

# Chapter 5

## Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery

Eugene Y. Chan

### Abstract

The advent of next-generation sequencing technologies has spurred remarkable progress in the field of genomics. Whereas traditional Sanger sequencing has yielded the first complete human genome sequence, next-generation methods have been able to resequence several human genomes. In this manner, next-generation approaches have powerful capabilities for understanding human variation. The throughput for these approaches is often measured in billions of base pairs per run, astounding numbers when compared with the millions of base pairs per day generated by automated capillary DNA sequencers. However, unlike traditional Sanger dideoxy sequencing, these methods have lower accuracy and shorter read lengths than the dideoxy gold standard. Are these limitations offset by the higher throughputs? An in-depth look at the single read and composite accuracy of these methods is presented. The stringent requirements for single nucleotide polymorphism (SNP) discovery utilizing these approaches is discussed along with a review of studies that have successfully employed next-generation sequencing methods for large-scale SNP discovery. Ultimately, the application of these ultra-high-throughput sequencing methods for SNP discovery will open up new horizons for understanding human genomic variation.

**Key words:** Next-generation sequencing, sequencing technology, single nucleotide discovery, accuracy, Sanger sequencing, polony sequencing, pyrosequencing, cycle extension, 454, Illumina, Helicos, SOLiD.

---

### 1. Introduction

Next-generation sequencing methods are those that employ massively parallel approaches in sequencing several hundred thousand to millions of reads simultaneously. This is an enormous increase in the number of reads compared with existing capillary sequencers that employ 96-capillary arrays. Currently, the four commercialized next-generation sequencing methods are parallel pyrosequencing in

picoliter reactors (1), Illumina sequencing (2–4), SOLiD or polony sequencing (5), and single molecule sequencing (6). These methods have enormous breadth in applications, including new applications such as ChIP-seq (7), transcript sequencing (8), microRNA discovery (9, 10), whole genome resequencing (11), and single nucleotide polymorphism (SNP) discovery (2, 11, 12). The use of these methods for SNP discovery is by far the most demanding application since it requires a high accuracy. In contrast to ChIP-seq or transcript sequencing applications, which rely on sequence identification for various reads, SNP discovery requires identification of individual base pairs and their differences. Any error in sequencing leads to incorrect SNP identification since SNPs occur at a high frequency, at every few hundred to thousand base pairs (11, 13, 14). Furthermore, the abundance of other variations in the human genome, including copy number variations (15), insertions/deletions (indels), and rearrangements, makes the task even more challenging. The shorter read lengths of next-generation methods and the abundance of repeated sequences in the human genome create unique issues in being able to accurately cover all parts of the genome for SNP discovery. Nonetheless, some next-generation sequencing methods have been able to bypass these challenges and have been able to add many new SNPs to dbSNP (*see Note 1*) (16). The following describes the accuracy, error, and coverage of next-generation sequencing methods as required for SNP discovery.

---

## 2. Sequencing Accuracy

What is the required accuracy for SNP discovery? Another way to think of the question is to understand how many errors can be tolerated in a given stretch of DNA. Since single base changes can give rise to significant phenotypic change, errors cannot be missed. One error in 100,000 bp would yield an error rate of 99.999%, a value set by the Archon Genomics X Prize (*see Note 2*) (17). Shendure et al. (5), in their polony paper, suggested an even more stringent number, a consensus error rate of one in 1,000,000 bp, or an error rate 99.9999%. Any error in sequencing can give rise to false positives or false negatives, leading to challenging downstream studies.

The accuracies of the next-generation sequencing methods are compared with the accuracy of Sanger dideoxy sequencing. The universally accepted accuracy metric for Sanger dideoxy sequencing is given in terms of Phred values (18, 19). Since each sequencing method has its own quality scores, the raw accuracy of

each sequencing method needs to be presented in a standard manner. Each method's performance relative to different errors, including in relation to homopolymers, base substitutions, and deletion–insertion polymorphisms, needs to be assessed. The massively parallel capabilities of next-generation methods allows for vast oversampling to correct for errors. Their short read lengths, however, lead to some parts of the genome being underrepresented. The inability to cover certain parts of the genome leads to large unanalyzed stretches of SNPs. The degree of this problem is explored for each of the methods, as is the redundancy required for calling SNPs in genome-scale efforts.

---

### 3. Raw Sequencing Accuracy

The accuracy of a single-pass sequencing read is its raw sequencing accuracy. Any accuracy figure that is a result of a composite or requires averaging is not a raw read. Raw accuracy allows the methods to be compared directly with Sanger dideoxy sequencing. It is an excellent measurement of each method's chemistry, fluorescence readout, process, and base-calling software. Ultimately, the more robust the raw reads, the fewer the required redundant reads. The limit to raw accuracy is the fidelity attained by nature's polymerases, which is on the order of  $10^{-5}$ – $10^{-7}$  per base pair for commercially available polymerases (20, 21). Polymerases with  $3' \rightarrow 5'$  exonuclease activity have fidelity values closer to  $10^{-7}$  errors per base pair. Given that all existing commercially available methods utilize a polymerase for readout or amplification, it is reasonable to say that this represents the ideal sequencing raw accuracy.

Sanger sequencing, as performed by existing capillary sequencers with fluorescent dye terminators, sets a high bar for raw accuracy. All base pairs that are accepted as reads have an accuracy greater than Phred Q20, or 99% accuracy (18, 19). For instance, the Applied Biosystems 3730xl in its long read mode can generate more than 800 bp of sequence at or above the Phred Q20 threshold (22). In fact the majority of base pairs in any given long read run, especially those between 100 and 700 bp, have Phred scores that approach Q50, or 99.999% (22, 23). Its performance is fairly constant across a broad range of different templates (23). The high confidence in the capillary readout makes the approach robust for sequencing and SNP discovery.

PCR is utilized in Sanger sequencing and the majority of next-generation sequencing methods. It is important to understand the impact of this step on overall sequencing accuracy. After PCR, a population of different molecules are generated, a fraction of

which contain errors. The final fraction of DNA containing at least one PCR-derived mutation, given as  $F$ , determines the significance of the error. For instance, a PCR that has  $F = 0.50$ , with an error propagated at one position early in the reaction, would yield a heterozygote call upon Sanger sequencing. In contrast, the same reaction with  $F = 0.1$  would yield a correct call since 90% of the signal would be from the original wild-type sequence. The relationship between  $F$  and polymerase fidelity ( $f$ ), PCR product size ( $b$ ), and number of DNA doublings ( $d$ ) is given by the equation  $F = 1 - e^{-bfd}$ , as outlined by Keohavong et al. (24). For a 1,000-bp target and 20 doublings, a polymerase with fidelity equal to  $10^{-7}$  would yield  $F \approx 0.2\%$ . For  $10^{-6}$ ,  $F \approx 2.0\%$ ; for  $10^{-5}$ ,  $F \approx 18\%$ ; and for  $10^{-4}$ ,  $F \approx 86\%$ . For high-fidelity polymerases, such as Phusion polymerase, which performs with  $4.4 \times 10^{-7}$  errors per base pair, about 0.9% of the fragments would yield an error. Although the use of a high-fidelity polymerase is advantageous, errors are still present, especially in next-generation sample preparation methods that have two different PCR steps. In these methods, the first PCR step generates linkers on the ends of genomic DNA fragments. A second PCR step then utilizes these linker-amplified molecules to clonally amplify each one of them. Errors in the first PCR step are propagated in the second step. For Phusion polymerase, about 0.9% of the 1,000-bp fragments have an error, corresponding to  $10^{-5}$  errors per base pair. From this calculation, it is clear that sample preparation errors are present when high-fidelity polymerases are used. However, this error is small when compared with readout errors, which can be between  $10^{-1}$  and  $10^{-2}$  for some methods.

**Table 5.1** summarizes the raw accuracy for the four next-generation sequencing methods compared with Sanger sequencing. The values represent data from single reads and exclude those derived from averaging, bidirectional reads, and filtered data. For each method, a range is reported. The range corresponds to raw accuracy

**Table 5.1**  
**Comparison of raw accuracies and tabulation of read lengths**

Method	Raw accuracy range
Sanger	99.0% to > 99.999% (22, 23)
454	96.0% (1, 31) to 97.0% (25)
Illumina	96.2% to 99.7 (28)
SOLiD	99.0% to > 99.9% (22)
Helicos	93.0 to 97.0% (6)

values within a sequencing read. For instance, the accuracy in automated dideoxy sequencing is greater than Phred Q50, or 99.999%, for most of its bases. Dideoxy sequencing accuracy at positions less than 100 bp and greater than 700 bp decreases to Phred Q20, or 99.0%. Errors increase with base position in all the methods. For Sanger sequencing, larger DNA fragments are more difficult to resolve, leading to decreased quality scores at higher positions, and ultimately limiting read length. For the next-generation sequencing methods, errors accumulate because of asynchronous chemistries or imperfect fluorescence detection, as illustrated in **Fig. 5.1** for the Illumina method. For instance, a 98% readout efficiency at each base pair leads to difficult calls with increasing length. At the 35th base pair, the readout for fewer than half the molecules would be correct, since  $0.98^{35} = 0.49$ . Higher readout efficiencies are challenging because, in fact, there are numerous steps in the read out of each base pair. The Helicos and 454 methods require at least eight steps to read out each base pair, including multiple wash and nucleotide addition steps. The large number of steps that need to be optimized makes it difficult to get more accurate results at longer reads. The read length limit for these methods is ultimately dictated by the drop off in base pair quality, which is 250 bp for parallel pyrosequencing, 36 bp for Illumina, 36 bp for SOLiD, and 25 bp for Helicos.

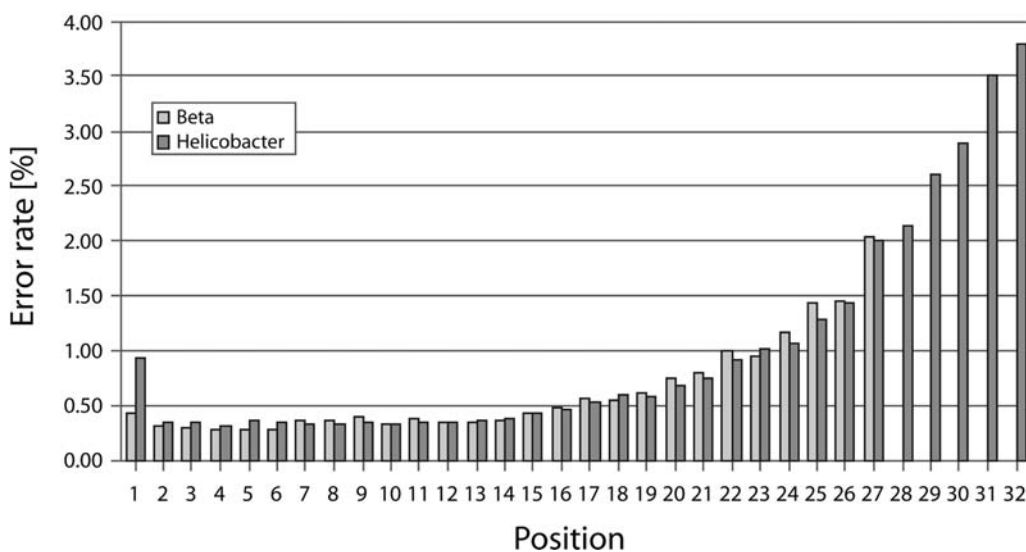


Fig. 5.1. Error rate increases with base pair position for Illumina Genome Analyzer. Data are from 27-mer *Bacillus vulgaris* reads and 32-mer *Helicobacter acinonychis* genome sequencing. (From Dohm et al. (28); reprinted with permission from Oxford University Press).

Both Sanger sequencing and SOLiD sequencing have the highest raw accuracy values, at greater than 99% for all the base pairs, corresponding to a minimum of Phred Q20. The higher accuracy for SOLiD is achieved by redundantly

reading each base pair twice by a two-base encoding scheme. While it can be argued that it is in fact two separate reads for the overlapping two-base reads, the read being performed on one sequence makes it an individual read. It is interesting to note that the uncorrected raw accuracy for SOLiD without two-base encoding is between 90 and 99% (22), which is similar to the single pass error for the Helicos method, which utilizes a second pass to attain accuracies of 0.1–0.3%. In the Helicos two-pass sequencing approach, the inefficiencies of single fluorophores are bypassed with bidirectional reading.

---

#### 4. Homopolymer Errors

A problem unique to next-generation sequencing methods is contiguous runs of the same base pair, called “homopolymer repeats.” For instance, the sequence 5' AAAAA 3' is an A5 homopolymer. Large stretches of homopolymers exist in the human genome. Homopolymer runs greater than 5 bp in length or more comprise more than 4% of the genome, according to human bacterial artificial chromosome data (25). Up to now, the majority of these have been placed in GenBank via Sanger sequencing, which resolves homopolymers readily in a gel matrix. The ability to read these sequences accurately allows more complete coverage of the genome and thus more complete SNP discovery efforts. The inability to read these sequences leads to a biased set of SNPs that may exclude some potentially important information. The extent of this problem among the next-generation sequencing methods is examined.

Both parallel pyrosequencing and Helicos single molecule sequencing have significant difficulties with homopolymer stretches (1, 6). For instance, it has been estimated that 39% of parallel pyrosequencing's errors are from incorrect analysis of homopolymer regions (26). An analysis of error at each homopolymer length is revealing, as documented by Wicker et al. (27). For A5 motifs, the 454 method has a 3.3% error rate. The error rate increases with longer motifs. For A8 motifs, the method has a 50% error rate. For the Helicos method, Harris et al. (6) described being able to call between 45 and 65% of four-mer homopolymers in single pass reads. The method, however, addresses these issues with its two-pass sequencing approach that employs a voting scheme. In this manner, more than 80% of A, G, and T four-mers are called correctly (**Fig. 5.2**).

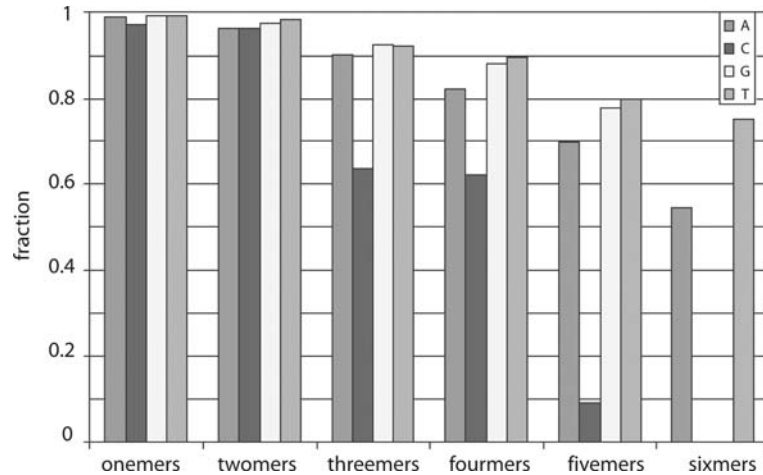


Fig. 5.2. Two-pass homopolymer error rate for the Helicos single molecule sequencing method. The fraction of homopolymers called correctly decreases with homopolymer length. (From Harris et al. (6); reprinted with permission from AAAS).

Homopolymer errors are unique to next-generation sequencing methods that read signal intensity as a measure of homopolymer length. In parallel pyrosequencing (*see Chapter 7* for the details of the method), when a homopolymer stretch is encountered, multiple incorporations of that base occur, releasing a proportionate amount of pyrophosphate that is detected as an intensity signal. The accuracy of the light intensity signal is diminished by asynchronous molecules and by photon shot noise. Photon shot noise follows a Poisson distribution, with the standard deviation equal to the square root of the average photon signal. As longer homopolymers are read, the probability of their Poisson distributions overlapping increases owing to the greater photon shot noise. These two impediments together lead to the errors observed for parallel pyrosequencing. Helicos distinguishes homopolymers on the basis of the inverse of their intensity since adjacent incorporated fluorescent nucleotides have increased quenching. Although the single molecule method does not have the problems of asynchronous reads, it does face a challenge in that it does not read more than a 3-bp homopolymer motif at a time. Longer homopolymers are stitched together in separate read cycles.

In contrast, Illumina and SOLiD address the homopolymer challenge through reversible terminators and two-base encoding. Illumina sequencing utilizes chemistry that caps the 3'-OH of modified fluorescent bases so that upon encountering homopolymer stretches only one nucleotide is incorporated at a time. In this manner, bases are only read out one at a time, allowing homopolymer stretches to be counted positionally as opposed to their being measured via measurement intensity. In a detailed analysis of Illumina's method, Dohm et al. (28) concluded that the method

handled homopolymer stretches without increased errors. Given the recent commercial introduction of the SOLiD system, confirmatory studies of the system's homopolymer accuracy have yet to be performed. Conceptually, and according to the manufacturer (22), the approach should not have any issues. The SOLiD system's chemistry positionally interrogates two bases at each cycle so there is no ambiguity about homopolymer length during a particular run.

---

## 5. Base Substitution Errors and Indels

Knowing the specific type of error introduced by the methods allows selection of the appropriate next-generation sequencing method for a given application. For SNP discovery, the method's accuracy with respect to base substitution errors is the most important. For discovery of deletion–insertion polymorphisms, microsatellites, and other length-based mutations, accuracy with respect to indels is paramount. The various different methods perform differently with respect to types of errors generated.

Illumina derives the majority of its errors from base substitutions. Dohm et al. (28) studied the accuracy of the method across 2.8 million 27-mer reads from *Bacillus vulgaris* and 12.3 million 36-mer reads from *Helicobacter acinonychis*. Base substitutions accounted for more than half of the errors. In particular, these were the transversions  $A \rightarrow C$ ,  $G \rightarrow T$ , and  $A \rightarrow T$ . Two possible explanations for these base substitution biases are possible. First, the particular polymerase and chemistry utilized by the method could favor the incorporation of pyrimidines in these contexts. Second, as suggested by Dohm et al. (28), there is insufficient spectral discrimination between the fluorophores, leading to bleed-through and miscalled bases. On the other hand, indel errors were infrequent for Illumina sequencing, occurring at a rate of less than 0.01% (28). On the rare chance that a base insertion occurred, it was most likely to be correlated with homopolymer runs of greater than four nucleotides. The addition of a base at a time with reversible terminator chemistry contributes to the accuracy of the method with respect to indels.

For parallel pyrosequencing and the Helicos method, the predominant errors are indel errors that arise from incorrect knowledge about readout position. For 454, 63% of total errors were indels, 36% from insertions and 27% from deletions (26). The high rate of indel errors, when compared with the 39% of homopolymer errors (26), suggests that indel errors occur anywhere and this is a fundamental challenge for parallel pyrosequencing's readout method. The method did better with mismatches, with 16% of total errors from this category (26). The remaining 21% errors were ambiguous base calls and could have led to indel or base substitution errors. For the Helicos

single molecule sequencing method, the major source of error is deletions, which account for 3–7% of errors on single pass reads. Insertion errors occur at a rate of 0.02–1.1% and base substitution errors at a rate of 0.01–1.0% (6). Of the deletion errors, 3–4% arise from nonemitting nucleotides and undetected events, suggesting that detection efficiency is an area that can be improved through brighter dyes and better chemistries.

SOLiD sequencing appears to perform equally well with respect to both types of errors, although independently verified studies are required to fully assess the capabilities of the method. At present, data are sparse for detailed accuracy analyses. Since the two-base encoding strategy of SOLiD sequencing gives good results with homopolymers, the majority of its errors are most likely to arise from base substitution errors.

---

## 6. Consensus Accuracy, Redundancy and Coverage

The throughput of the next-generation sequencers allows for more reads and thus more opportunities to correct innate errors in raw accuracy. Overall consensus accuracy is the composite accuracy from sequencing a genomic region many times. Even a sequencing method with 90% raw accuracy can attain 99.999% consensus accuracy by providing fivefold redundancy. In practice, attaining sufficient redundancy with short reads provides challenges since half of the human genome is repeat sequences (13, 14). Repeat sequences prevent short reads from being accurately mapped to the genome. Consequently, areas of the genome remain without adequate coverage, excluding potentially important areas of the genome from analysis.

In practice, the actual redundancy required to attain high consensus accuracies (greater than 99.99%) is higher than the theoretically predicated number. For instance, for Illumina sequencing, the required redundancy to attain close to no errors per kilobase pair is 20-fold for bacterial genomes (28). The SOLiD method requires 12-fold sequence redundancy in human germline genome sequencing (22). The Helicos method requires at least 20-fold redundancy for good mutation detection. In their M13 viral genome study, they attained 150-fold redundancies (6). Parallel pyrosequencing requires 7.4-fold redundancy to sequence a diploid human genome (11). For Sanger sequencing, it is commonly accepted that threefold redundancy is sufficient to attain the required accuracies for diploid genomes. It should be noted that these values for parallel pyrosequencing and Sanger sequencing are for diploid genomes, whereas for the others, they are for haploid genomes.

Genome coverage is perhaps the most significant issue for the next-generation sequencing methods. Short read length methods, those with 20–30-bp reads, are at a significant disadvantage in analyzing complex genomes and may leave parts of the genome inadequately covered. Short reads do well with smaller, haploid genomes, as seen from the use of the Helicos method with the M13 viral genome, where it attained 100% coverage (6). With bacterial-sized genomes, coverage for short read next-generation methods starts to decrease. In the sequencing of a strain of *Escherichia coli* with polony technology and a mate-paired library, 30.1 Mb of sequence was generated, with 83.3% of the genome having a twofold or greater redundancy (5). The numbers quickly dropped to 66.9% of the genome with fourfold or greater redundancy (5). The coverage issue for small bacterial genomes can be bypassed by higher throughput. Dohm et al. (28) showed complete coverage utilizing the Illumina platform, but some regions of the genome ended up having more than 350-fold redundancy to ensure full coverage given the GC-rich read biases. Ultimately, the average redundancy required for full bacterial genome coverage will be dictated by sample preparation biases, which are yet to be determined for the SOLiD and Helicos methods. In a whole-genome sequencing effort of *Caenorhabditis elegans* by Hillier et al. (2), the genome was repeat-masked for short repeats. In this manner, approximately 23% of the genome was excluded from analysis (Fig. 5.3).

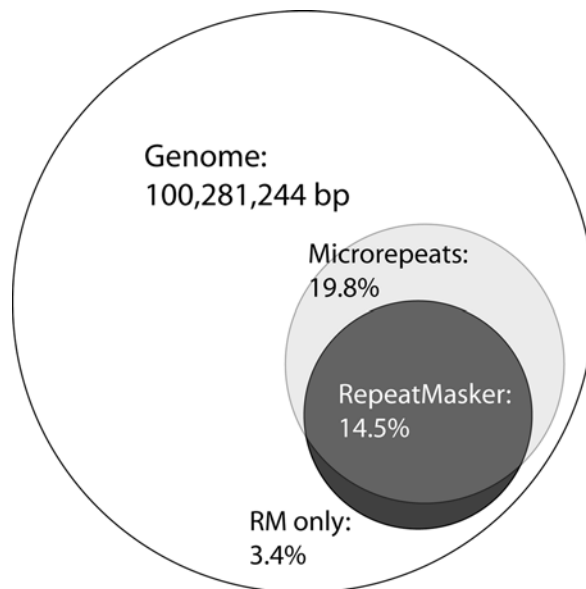


Fig. 5.3. Repetitive content in *Caenorhabditis elegans*. The genome has a large percentage of microrepeats and sequences identified by RepeatMasker. (From Hillier et al. (2); reprinted by permission from Macmillan Publishers Ltd).

Genome coverage is improved with longer read lengths. Parallel pyrosequencing, with average read lengths of 250 bp, sequenced a human genome with average 7.4-fold redundancy (11). Unlike for biased short reads, the method gave fairly uniform coverage of the human genome in a Poisson distribution (Fig. 5.4). Average redundancy was 3.7-fold for the X chromosome. Twenty nine megabases in 110,000 contigs did not match back to the genome, with 65% of these sequences being identified as satellite DNA and the others in repeat-rich regions such as ALU, LINE1, and LINE2 repeat sequences. When compared with the reference genome and excluding centromere sequences, the data showed that more than 98% of chromosome 1 was covered, showing that average read lengths of 250 bp could be utilized for whole human genome resequencing.

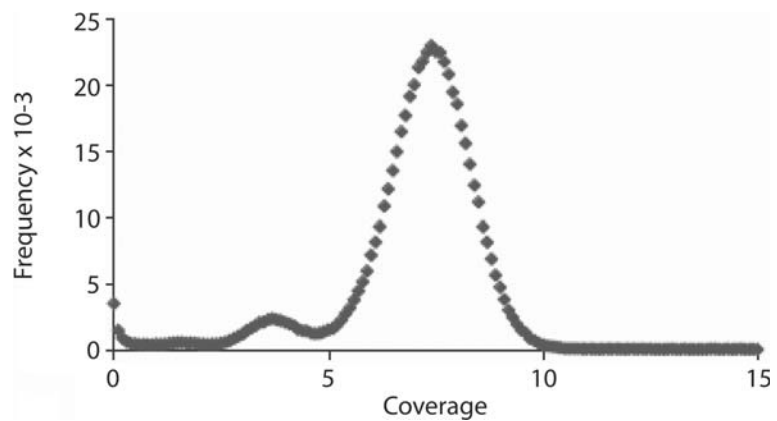


Fig. 5.4. Average human genome sequencing redundancy for parallel pyrosequencing. For all chromosomes, average redundancy was 7.4-fold, and for the X chromosome, it was 3.7-fold. (From Wheeler et al. (11); reprinted by permission from Macmillan Publishers Ltd).

---

## 7. Sensitivity and Specificity of SNP Discovery

The sensitivity and specificity of SNP discovery, particularly compared with those of current accepted technologies, ultimately reflect the performance of each of the next-generation sequencing methods. These are the ultimate metrics since they are the product of each method's inherent error and coverage. A direct comparison with existing SNP analysis technologies allows next-generation methods to be thoroughly evaluated. Existing SNP analysis technologies include dideoxy sequencing, Affymetrix genotyping arrays, and Illumina genotyping bead arrays.

Of the existing next-generation sequencing methods, parallel pyrosequencing has made the greatest strides in being able to accurately document its performance in human SNP discovery

through sequencing of an individual genome (11). In comparing the parallel pyrosequencing method with validated Affymetrix genotyping arrays, the approach showed 99.4% specificity in homozygous reference calls, 95.1% with homozygous variants, and 75.8% with heterozygous calls. In their analysis, the authors concluded that heterozygotes require at least a 13-fold redundancy to accurately call 99% of heterozygous SNPs (Fig. 5.5). On the basis of the method's Poisson distribution, close to 20-fold mean redundancy would need to be attained to ensure that most of the human genome has at least 13-fold redundancy, more than doubling the cost of the \$1 million genome. The specificity was 93.3% for homozygotes and 97.8% for heterozygotes. Utilizing this approach, parallel pyrosequencing was able to identify 3.32 million SNPs, with 606,797 of those as novel SNPs. This compared well with the shotgun-sequenced Venter genome that had 3.47 million SNPs, with 647,767 of those being novel (11).

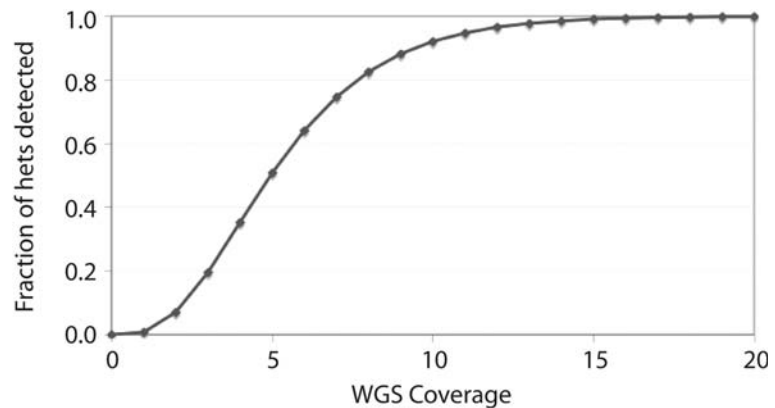


Fig. 5.5. Percentage of heterozygotes called corrected versus redundancy in parallel pyrosequencing. At least 13-fold redundancy is required to correctly call 99% of heterozygotes. (From Wheeler et al. (11); reprinted by permission from Macmillan Publishers Ltd).

The short read methods, having arrived at the marketplace later than parallel pyrosequencing, are showing that they can be effective for SNP discovery. The SOLiD method has been utilized to sequence a Yoruba germline genome and the data are being analyzed to demonstrate its performance in SNP discovery (22). The Illumina method has been utilized for a smaller diploid *C. elegans* genome for the N2 Bristol and CB4858 strains (2). For the published *C. elegans* study, a detailed analysis was performed to assess the capabilities of the platform for SNP discovery. Putative SNPs identified with the approach were verified by a PCR/dideoxy sequencing approach. A SNP validation rate of 93.8% was attained. The conversion rate, or the number of confirmed variants divided by the number of submitted variants, was lower at 87.7%,

presumably owing to difficulties with the PCR/dideoxy validation step. With use of this approach, a total of 45,539 *C. elegans* SNPs were identified across nonrepetitive regions of the genome. Additional studies with these approaches will solidify next-generation sequencing's performance in SNP discovery.

---

## 8. Complete Polymorphism Discovery

Next-generation methods have a potential for broad utility in understanding human variation through SNPs and other polymorphisms, such as copy number variations and indels. Insight into the capabilities of these methods with respect to these polymorphisms are beginning to emerge. To assess copy number variations properly, the sequencing method needs to be able to have unbiased coverage of sequences in the genome. Parallel pyrosequencing has relatively unbiased coverage (except for the X chromosome) and was shown to match comparative genome hybridization studies in 18 of 23 regions in the human genome (11). Deletions were also identified using this method, including sequencing of breakpoint regions. The Illumina method has been shown to have coverage biases based on GC content (2, 28), which would undoubtedly impact copy number variation measurements. Despite this GC bias, Hillier et al. (2) were able to find that there was correlation between the number of ribosomal DNA sequences and the number of 32-mer reads. As for indels, the underlying homopolymer error rate for each method dictates its performance in being able to identify indels, an important emerging class of polymorphisms (29). Parallel pyrosequencing identified over 12.5 million one-base indels in the human genome sequencing data set. A total of 10.4 million of these were associated with homopolymeric runs two to 20 bases in length, leading to considerable ambiguity about the accuracy of these events (11). The Helicos method showed between 92 and 100% success in calling mock indels in the smaller 7.2 M13 genome. From these studies, it is clear that the potential is there for next-generation sequencing methods to play a role in understanding any human variation at its most fundamental level.

---

## 9. Outlook for Next-Generation Sequencing in SNP Discovery

SNP discovery will undoubtedly benefit from the enormous data sets generated by next-generation sequencing technologies. However, there are stringent accuracy requirements that need to be met

for these methods to be effective. Errors with respect to raw data, homopolymer stretches, indels, redundancy, and coverage ultimately dictate the utility of each technology for SNP discovery. Since no technology is error-free, it is important to understand the limitations of each so that effective SNP discovery can be performed.

Next-generation sequencing methods can attain adequate consensus accuracies through adequate depth of coverage for SNP discovery. An average of approximately 20-fold redundancy is required for the 454 and Illumina methods to accurately call SNPs in a diploid genome, which includes calling homozygous and heterozygous SNPs. Throughput and cost ultimately factor into attaining these levels of redundancy. For the 454 method, it would cost more than \$2 million and close to 6 months to attain a 20-fold redundancy. For the Illumina method, sequencing 60 Gb of DNA would cost approximately \$160,000 in reagents at \$4000 per run. This estimate is simplistic because its reads are biased towards GC-rich sequences (28), meaning that a much higher average redundancy would need to be attained to have AT-rich sequences represented at least 20-fold. Furthermore, the method has short read lengths, which predisposes it to limitations when sequencing complex genomes.

The predominate error in short read length methods, defined as those with 20–30-bp reads, is genome coverage. Even for a moderately complex genome such as that of *C. elegans*, about 25% of the genome is excluded from analysis owing to repeats (2). Given that the human genome is about 50% repeats, short read methods would not be able to discover SNPs for a significant portion of the genome. Improvements in raw base and homopolymer accuracy would only allow those regions accessible to short reads to be analyzed. The only way to improve upon genome coverage is to attain longer reads. Doing so would be challenging given that the limiting factor for these methods is their innate chemistry.

Sample preparation contributes a small amount of error to the overall workflow. The major difference in sample preparation techniques is the use of PCR. In amplification with PCR, errors are propagated, especially when there is more than one PCR step involved. In other words, clonal amplification of a PCR-amplified sample leads to errors. With a high-fidelity polymerase, sample preparation leads to an error rate of  $10^{-5}$  for clonally amplified molecules. Sample preparation with one PCR step, such as for some Sanger sequencing reactions, would have no errors since the predominant molecules in the reaction are the correct sequence. Single molecule approaches, such as the Helicos method, utilize direct reads off single molecules to bypass sample preparation errors.

Homopolymers and indel errors present challenges for SNP discovery, particularly for parallel pyrosequencing and single molecule sequencing. Even with oversampling, the inherent errors in

these methods limit their utility in these genomic motifs. These methods measure homopolymer length via measured intensity, which is prone to errors due to photon shot noise, fluorescence quenching, and asynchronous molecules. Indel errors also arise because of missed calls regarding base incorporations. Deletions are prominent for the Helicos method since single molecules are particularly prone to detection errors. In contrast, the Illumina and SOLiD technologies interrogate each base pair in a positional manner, with the base pair positions always recorded.

It is clear from this accuracy analysis that SNP discovery is the most stringent application for next-generation sequencing methods. The 454 and Illumina technologies have been first in studies demonstrating their use in SNP discovery (2, 11). As for Helicos and SOLiD methods, their use in SNP discovery will be born out with time since they were only recently introduced into the marketplace. An abundance of applications exist for next-generation sequencing methods, most of which require less accuracy. Among these are transcriptome sequencing, ChIP-seq, and microRNA studies. While these are the applications that will enjoy rapid adoption, there is significant interest to fully utilize next-generation sequencing technologies for SNP discovery.

One such effort is the 1,000 Genomes Project, which is the most significant ongoing project to characterize human genomic variation. Next-generation sequencers will play a key role in data generation (30). The project's goal is to create a new map of the human genome that will catalog human variation more comprehensively than existing maps. In addition to an international consortium of scientists, Illumina, 454, and Applied Biosystems will participate by contributing their technologies to the effort. Each company will sequence at least 75 Gb, with Applied Biosystems contributing a total of 275 Gb. The vast amounts of data that will arise from this project will likely lead to the discovery of many SNPs and other important human variations. It will rigorously test the capability of each of these approaches for large-scale SNP discovery. More importantly, the completed map will be a composite of the various approaches, potentially allowing the diversity of errors specific to each approach to be averaged out.

Exciting developments in the field of next-generation sequencing has led to a deeper understanding of variation across individual genomes. These approaches leverage off the successful completion of the Human Genome Project to resequence entire genomes, enabling large-scale SNP discovery on an unprecedented scale. Data from individual genomes have already shown that next-generation sequencers are up to the demands of SNP discovery. As progress is made towards routine human genome resequencing, further improvements in next-generation sequencing raw accuracy, homopolymer reads, and coverage are inevitable and will undoubtedly change how we understand SNPs in the years to come.

## 10. Notes

1. The SNP database (dbSNP) is a public-domain archive for a collection of SNPs provided by the National Center for Biotechnology Information (NCBI). For more information on various SNP databases, *see* **Chapter 3**.
2. The private X-Prize Foundation is offering US \$10 million to the first team of researchers to sequence 100 human genomes in 10 days for less than \$10,000 a genome (17).

## References

1. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
2. Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G. et al. (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183–188.
3. Bennett, S. T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373–382.
4. Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics* **5**, 433–438.
5. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P. et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732.
6. Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J. et al. (2008) Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109.
7. Mardis, E. R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods* **4**, 613–614.
8. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.
9. Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y. et al. (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res.* **10**, 997–1006.
10. Glazov, E. A., Cottee, P. A., Barris, W. C., Moore, R. J., Dalrymple, B. P. et al. (2008) A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome Res.* **18**, 957–964.
11. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L. et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876.
12. Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P. et al. (2008) Genetic variation in an individual human exome. *PLoS Genet.* **4**, e1000160.
13. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
14. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.
15. McCarroll, S. A. and Altshuler, D. M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–42.
16. <http://www.ncbi.nlm.nih.gov/projects/SNP/>
17. <http://genomics.xprize.org/>
18. Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.
19. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.
20. Paez, J. G., Lin, M., Beroukhi, R., Lee, J. C., Zhao, X. et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71.

21. Huang, H. and Keohavong, P. (1996) Fidelity and predominant mutations produced by deep vent wild-type and exonuclease-deficient DNA polymerases during in vitro DNA amplification. *DNA Cell Biol.* **15**, 589–594.
22. <http://www.appliedbiosystems.com>
23. Heiner, C. R., Hunkapiller, K. L., Chen, S. M., Glass, J. I. and Chen, E. Y. (1998) Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res.* **8**, 557–561.
24. Keohavong, P. and Thilly, W. G. (1989) Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9253–9257.
25. Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G. et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18**, 763–770.
26. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. and Welch, D. M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143.
27. Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. et al. (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, 275.
28. Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105.
29. Bhangale, T. R., Stephens, M. and Nickerson, D. A. (2006) Automating resequencing-based detection of insertion-deletion polymorphisms *Nat. Genet.* **38**, 1457–1462.
30. <http://www.1000genomes.org>
31. Chen, F., Alessi, J., Kirton, E., Singan, V. and Richardson, P. (2006) Comparison of 454 sequencing platform with traditional Sanger sequencing: a case study with de novo sequencing of *Prochlorococcus marinus* NATL2A genome. Poster LBNL 59003. Plant & Animal Genome XIV Conference, January 14–18, 2006 (San Diego, CA).