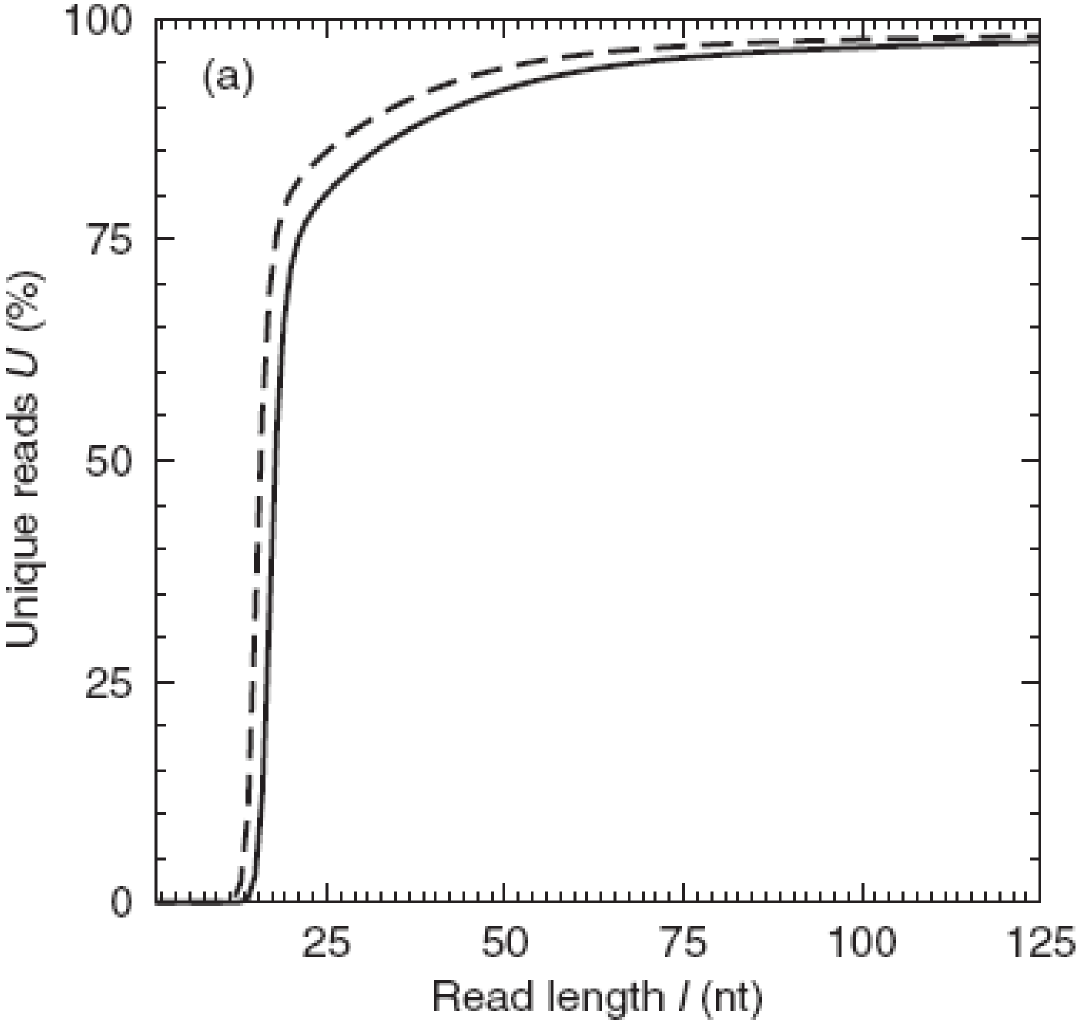


Figure 7.1. Percentage of unique sub-sequences in the human genome. (modified from Whiteford et al, 2005). As the theoretical sequence read length increases (x-axis), so does the fraction of human chromosome 1 (dashed line) and the entire human genome (solid line). Longer reads clearly enable more complete genome resequencing.

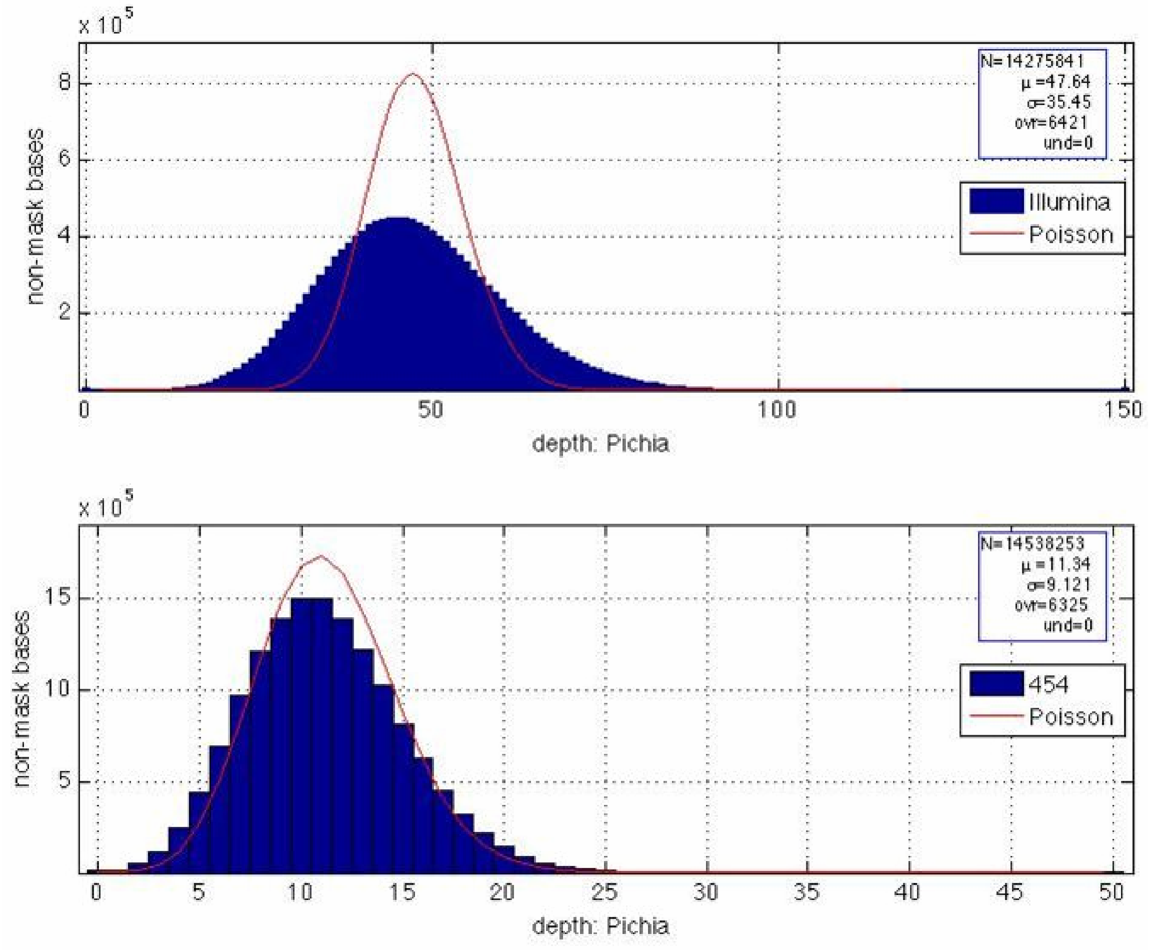


Figure 7.2. Representational biases in the depth of sequence coverage (Donald Stewart, personal communication). Overall depth of coverage is shown for Illumina (top) and 454 Life Sciences (bottom) sequencing runs of the *Pichia* stipitis genome. The expected Poisson distribution of sequence coverage (pink) is shown in comparison to the observed distribution of sequence coverage (blue). The degree of difference in these distributions varies between the sequencing technologies and reflects sequence-biases in the DNA library fragmentation and amplification steps prior to sequencing.

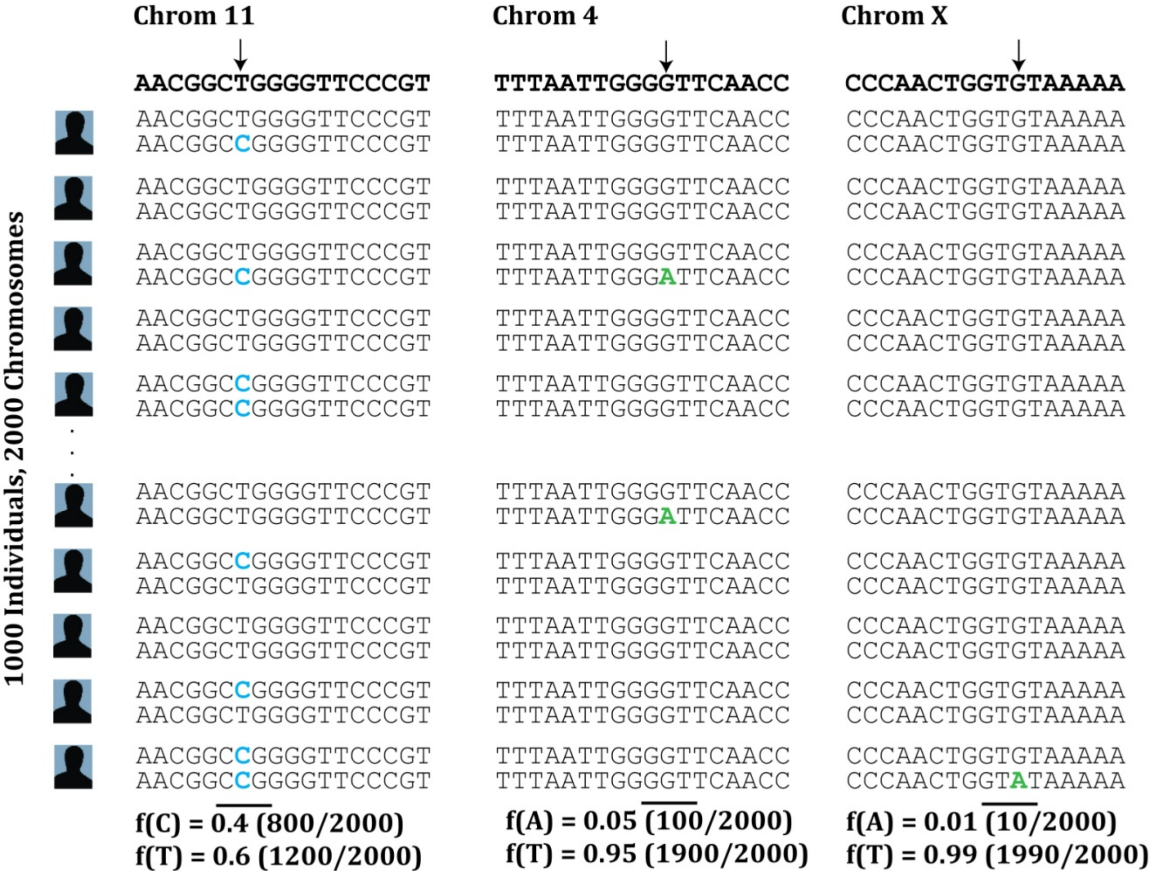

1000 Individuals, 2000 Chromosomes
Chrom 11
Chrom 4
Chrom X
AACGGCTGGGGTTCCCGT
TTTAATTGGGGTTCAACC
CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGATTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGATTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCTGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTGTAAAAA
AACGGCCGGGGTTCCCGT TTTAATTGGGGTTCAACC CCCAACTGGTATAAAAA
f(C) = 0.4 (800/2000)
f(T) = 0.6 (1200/2000)
f(A) = 0.05 (100/2000)
f(T) = 0.95 (1900/2000)
f(A) = 0.01 (10/2000)
f(T) = 0.99 (1990/2000)

Figure 7.3. Examples of polymorphisms with varying minor allele frequencies in a human population. Cartoons indicating **"**common" polymorphisms (left; minor allele frequency of 40%, middle; minor allele frequency of 5%) and a "rare" polymorphism (right; minor allele frequency of less than 1%) are shown. The cumulative contribution of rare alleles to phenotype is unknown.

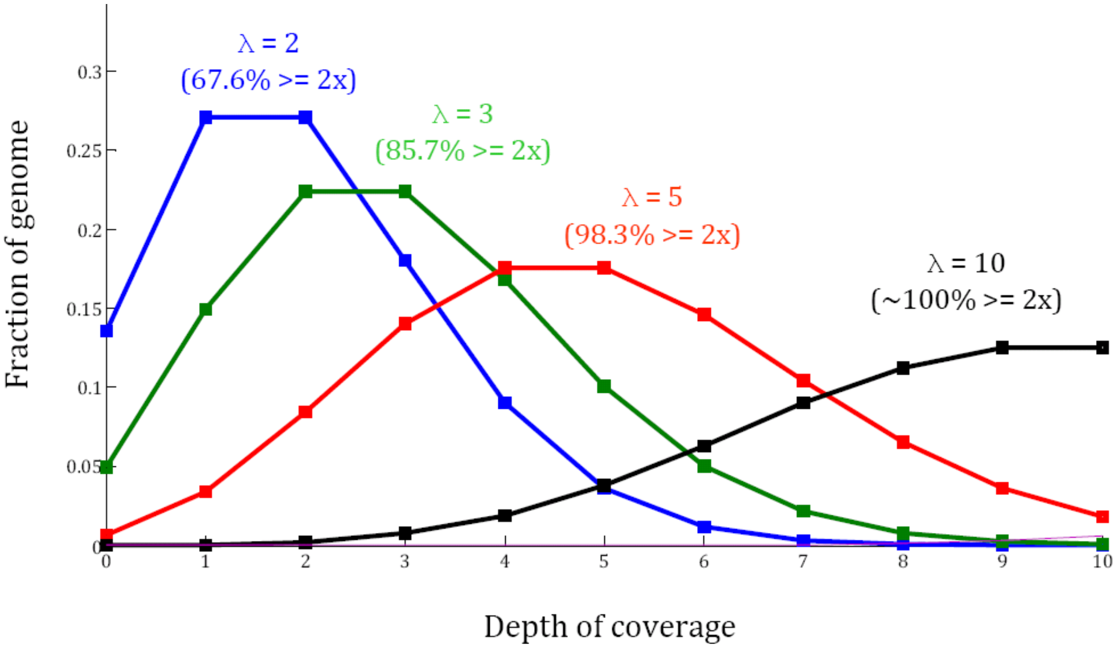


Figure 7.4. Sequence coverage distribution based on an unbiased Poisson expectation. Assuming no representation biases, the expected distribution of sequence coverage is shown for various sequence coverage (lambda). The fraction of positions with greater than 2X coverage at each average coverage (lambda) are shown.

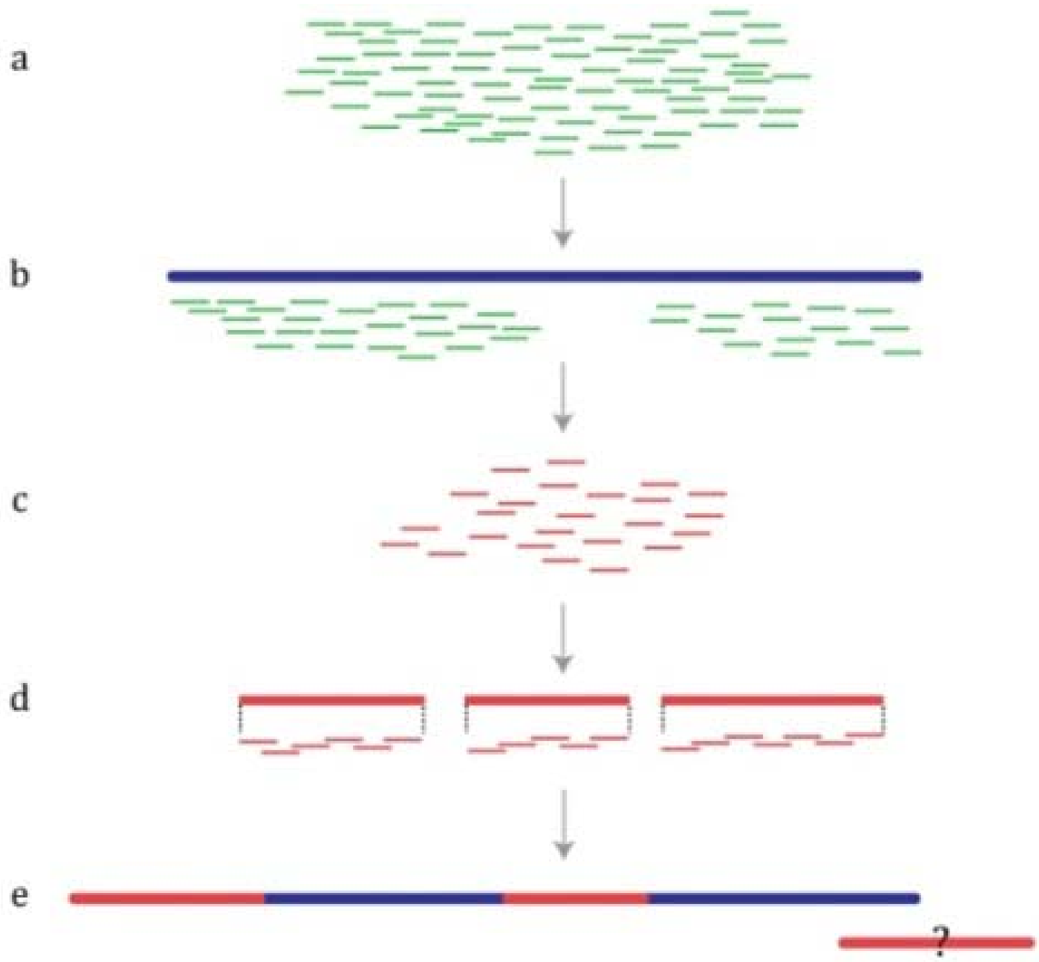


Figure 7.5. Guided *de novo* genome assembly. Shotgun sequences from a yet-unsequenced species (a) are aligned to a known reference genome from a closely-related species (b). The unaligned reads (c) are assembled into contigs (d) using short-read assembly methods. The resulting contigs are then mapped to the original reference genome sequence to produce a draft genome sequence for the species in question (e). However, not all contigs can be unambiguously oriented in the new draft sequence.

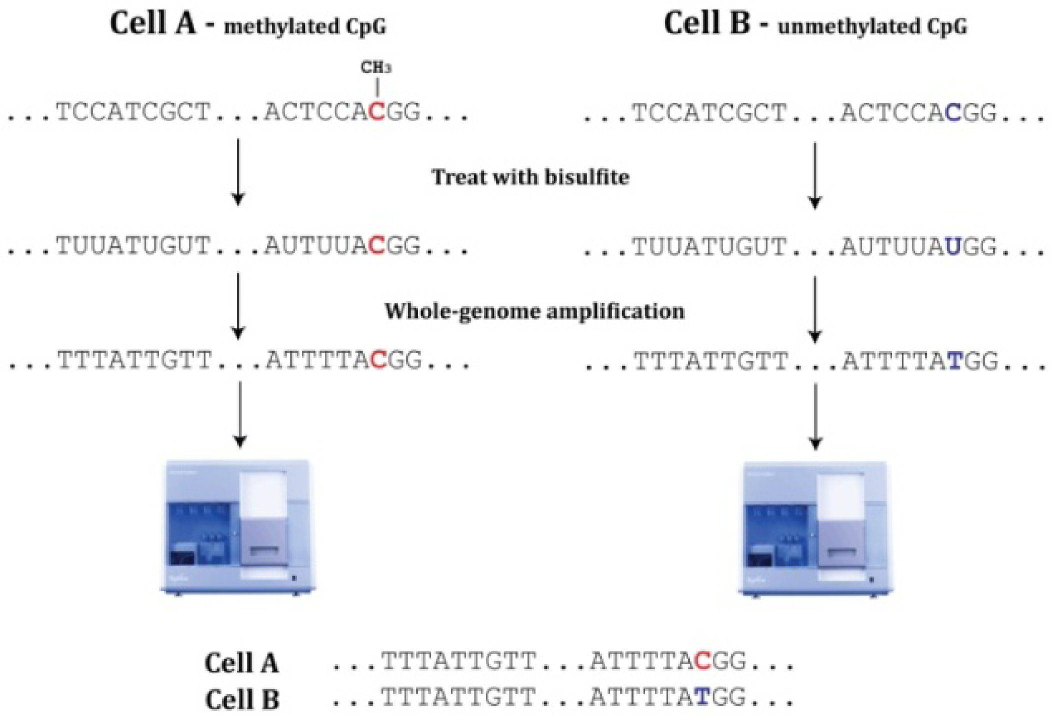

Cell A - methylated CpG
Cell B - unmethylated CpG
CH3
...TCCATCGCT...ACTCCACGG...
...TCCATCGCT...ACTCCACGG...
Treat with bisulfite
...TUUATUGUT...AUTUUACGG...
...TUUATUGUT...AUTUUAUGG...
Whole-genome amplification
...TTTATTGTT...ATTTTACGG...
...TTTATTGTT...ATTTTATGG...
Cell A ...TTTATTGTT...ATTTTACGG...
Cell B ...TTTATTGTT...ATTTTATGG...

Figure 7.6. High-throughput bisulfite sequencing with new sequencing technologies. Bisulfite treatment converts all unmethylated cytosines to uracil yet leaves methylated CpG cytosines unchanged. After PCR amplification, uracils are converted to thymine. After conversion, the relative methylation states of different cells or individuals can be compared via sequence alignment.

References

Adams, M.D. S.E. Celniker R.A. Holt C.A. Evans J.D. Gocayne P.G. Amanatides S.E. Scherer P.W. Li R.A. Hoskins R.F. Galle R.A. George S.E. Lewis S. Richards M. Ashburner S.N. Henderson G.G. Sutton J.R. Wortman M.D. Yandell Q. Zhang L.X. Chen R.C. Brandon Y.H. Rogers R.G. Blazej M. Champe B.D. Pfeiffer K.H. Wan C. Doyle E.G. Baxter G. Helt C.R. Nelson G.L. Gabor J.F. Abril A. Agbayani H.J. An C. Andrews-Pfannkoch D. Baldwin R.M. Ballew A. Basu J. Baxendale L. Bayraktaroglu E.M. Beasley K.Y. Beeson P.V. Benos B.P. Berman D. Bhandari S. Bolshakov D. Borkova M.R. Botchan J. Bouck P. Brokstein P. Brottier K.C. Burtis D.A. Busam H. Butler E. Cadieu A. Center I. Chandra J.M. Cherry S. Cawley C. Dahlke L.B. Davenport P. Davies B. de Pablos A. Delcher Z. Deng A.D. Mays I. Dew S.M. Dietz K. Dodson L.E. Doup M. Downes S. Dugan-Rocha B.C. Dunkov P. Dunn K.J. Durbin C.C. Evangelista C. Ferraz S. Ferriera W. Fleischmann C. Fosler A.E. Gabrielian N.S. Garg W.M. Gelbart K. Glasser A. Glodek F. Gong J.H. Gorrell Z. Gu P. Guan M. Harris N.L. Harris D. Harvey T.J. Heiman J.R. Hernandez J. Houck D. Hostin K.A. Houston T.J. Howland M.H. Wei C. Ibegwam M. Jalali F. Kalush G.H. Karpen Z. Ke J.A. Kennison K.A. Ketchum B.E. Kimmel C.D. Kodira C. Kraft S. Kravitz D. Kulp Z. Lai P. Lasko Y. Lei A.A. Levitsky J. Li Z. Li Y. Liang X. Lin X. Liu B. Mattei T.C. McIntosh M.P. McLeod D. McPherson G. Merkulov N.V. Milshina C. Mobarry J. Morris A. Moshrefi S.M. Mount M. Moy B. Murphy L. Murphy D.M. Muzny D.L. Nelson D.R. Nelson K.A. Nelson K. Nixon D.R. Nusskern J.M. Pacleb M. Palazzolo G.S. Pittman S. Pan J. Pollard V. Puri M.G. Reese K. Reinert K. Remington R.D. Saunders F. Scheeler H. Shen B.C. Shue I. Siden-Kiamos M. Simpson M.P. Skupski T. Smith E. Spier A.C. Spradling M. Stapleton R. Strong E. Sun R. Svirskas C. Tector R. Turner E. Venter A.H. Wang X. Wang Z.Y. Wang D.A. Wassarman G.M. Weinstock J. Weissenbach S.M. Williams WoodageT K.C. Worley D. Wu S. Yang Q.A. Yao J. Ye R.F. Yeh J.S. Zaveri M. Zhan G. Zhang Q. Zhao L. Zheng X.H. Zheng F.N. Zhong W. Zhong X. Zhou S. Zhu X. Zhu H.O. Smith R.A. Gibbs E.W. Myers G.M. Rubin and J.C. Venter. 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

Ahmadian, A., M. Ehn, and S. Hober. 2006. Pyrosequencing: history, biochemistry and future. *Clin Chim Acta* **363:** 83-94.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Altshuler, D., V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, and E.S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513-516.

Audrezet, M.P., J.M. Chen, O. Raguenes, N. Chuzhanova, K. Giteau, C. Le Marechal, I. Quere, D.N. Cooper, and C. Ferec. 2004. Genomic rearrangements in the CFTR gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Hum Mutat* **23:** 343-357.

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7:** 781-791.

Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823-837.

Bentley, D.R. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16:** 545-552.

Birney, E. J.A. Stamatoyannopoulos A. Dutta R. Guigo T.R. Gingeras E.H. Margulies Z. Weng M. Snyder E.T. Dermitzakis R.E. Thurman M.S. Kuehn C.M. Taylor S. Neph C.M. Koch S. Asthana A. Malhotra I. Adzhubei J.A. Greenbaum R.M. Andrews P. Flicek P.J. Boyle H. Cao N.P. Carter G.K. Clelland S. Davis N. Day P. Dhami S.C. Dillon M.O. Dorschner H. Fiegler P.G. Giresi J. Goldy M. Hawrylycz A. Haydock R. Humbert K.D. James B.E. Johnson E.M. Johnson T.T. Frum E.R. Rosenzweig N. Karnani K. Lee G.C. Lefebvre P.A. Navas F. Neri S.C. Parker P.J. Sabo R. Sandstrom A. Shafer D. Vetrie M. Weaver S. Wilcox M. Yu F.S. Collins J. Dekker J.D. Lieb T.D. Tullius G.E. Crawford S. Sunyaev W.S. Noble I. Dunham F. Denoeud A. Reymond P. Kapranov J. Rozowsky D. Zheng R. Castelo A. Frankish J. Harrow S. Ghosh A. Sandelin I.L. Hofacker R. Baertsch D. Keefe S. Dike J. Cheng H.A. Hirsch E.A. Sekinger J. Lagarde J.F. Abril A. Shahab C. Flamm C. Fried J. Hackermuller J. Hertel M. Lindemeyer K. Missal A. Tanzer S. Washietl J. Korbel O. Emanuelsson J.S. Pedersen N. Holroyd R. Taylor D. Swarbreck N. Matthews M.C. Dickson D.J. Thomas M.T. Weirauch J. Gilbert J. Drenkow I. Bell X. Zhao K.G. Srinivasan W.K. Sung H.S. Ooi K.P. Chiu S. Foissac T. Alioto M. Brent L. Pachter M.L. Tress A. Valencia S.W. Choo C.Y. Choo C. Ucla C. Manzano C. Wyss E. Cheung T.G. Clark J.B. Brown M. Ganesh S. Patel H. Tammana J. Chrast C.N. Henrichsen C. Kai J. Kawai U. Nagalakshmi J. Wu Z. Lian J. Lian P. Newburger X. Zhang P. Bickel J.S. Mattick P. Carninci Y. Hayashizaki S. Weissman T. Hubbard R.M. Myers J. Rogers P.F. Stadler T.M. Lowe C.L. Wei Y. Ruan K. Struhl M. Gerstein S.E. Antonarakis Y. Fu E.D. Green U. Karaoz A. Siepel J. Taylor L.A. Liefer K.A. Wetterstrand P.J. Good E.A. Feingold M.S. Guyer G.M. Cooper G. Asimenos C.N. Dewey M. Hou S. Nikolaev J.I. Montoya-Burgos A. Loytynoja S. Whelan F. Pardi T. Massingham H. Huang N.R. Zhang I. Holmes J.C. Mullikin A. Ureta-Vidal B. Paten M. Seringhaus D. Church K. Rosenbloom W.J. Kent E.A. Stone S. Batzoglou N. Goldman R.C. Hardison D. Haussler W. Miller A. Sidow N.D. Trinklein Z.D. Zhang L. Barrera R. Stuart D.C. King A. Ameur S. Enroth M.C. Bieda J. Kim A.A. Bhinge N. Jiang J. Liu F. Yao V.B. Vega C.W. Lee P. Ng A. Shahab A. Yang Z. Moqtaderi Z. Zhu X. Xu S. Squazzo M.J. Oberley D. Inman M.A. Singer T.A. Richmond K.J. Munn A. Rada-Iglesias O. Wallerman J. Komorowski J.C. Fowler P. Couttet A.W. Bruce O.M. Dovey P.D. Ellis C.F. Langford D.A. Nix G. Euskirchen S. Hartman A.E. Urban P. Kraus S. Van Calcar N. Heintzman T.H. Kim K. Wang C. Qu G. Hon R. Luna C.K. Glass M.G. Rosenfeld S.F. Aldred S.J. Cooper A. Halees J.M. Lin H.P. Shulha X. Zhang M. Xu J.N. Haidar Y. Yu Y. Ruan V.R. Iyer R.D. Green C. Wadelius P.J. Farnham B. Ren R.A. Harte A.S. Hinrichs H. Trumbower H. Clawson J. Hillman-Jackson A.S. Zweig K. Smith A. Thakkapallayil G. Barber R.M. Kuhn D. Karolchik L. Armengol C.P. Bird P.I. de Bakker A.D. Kern N. Lopez-Bigas J.D. Martin B.E. Stranger A. Woodroffe E. Davydov A. Dimas E. Eyras I.B. Hallgrimsdottir J. Huppert M.C. Zody G.R. Abecasis X. Estivill G.G. Bouffard X. Guan N.F. Hansen J.R. Idol V.V. Maduro B. Maskeri J.C. McDowell M. Park P.J. Thomas A.C. Young R.W. Blakesley D.M. Muzny E. Sodergren

D.A. Wheeler K.C. Worley H. Jiang G.M. Weinstock R.A. Gibbs T. Graves R. Fulton E.R. Mardis R.K. Wilson M. Clamp J. Cuff S. Gnerre D.B. Jaffe J.L. Chang K. Lindblad-Toh E.S. Lander M. Koriabine M. Nefedov K. Osoegawa Y. Yoshinaga B. Zhu and P.J. de Jong. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799-816.

Boeke, J.D., F. LaCroute, and G.R. Fink. 1984. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol Gen Genet* **197:** 345-346.

Braverman, J.M., R.R. Hudson, N.L. Kaplan, C.H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140:** 783-796.

Brockman, W., P. Alvarez, S. Young, M. Garber, G. Giannoukos, W.L. Lee, C. Russ, E.S. Lander, C. Nusbaum, and D.B. Jaffe. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res*.

Carter, N.P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39:** S16-21.

Chakravarti, A. 1999. Population genetics--making sense out of sequence. *Nat Genet* **21:** 56-60.

Chen, N., T.W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C.K. Chen, W.J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E.M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P.W. Sternberg, and L.D. Stein. 2005. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res* **33:** D383-389.

Cho, Y.G., J.H. Song, C.J. Kim, S.W. Nam, N.J. Yoo, J.Y. Lee, and W.S. Park. 2007. Genetic and epigenetic analysis of the KLF4 gene in gastric cancer. *Apmis* **115:** 802-808.

Clark, T.G., T. Andrew, G.M. Cooper, E.H. Margulies, J.C. Mullikin, and D.J. Balding. 2007. Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biol* **8:** R180.

Collins, F.S., L.D. Brooks, and A. Chakravarti. 1998a. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8:** 1229-1231.

Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998b. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282:** 682-689.

Conrad, D.F., T.D. Andrews, N.P. Carter, M.E. Hurles, and J.K. Pritchard. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38:** 75-81.

Daly, M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* **29:** 229-232.

Denver, D.R., K. Morris, and W.K. Thomas. 2003. Phylogenetics in Caenorhabditis elegans: an analysis of divergence and outcrossing. *Mol Biol Evol* **20:** 393-400.

Dohm, J.C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res* **17:** 1697-1706.

Eckhardt, F., S. Beck, I.G. Gut, and K. Berlin. 2004. Future potential of the Human Epigenome Project. *Expert Rev Mol Diagn* **4:** 609-618.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Fiegler, H., R. Redon, D. Andrews, C. Scott, R. Andrews, C. Carder, R. Clark, O. Dovey, P. Ellis, L. Feuk, L. French, P. Hunt, D. Kalaitzopoulos, J. Larkin, L. Montgomery, G.H. Perry, B.W. Plumb, K. Porter, R.E. Rigby, D. Rigler, A. Valsesia, C. Langford, S.J. Humphray, S.W. Scherer, C. Lee, M.E. Hurles, and N.P. Carter. 2006. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res* **16:** 1566-1574.

Frazer, K.A. D.G. Ballinger D.R. Cox D.A. Hinds L.L. Stuve R.A. Gibbs J.W. Belmont A. Boudreau P. Hardenbol S.M. Leal S. Pasternak D.A. Wheeler T.D. Willis F. Yu H. Yang C. Zeng Y. Gao H. Hu W. Hu C. Li W. Lin S. Liu H. Pan X. Tang J. Wang W. Wang J. Yu B. Zhang Q. Zhang H. Zhao H. Zhao J. Zhou S.B. Gabriel R. Barry B. Blumenstiel A. Camargo M. Defelice M. Faggart M. Goyette S. Gupta J. Moore H. Nguyen R.C. Onofrio M. Parkin J. Roy E. Stahl E. Winchester L. Ziaugra D. Altshuler Y. Shen Z. Yao W. Huang X. Chu Y. He L. Jin Y. Liu Y. Shen W. Sun H. Wang Y. Wang Y. Wang X. Xiong L. Xu M.M. Waye S.K. Tsui H. Xue J.T. Wong L.M. Galver J.B. Fan K. Gunderson S.S. Murray A.R. Oliphant M.S. Chee A. Montpetit F. Chagnon V. Ferretti M. Leboeuf J.F. Olivier M.S. Phillips S. Roumy C. Sallee A. Verner T.J. Hudson P.Y. Kwok D. Cai D.C. Koboldt R.D. Miller L. Pawlikowska P. Taillon-Miller M. Xiao L.C. Tsui W. Mak Y.Q. Song P.K. Tam Y. Nakamura T. Kawaguchi T. Kitamoto T. Morizono A. Nagashima Y. Ohnishi A. Sekine T. Tanaka T. Tsunoda P. Deloukas C.P. Bird M. Delgado E.T. Dermitzakis R. Gwilliam S. Hunt J. Morrison D. Powell B.E. Stranger P. Whittaker D.R. Bentley M.J. Daly P.I. de Bakker J. Barrett Y.R. Chretien J. Maller S. McCarroll N. Patterson I. Pe'er A. Price S. Purcell D.J. Richter P. Sabeti R. Saxena S.F. Schaffner P.C. Sham P. Varilly D. Altshuler L.D. Stein L. Krishnan A.V. Smith M.K. Tello-Ruiz G.A. Thorisson A. Chakravarti P.E. Chen D.J. Cutler C.S. Kashuk S. Lin G.R. Abecasis W. Guan Y. Li H.M. Munro Z.S. Qin D.J. Thomas G. McVean A. Auton L. Bottolo N. Cardin S. Eyheramendy C. Freeman J. Marchini S. Myers C. Spencer M. Stephens P. Donnelly L.R. Cardon G. Clarke D.M. Evans A.P. Morris B.S. Weir T. Tsunoda J.C. Mullikin S.T. Sherry M. Feolo A. Skol H. Zhang C. Zeng H. Zhao I. Matsuda Y. Fukushima D.R. Macer E. Suda C.N. Rotimi C.A. Adebamowo I. Ajayi T. Aniagwu P.A. Marshall C. Nkwodimmah C.D. Royal M.F. Leppert M. Dixon A. Peiffer R. Qiu A. Kent K. Kato N. Niikawa I.F. Adewole B.M. Knoppers M.W. Foster E.W. Clayton J. Watkin R.A. Gibbs J.W. Belmont D. Muzny L. Nazareth E. Sodergren G.M. Weinstock D.A. Wheeler I. Yakub S.B. Gabriel R.C. Onofrio D.J. Richter L. Ziaugra B.W. Birren M.J. Daly D. Altshuler R.K. Wilson L.L. Fulton J. Rogers J. Burton N.P. Carter C.M. Clee M. Griffiths M.C. Jones K. McLay R.W. Plumb M.T. Ross S.K. Sims D.L. Willey Z. Chen H. Han L. Kang M. Godbout J.C. Wallenburg P. L'Archeveque G. Bellemare K. Saeki H. Wang D. An H. Fu Q. Li Z. Wang R. Wang A.L. Holden L.D. Brooks J.E. McEwen M.S. Guyer V.O. Wang J.L. Peterson M. Shi J. Spiegel L.M. Sung L.F. Zacharia F.S. Collins K. Kennedy R. Jamieson and J. Stewart. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851-861.

Freedman, M.L., K.L. Penney, D.O. Stram, L. Le Marchand, J.N. Hirschhorn, L.N. Kolonel, D. Altshuler, B.E. Henderson, and C.A. Haiman. 2004. Common variation in BRCA2 and breast cancer risk: a haplotype-based analysis in the Multiethnic Cohort. *Hum Mol Genet* **13:** 2431-2441.

Fu, Y.X. 1995. Statistical properties of segregating sites. *Theor Popul Biol* **48:** 172-197.

Gibbs, R.A. 2003. The International HapMap Project. *Nature* **426:** 789-796.

Girard, A., R. Sachidanandam, G.J. Hannon, and M.A. Carmell. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199-202.

Goldstein, D.B. and M.E. Weale. 2001. Population genomics: linkage disequilibrium holds the key. *Curr Biol* **11:** R576-579.

Gordon, D., C. Abajian, and P. Green. 1998a. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Gordon, D., C. Abajian, and P. Green. 1998b. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Harris, T.W., N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, J. Chan, C.K. Chen, W.J. Chen, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E.M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P.W. Sternberg, and L.D. Stein. 2004. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res* **32:** D411-417.

Harris, T.W., R. Lee, E. Schwarz, K. Bradnam, D. Lawson, W. Chen, D. Blasier, E. Kenny, F. Cunningham, R. Kishore, J. Chan, H.M. Muller, A. Petcherski, G. Thorisson, A. Day, T. Bieri, A. Rogers, C.K. Chen, J. Spieth, P. Sternberg, R. Durbin, and L.D. Stein. 2003. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res* **31:** 133-137.

He, L., X. He, S.W. Lowe, and G.J. Hannon. 2007. microRNAs join the p53 network--another piece in the tumour-suppression puzzle. *Nat Rev Cancer* **7:** 819-822.

Hillier, L.W., G.T. Marth, A.R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J.I. Glasscock, M. Hickenbotham, W. Huang, V.J. Magrini, R.J. Richt, S.N. Sander, D.A. Stewart, M. Stromberg, E.F. Tsung, T. Wylie, T. Schedl, R.K. Wilson, and E.R. Mardis. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods*.

Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072-1079.

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon, and W.R. McCombie. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39:** 1522-1527.

Hoskins, R.A., A.C. Phan, M. Naeemuddin, F.A. Mapa, D.A. Ruddy, J.J. Ryan, L.M. Young, T. Wells, C. Kopczynski, and M.C. Ellis. 2001. Single nucleotide polymorphism markers for genetic mapping in Drosophila melanogaster. *Genome Res* **11:** 1100-1113.

Huse, S.M., J.A. Huber, H.G. Morrison, M.L. Sogin, and D.M. Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8:** R143.

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

Ibarra, I., Y. Erlich, S.K. Muthuswamy, R. Sachidanandam, and G.J. Hannon. 2007. A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. *Genes Dev* **21:** 3238-3243.

Ikegawa, S., A. Mabuchi, M. Ogawa, and T. Ikeda. 2002. Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? *Hum Genet* **110:** 606-608.

Jeffries, T.W., I.V. Grigoriev, J. Grimwood, J.M. Laplaza, A. Aerts, A. Salamov, J. Schmutz, E. Lindquist, P. Dehal, H. Shapiro, Y.S. Jin, V. Passoth, and P.M. Richardson. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. *Nat Biotechnol* **25:** 319-326.

Ke, X., S. Hunt, W. Tapper, R. Lawrence, G. Stavrides, J. Ghori, P. Whittaker, A. Collins, A.P. Morris, D. Bentley, L.R. Cardon, and P. Deloukas. 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* **13:** 577-588.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kim, J., K. Inoue, J. Ishii, W.B. Vanti, S.V. Voronov, E. Murchison, G. Hannon, and A. Abeliovich. 2007. A MicroRNA feedback circuit in midbrain dopamine neurons. *Science* **317:** 1220-1224.

Kim, M., H.R. Jang, J.H. Kim, S.M. Noh, K.S. Song, J.S. Cho, H.Y. Jeong, J.C. Norman, P.T. Caswell, G.H. Kang, S.Y. Kim, H.S. Yoo, and Y.S. Kim. 2008. Epigenetic inactivation of Protein Kinase D1 in gastric cancer and its role in gastric cancer cell migration and invasion. *Carcinogenesis*.

Koch, R., H.G. van Luenen, M. van der Horst, K.L. Thijssen, and R.H. Plasterk. 2000. Single nucleotide polymorphisms in wild isolates of Caenorhabditis elegans. *Genome Res* **10:** 1690-1696.

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420-426.

Kwok, P.Y. and X. Chen. 1998. Detection of single nucleotide variations. *Genet Eng (N Y)* **20:** 125-134.

Kwok, P.Y. and Z. Gu. 1999. Single nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today* **5:** 538-543.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh R. Funke D. Gage K. Harris A. Heaford J. Howland L. Kann J. Lehoczky R. LeVine P. McEwan K. McKernan J. Meldrim J.P. Mesirov C. Miranda W. Morris J. Naylor C. Raymond M. Rosetti R. Santos A. Sheridan C. Sougnez N. Stange-Thomann N. Stojanovic A. Subramanian D. Wyman J. Rogers J. Sulston R. Ainscough S. Beck D. Bentley J. Burton C. Clee N. Carter A. Coulson R. Deadman P. Deloukas A. Dunham I. Dunham R. Durbin L. French D. Grafham S. Gregory T. Hubbard S. Humphray A. Hunt M. Jones C. Lloyd A. McMurray L. Matthews S. Mercer S. Milne J.C. Mullikin A. Mungall R. Plumb M. Ross R. Shownkeen S. Sims R.H. Waterston R.K. Wilson L.W. Hillier J.D. McPherson M.A. Marra E.R. Mardis L.A. Fulton A.T. Chinwalla K.H. Pepin W.R. Gish S.L. Chissoe M.C. Wendl K.D. Delehaunty T.L. Miner A. Delehaunty J.B. Kramer L.L. Cook R.S. Fulton D.L. Johnson P.J. Minx S.W. Clifton T. Hawkins E. Branscomb P. Predki P. Richardson S. Wenning T. Slezak N. Doggett J.F. Cheng A. Olsen S. Lucas C. Elkin E. Uberbacher M. Frazier R.A. Gibbs D.M. Muzny S.E. Scherer J.B. Bouck E.J. Sodergren K.C. Worley C.M. Rives J.H. Gorrell M.L. Metzker S.L.

Naylor R.S. Kucherlapati D.L. Nelson G.M. Weinstock Y. Sakaki A. Fujiyama M. Hattori T. Yada A. Toyoda T. Itoh C. Kawagoe H. Watanabe Y. Totoki T. Taylor J. Weissenbach R. Heilig W. Saurin F. Artiguenave P. Brottier T. Bruls E. Pelletier C. Robert P. Wincker D.R. Smith L. Doucette-Stamm M. Rubenfield K. Weinstock H.M. Lee J. Dubois A. Rosenthal M. Platzer G. Nyakatura S. Taudien A. Rump H. Yang J. Yu J. Wang G. Huang J. Gu L. Hood L. Rowen A. Madan S. Qin R.W. Davis N.A. Federspiel A.P. Abola M.J. Proctor R.M. Myers J. Schmutz M. Dickson J. Grimwood D.R. Cox M.V. Olson R. Kaul N. Shimizu K. Kawasaki S. Minoshima G.A. Evans M. Athanasiou R. Schultz B.A. Roe F. Chen H. Pan J. Ramser H. Lehrach R. Reinhardt W.R. McCombie M. de la Bastide N. Dedhia H. Blocker K. Hornischer G. Nordsiek R. Agarwala L. Aravind J.A. Bailey A. Bateman S. Batzoglou E. Birney P. Bork D.G. Brown C.B. Burge L. Cerutti H.C. Chen D. Church M. Clamp R.R. Copley T. Doerks S.R. Eddy E.E. Eichler T.S. Furey J. Galagan J.G. Gilbert C. Harmon Y. Hayashizaki D. Haussler H. Hermjakob K. Hokamp W. Jang L.S. Johnson T.A. Jones S. Kasif A. Kaspryzk S. Kennedy W.J. Kent P. Kitts E.V. Koonin I. Korf D. Kulp D. Lancet T.M. Lowe A. McLysaght T. Mikkelsen J.V. Moran N. Mulder V.J. Pollara C.P. Ponting G. Schuler J. Schultz G. Slater A.F. Smit E. Stupka J. Szustakowki D. Thierry-Mieg J. Thierry-Mieg L. Wagner J. Wallis R. Wheeler A. Williams Y.I. Wolf K.H. Wolfe S.P. Yang R.F. Yeh F. Collins M.S. Guyer J. Peterson A. Felsenfeld K.A. Wetterstrand A. Patrinos and M.J. Morgan. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2:** 231-239.

Levy, S., G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov, Y. Lin, J.R. MacDonald, A.W. Pang, M. Shago, T.B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S.A. Kravitz, D.A. Busam, K.Y. Beeson, T.C. McIntosh, K.A. Remington, J.F. Abril, J. Gill, J. Borman, Y.H. Rogers, M.E. Frazier, S.W. Scherer, R.L. Strausberg, and J.C. Venter. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254.

Mackelprang, R., R.J. Livingston, M.A. Eberle, C.S. Carlson, Q. Yi, J.M. Akey, and D.A. Nickerson. 2006. Sequence diversity, natural selection and linkage disequilibrium in the human T cell receptor alpha/delta locus. *Hum Genet* **119:** 255-266.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Marnellos, G. 2003. High-throughput SNP analysis for genetic association studies. *Curr Opin Drug Discov Devel* **6:** 317-321.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala, D. Church, S. Wheelan, J. Baker, M. Ward, M. Kholodov, L. Phan, E. Czabarka, J. Murvai, D. Cutler, S. Wooding, A. Rogers, A. Chakravarti, H.C. Harpending, P.Y. Kwok, and S.T. Sherry. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci U S A* **100:** 376-381.

Marth, G., R. Yeh, M. Minton, R. Donaldson, Q. Li, S. Duan, R. Davenport, R.D. Miller, and P.Y. Kwok. 2001. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* **27:** 371-372.

Marth, G.T., I. Korf, M.D. Yandell, R.T. Yeh, Z. Gu, H. Zakeri, N.O. Stitziel, L. Hillier, P.Y. Kwok, and W.R. Gish. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat Genet* **23:** 452-456.

Matsuzaki, H., S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G.C. Kennedy, T.A. Webster, S. Cawley, P.S. Walsh, K.W. Jones, S.P. Fodor, and R. Mei. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1:** 109-111.

Mikkelsen, T.S., M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E.S. Lander, and B.E. Bernstein. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*.

Mullikin, J.C., S.E. Hunt, C.G. Cole, B.J. Mortimore, C.M. Rice, J. Burton, L.H. Matthews, R. Pavitt, R.W. Plumb, S.K. Sims, R.M. Ainscough, J. Attwood, J.M. Bailey, K. Barlow, R.M. Bruskiewich, P.N. Butcher, N.P. Carter, Y. Chen, C.M. Clee, P.C. Coggill, J. Davies, R.M. Davies, E. Dawson, M.D. Francis, A.A. Joy, R.G. Lamble, C.F. Langford, J. Macarthy, V. Mall, A. Moreland, E.K. Overton-Larty, M.T. Ross, L.C. Smith, C.A. Steward, J.E. Sulston, E.J. Tinsley, K.J. Turney, D.L. Willey, G.D. Wilson, A.A. McMurray, I. Dunham, J. Rogers, and D.R. Bentley. 2000. An SNP map of human chromosome 22. *Nature* **407:** 516-520.

Murrell, A., V.K. Rakyan, and S. Beck. 2005. From genome to epigenome. *Hum Mol Genet* **14 Spec No 1:** R3-R10.

Ng, P., J.J. Tan, H.S. Ooi, Y.L. Lee, K.P. Chiu, M.J. Fullwood, K.G. Srinivasan, C. Perbost, L. Du, W.K. Sung, C.L. Wei, and Y. Ruan. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res* **34:** e84.

O'Donald, P. 1967. The evolution of selective advantage in a deleterious mutation. *Genetics* **56:** 399-404.

Owen-Hughes, T. and M. Engeholm. 2007. Pyrosequencing positions nucleosomes precisely. *Genome Biol* **8:** 217.

Packard, C.J., R.G. Westendorp, D.J. Stott, M.J. Caslake, H.M. Murray, J. Shepherd, G.J. Blauw, M.B. Murphy, E.L. Bollen, B.M. Buckley, S.M. Cobbe, I. Ford, A. Gaw, M. Hyland, J.W. Jukema, A.M. Kamper, P.W. Macfarlane, J. Jolles, I.J. Perry, B.J. Sweeney, and C. Twomey. 2007. Association between apolipoprotein E4 and cognitive decline in elderly adults. *J Am Geriatr Soc* **55:** 1777-1785.

Petrov, D.A. and D.L. Hartl. 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc Natl Acad Sci U S A* **96:** 1475-1479.

Quinlan, A.R. and G.T. Marth. 2007. Primer-site SNPs mask mutations. *Nat Methods* **4:** 192.

Quinlan, A.R., D.A. Stewart, M.P. Stromberg, and G.T. Marth. 2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods*.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444-454.

Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Richly, E. and D. Leister. 2004. NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol* **21:** 1081-1084.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242:** 84-89.

Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers, D.R. Bentley, P.Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, and D. Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-933.

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

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.H. Lee, J. Hicks, S.J. Spence, A.T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P.K. Gregersen, J. Bregman, J.S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.C. King, D. Skuse, D.H. Geschwind, T.C. Gilliam, K. Ye, and M. Wigler. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445-449.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T.C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525-528.

Shendure, J., R.D. Mitra, C. Varma, and G.M. Church. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5:** 335-344.

Sherry, S.T., M. Ward, and K. Sirotkin. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9:** 677-679.

Shi, N.Q., B. Davis, F. Sherman, J. Cruz, and T.W. Jeffries. 1999. Disruption of the cytochrome c gene in xylose-utilizing yeast Pichia stipitis leads to higher ethanol production. *Yeast* **15:** 1021-1030.

Sjoblom, T., S. Jones, L.D. Wood, D.W. Parsons, J. Lin, T.D. Barber, D. Mandelker, R.J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S.D. Markowitz, J. Willis, D. Dawson, J.K. Willson, A.F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B.H.

Park, K.E. Bachman, N. Papadopoulos, B. Vogelstein, K.W. Kinzler, and V.E. Velculescu. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268-274.

Smit, A.F.A.G., P. REPEATMASKER.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.

Song, J.H., C.J. Kim, Y.G. Cho, H.J. Kwak, S.W. Nam, N.J. Yoo, J.Y. Lee, and W.S. Park. 2007. Genetic and epigenetic analysis of the EPHB2 gene in gastric cancers. *Apmis* **115:** 164-168.

Stein, L.D., Z. Bao, D. Blasiar, T. Blumenthal, M.R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D'Eustachio, D.H. Fitch, L.A. Fulton, R.E. Fulton, S. Griffiths-Jones, T.W. Harris, L.W. Hillier, R. Kamath, P.E. Kuwabara, E.R. Mardis, M.A. Marra, T.L. Miner, P. Minx, J.C. Mullikin, R.W. Plumb, J. Rogers, J.E. Schein, M. Sohrmann, J. Spieth, J.E. Stajich, C. Wei, D. Willey, R.K. Wilson, R. Durbin, and R.H. Waterston. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol* **1:** E45.

Stephens, M. and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73:** 1162-1169. Epub 2003 Oct 1120.

Stephens, M., J.S. Sloan, P.D. Robertson, P. Scheet, and D.A. Nickerson. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet* **38:** 375-381.

Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68:** 978-989. Epub 2001 Mar 2009.

Suh, S.O., C.J. Marshall, J.V. McHugh, and M. Blackwell. 2003. Wood ingestion by passalid beetles in the presence of xylose-fermenting gut yeasts. *Mol Ecol* **12:** 3137-3145.

Sunyaev, S.R., W.C. Lathe, 3rd, V.E. Ramensky, and P. Bork. 2000. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet* **16:** 335-337.

Tajima, F. 1989. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* **123:** 229-240.

The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282:** 2012-2018.

The ENCODE (ENCyclopedia Of DNA Elements) Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636-640.

The ENCODE Project Consortium. 2003. The International HapMap Project. *Nature* **426:** 789-796.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789-796.

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

Thomas, R.K., E. Nickerson, J.F. Simons, P.A. Janne, T. Tengs, Y. Yuza, L.A. Garraway, T. LaFramboise, J.C. Lee, K. Shah, K. O'Neill, H. Sasaki, N. Lindeman, K.K. Wong, A.M. Borras, E.J. Gutmann, K.H. Dragnev, R. DeBiasi, T.H. Chen, K.A. Glatt, H. Greulich, B. Desany, C.K. Lubeski, W. Brockman, P. Alvarez, S.K. Hutchison, J.H. Leamon, M.T. Ronan, G.S. Turenchalk, M. Egholm, W.R. Sellers, J.M. Rothberg, and M. Meyerson. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* **12:** 852-855.

Velicer, G.J., G. Raddatz, H. Keller, S. Deiss, C. Lanz, I. Dinkelacker, and S.C. Schuster. 2006. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc Natl Acad Sci U S A* **103:** 8107-8112.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt J.D. Gocayne P. Amanatides R.M. Ballew D.H. Huson J.R. Wortman Q. Zhang C.D. Kodira X.H. Zheng L. Chen M. Skupski G. Subramanian P.D. Thomas J. Zhang G.L. Gabor Miklos C. Nelson S. Broder A.G. Clark J. Nadeau V.A. McKusick N. Zinder A.J. Levine R.J. Roberts M. Simon C. Slayman M. Hunkapiller R. Bolanos A. Delcher I. Dew D. Fasulo M. Flanigan L. Florea A. Halpern S. Hannenhalli S. Kravitz S. Levy C. Mobarry K. Reinert K. Remington J. Abu-Threideh E. Beasley K. Biddick V. Bonazzi R. Brandon M. Cargill I. Chandramouliswaran R. Charlab K. Chaturvedi Z. Deng V. Di Francesco P. Dunn K. Eilbeck C. Evangelista A.E. Gabrielian W. Gan W. Ge F. Gong Z. Gu P. Guan T.J. Heiman M.E. Higgins R.R. Ji Z. Ke K.A. Ketchum Z. Lai Y. Lei Z. Li J. Li Y. Liang X. Lin F. Lu G.V. Merkulov N. Milshina H.M. Moore A.K. Naik V.A. Narayan B. Neelam D. Nusskern D.B. Rusch S. Salzberg W. Shao B. Shue J. Sun Z. Wang A. Wang X. Wang J. Wang M. Wei R. Wides C. Xiao C. Yan A. Yao J. Ye M. Zhan W. Zhang H. Zhang Q. Zhao L. Zheng F. Zhong W. Zhong S. Zhu S. Zhao D. Gilbert S. Baumhueter G. Spier C. Carter A. Cravchik T. Woodage F. Ali H. An A. Awe D. Baldwin H. Baden M. Barnstead I. Barrow K. Beeson D. Busam A. Carver A. Center M.L. Cheng L. Curry S. Danaher L. Davenport R. Desilets S. Dietz K. Dodson L. Doup S. Ferriera N. Garg A. Gluecksmann B. Hart J. Haynes C. Haynes C. Heiner S. Hladun D. Hostin J. Houck T. Howland C. Ibegwam J. Johnson F. Kalush L. Kline S. Koduru A. Love F. Mann D. May S. McCawley T. McIntosh I. McMullen M. Moy L. Moy B. Murphy K. Nelson C. Pfannkoch E. Pratts V. Puri H. Qureshi M. Reardon R. Rodriguez Y.H. Rogers D. Romblad B. Ruhfel R. Scott C. Sitter M. Smallwood E. Stewart R. Strong E. Suh R. Thomas N.N. Tint S. Tse C. Vech G. Wang J. Wetter S. Williams M. Williams S. Windsor E. Winn-Deen K. Wolfe J. Zaveri K. Zaveri J.F. Abril R. Guigo M.J. Campbell K.V. Sjolander B. Karlak A. Kejariwal H. Mi B. Lazareva T. Hatton A. Narechania K. Diemer A. Muruganujan N. Guo S. Sato V. Bafna S. Istrail R. Lippert R. Schwartz B. Walenz S. Yooseph D. Allen A. Basu J. Baxendale L. Blick M. Caminha J. Carnes-Stine P. Caulk Y.H. Chiang M. Coyne C. Dahlke A. Mays M. Dombroski M. Donnelly D. Ely S. Esparham C. Fosler H. Gire S. Glanowski K. Glasser A. Glodek M. Gorokhov K. Graham B. Gropman M. Harris J. Heil S. Henderson J. Hoover D. Jennings C. Jordan J. Jordan J. Kasha L. Kagan C. Kraft A. Levitsky M. Lewis X. Liu J. Lopez D. Ma W. Majoros J. McDaniel S. Murphy M. Newman T. Nguyen N. Nguyen M. Nodell S. Pan J. Peck M. Peterson W. Rowe R. Sanders J. Scott M. Simpson T. Smith A. Sprague T. Stockwell R. Turner E. Venter M. Wang M. Wen D. Wu M. Wu A. Xia A. Zandieh and X. Zhu. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Warren, R.L., G.G. Sutton, S.J. Jones, and R.A. Holt. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23:** 500-501.

Weber, J.L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7:** 524-530.

Weber, J.L., D. David, J. Heil, Y. Fan, C. Zhao, and G. Marth. 2002. Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* **71:** 854-862.

Whiteford, N., N. Haslam, G. Weber, A. Prugel-Bennett, J.W. Essex, P.L. Roach, M. Bradley, and C. Neylon. 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33:** e171.

Yoo, C.B. and P.A. Jones. 2006. Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov* **5:** 37-50.

Zhang, J., D.A. Wheeler, I. Yakub, S. Wei, R. Sood, W. Rowe, P.P. Liu, R.A. Gibbs, and K.H. Buetow. 2005. SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. *PLoS Comput Biol* **1:** e53.

Zhang, Z. and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31:** 5338-5348.

Zhao, Z. and E. Boerwinkle. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12:** 1679-1686.