CHAPTER 8

Genomics and Gene Expression

SUMMARY

Genomics and the corresponding techniques enabled researchers to sequence the entire human genome between 1986 and 2000, with the final draft completed in 2003 with the help of Celera Genomics led by Craig Venter. Throughout the many years for the Human Genome Project (and the unofficial human genome sequencing by Celera), other genomes were sequenced, including *Drosophila* and the mouse. Prior to the next-generation sequencing techniques, DNA sequencing involved constructing genome libraries using BACs and YACs, and then each construct was sequenced separately and compiled into the final sequence. New advances in sequencing technology allow sequencing of an organism's entire genome in a matter of a few days relatively cheaply when compared to the multiyear, multibillion-dollar Human Genome Project. Having an organism's DNA sequence is only a part of the challenge, however. Now, researchers are determining the function of the genes and other "junk" or noncoding sequences within the genomes.

Two types of projects are derived from the Human Genome Project. The first type is concerned with understanding the variations among genomes. The second type associated DNA sequence with function (or dysfunction). These projects include the HapMap Project, Human Epigenome Project, and the Human Microbiome Project.

Maps of the genome help orient the sequences when compiling them. Two types of maps exist: genetic maps and physical maps. Genetic maps plot various points, called genetic markers, relative to each other within the DNA sequence. The best genetic markers are genes that code for specific traits, which can be monitored during breeding experiments. Genetic markers in the human genome are based more on physical markers since mating experiments are unethical. Physical markers include restriction fragment length polymorphisms (RFLPs), which are easy to identify and follow throughout the generations. Naturally occurring tandem repeats that occur naturally within the genome, called variable number tandem repeats (VNTRs), are also used as physical markers. Microsatellite polymorphisms are similar to VNTRs, but the repeating units are not as lengthy. Additionally, single nucleotide polymorphisms (SNPs), which are substitutions of single nucleotides in the DNA, are also used as physical markers. They can occur anywhere in the DNA, including the coding region, and sometimes change the coding and can have an affect on protein structure and function.

Physical maps are more definitive in terms of the distance between features in the DNA sequence and are particularly useful for large genomes because they are more numerous. Sequence tagged sites (STSs) are 100–500 base-pair sequences. A specific type of STS is called an expression sequence tag (EST) because it actually codes for mRNA. Contig maps contain overlapping sequences that are useful to help piece together shorter regions into large segments.

Other types of mapping help solve some of the problems associated with library clones, such as two DNA fragments being inserted into the same vector. Radiation hybrid mapping results in the formation of a human–hamster hybrid cell with fragmented DNA that can then be examined for STSs and ESTs. Cytogenic maps utilize stains that indicate specific banding patterns on chromosomes, providing a comparison of genes on a larger scale.

Many techniques exist to sequence genes or whole genomes of organisms. Chromosome walking begins with the identification and sequencing of an STS or RFLP. Once that sequence is known, overlapping clones to those regions can be identified and sequenced. Eventually, the entire region of interest is sequenced by continuously walking one direction or the other along the chromosome. Shotgun sequencing is a more appropriate method for whole genome sequence. Clones from the library are sequenced and the data is compiled and ordered by a computer based on the overlapping sequences. The entire 3 billion base-pair human genome was sequenced using the shotgun approach. Unfortunately, because of highly condensed regions of the genome called heterochromatin, gaps in the sequence information of the human genome still exist.

The human genome consists of about 3.2 gigabase pairs (Gbp) of only adenine, cytosine, guanine, and thymine in the sequence, with 20,805 protein coding genes. Some of the protein products from these genes share similarity to other animals, such as worms and flies. The total gene numbers are estimated and subject to change. The human genome sequence is also broken up by the presence of introns, which can be tens of thousands of base pairs in lengths. The interceding exons might be mistaken for separate genes. Another factor that affects the estimation of gene number is the presence of pseudogenes, which are duplications of genes and are no longer functional. Additionally, within the human genome are large noncoding sequences. Many of these areas are conserved when compared to the mouse genome, implying an undiscovered function. Over 10,000 conserved noncoding elements (CNEs) exist, with 500 regions that have 200 or more nucleotides that are perfect matches to the mouse genome. Some are even conserved in nonmammalian animals. Proposed functions for these regions include enhancer elements for gene expression or insulator sequences. Repetitive LINE and SINE sequences also exist within the noncoding regions of the genome. LINE sequences autonomously move locations within the genome. SINE elements cannot move unless they have help from LINE proteins. The most common type of SINE is the Alu element.

Bioinformatics is the use of computers to handle the vast amount of data associated with biology, such as information regarding DNA sequences, macromolecules, and analyses of genome functions. Researchers are able to use multiple websites to examine the immense data using a more controlled and searchable method. Because of the immensity of biological data, computer programs use a process called data mining to search, sort, and interpret the data. Researchers can also search for DNA sequence similarities among different organisms and generate reports on the outputs. Other programs exist to help identify specific sequences or motifs within specific regions of the genome.

In terms of application of genomics and impact to humans, the greatest impact is in regards to disease identification and diagnosis. The identification of gene defects associated with a disease opens up the possibility of genetic testing, where an individual is tested for the presence of the defect even before symptoms of the disease appear. Muscular dystrophy, cystic fibrosis, and sickle cell anemia are just a few examples of diseases linked to a genetic defect. For other diseases, such as many cancers, an environmental factor influences the progression of the disease even when a genetic defect is present. The identification of the genetic defect might allow an individual to live his or her life differently in the hopes of preventing the disease.

Mutations are changes that occur in the sequence of DNA. Some mutations can be inherited if they occurred in the cells that give rise to gametes. Mutations can be classified as base substitutions, which, as the name applies, means that a single base in the DNA has been changed. If the base substitutions alters the code at that particular position and the amino acid is changed, this is called a missense mutation. In some cases, the amino acid is changed to an amino acid that has the same property, which is called a conservative substitution. In other cases, the change results in an amino acid with a different property that drastically changes the structure and function of the protein. A nonsense mutation results when the amino acid sequence is changed to a stop codon, which prematurely ends the polypeptide at that position. In addition to base substitutions, insertions or deletions may occur. Larger mutations also exist. Inversions cause large pieces of the chromosome to turn around and reinsert in the opposite direction. Large pieces of DNA can also be cut out and reinserted into other locations, sometimes on another chromosome. This is called translocation. Duplications are large pieces of DNA that get copied and moved to another location. Recombination hot spots account for much of the genetic variation in the human genome. Haplotype blocks, or hapblocks, are passed from parents to offspring and are defined regions that are linked as a group or block.

Molecular phylogenetics is the study of evolutionary relationships through comparisons in DNA or protein sequences. Essential genes or proteins often do not have as many changes within them in contrast to nonessential genes or proteins. Comparing the sequences from

two or more species for a given gene or protein generates an alignment of the sequences where gaps or differences can be observed. This alignment data can then be used to build phylogenetic trees that help determine relatedness of the individuals. The fewer the changes, the closer the organisms are related to each other.

Pharmacology has also evolved rapidly from the study of the human genome. Pharmacogenetics attempts to identify and link genes to a certain drug response to reduce the numbers of adverse drug reactions. Populations of individuals that do not respond to a drug as predicted could be genetically screened to identify any sequences, particularly in SNPs, that could account for the differences in drug metabolism. For example, an asthma patient's response to albuterol, a drug used to treat asthma, is partially dictated by the presence or absence of mutations in a gene that encodes a receptor affected by the albuterol.

Whether or not a gene is in fact transcribed into mRNA can be investigated at a global level using a DNA microarray (DNA chip) with representations of each gene in the genome printed onto it. The sequences on the chip act as the probe and hybridize with complementary sequences present in the sample. To examine gene expression, scientists extract total RNA from the cells grown under specific conditions and label it with fluorescent dyes. Fluorescence is monitored and compared with control experiments to indicate which RNA molecules are present at specific times and under the tested conditions. The results give an enormous amount of data from a global perspective.

There are actually two different types of DNA microarrays. The cDNA microarrays contain larger fragments than the oligonucleotide microarrays. Additionally, cDNA microarrays require the DNA fragments to be produced separately, purified, and then attached to the chip. On the other hand, the DNA is chemically synthesized directly on the chip for oligonucleotide microarrays. The experimental sample is hybridized to the complementary sequences present on the chips. If secondary structures are present in those sequences bound to the chip, hybridization is more difficult to obtain. Oligonucleotide probes tend to have more problems with secondary structures. Tiling arrays are oligonucleotide microarrays that have smaller sequences back to back, which cover the entire genome of an organism. They are useful to identify binding sites for transcription factors as well as many genetic markers.

Monitoring gene expression on a global scale provides valuable information and a starting point from which to develop more experiments. RNA-Seq creates complementary DNA libraries from fragmented whole cell mRNA. The cDNA is then sequenced and the sequences aligned. The copy number of cDNAs for a particular gene indicates the level of gene expression. There are numerous advantages to using RNA-Seq over microarrays. RNA-Seq can even be used to compare gene expression profiles of the same cells under different conditions. In clinical applications, RNA-Seq can identify SNPs and post-translational editing of mRNA.

Single gene monitoring is important to determine very specific conditions for gene expression. Gene fusions are genetic constructs that link the regulatory region for a gene of interest to the coding region for a reporter gene, such as *lacZ* or GFP. The *lacZ* gene encodes β-galactosidase, which can be assayed *in vitro* to monitor expression. GFP can also be monitored by assaying for fluorescence, either *in vivo* or *in vitro*. Other candidate reporter genes include *phoA* for alkaline phosphatase and *lux* or *luc* for monitoring luminescence. The level of expression can be determined for the gene of interest by monitoring the product of the reporter gene.

Epigenetic changes are inherited changes in the DNA and chromosome structure that do not change the sequence of nucleotides. These changes include DNA methylation patterns and histone modifications. In some organisms, RNAi and prions are even inherited, although these are not considered true epigenetic effects unless gene expression is altered. True epigenetic events must be inherited, either from parents to offspring or from between cells within the same organism. The *papBA* operon of *Escherichia coli* is an example of epigenetics within a single-celled organism. In higher organisms, DNA methylation plays major roles

in gene regulation of development. Methylation in humans typically occurs at CpG motifs to turn off gene expression. Changes in methylome as humans age contribute to aging and acquisition of age-related diseases, such as cancer. The function of epigenetics in mammals is significant and is involved in genome stability, X chromosome inactivation, parental imprinting, and development and differentiation. Environmental factors, including the diet of grandparents and parents, can potentially influence gene expression in future generations. From an evolutionary standpoint, this can "pre-adapt" the newer generations to the current environmental state.

Case Study The Next-Generation Sequencing Revolution and Its Impact on Genomics

Daniel C. Koboldt et al. (2013). *Cell* 155, 27–38.

Initially, genomics focused on size and repetitive content of bacterial or viral genomes or single genes. Improvements in sequencing technologies expanded genomics to include entire chromosomes and eventually decoding larger genomes, including the human genome. Vast improvements in sequencing technologies (next-generation sequencing) now enable researchers to use sequencing in clinical diagnostics and testing and to examine biological phenomena on a genome-wide scale.

As stated in this review, what are the two technical preparatory approaches to investigate selected genomic regions?

The first approach uses PCR with multiplex primers to couple high throughput of newer sequencing platforms with the individual single DNA products obtained in the mixture. The adapters added to the ends of the PCR products create a library that is ready to be sequenced. The second approach uses a hybrid capture system that probes whole-genome libraries with a mixture of highly specific probes. The probes are labeled, enabling a researcher to capture a specific probe:library complex.

Techniques to analyze sequenced genomes have also improved. What is GWAS? What are some applications of GWAS?

Genome-wide association studies (GWAS) compare allele frequencies across the genomes of experimental and control groups to define alleles that are common markers for the disease. Thousands of individuals could be screened and genotyped using single-nucleotide polymorphisms on microarrays. GWAS are also used to study diseases, such as Crohn's disease and age-related macular degeneration, and pharmacogenetics.

When genomes are analyzed, sequences are often lined up against a reference sequence. Since no two individuals in any population are genetically identical, what methods are utilized to examine variants?

Regardless of rarity, most variants can be discovered if the appropriate technique and analysis are employed. The human genome reference sequences lacks novel genome content across all humans. Single nucleotide polymorphisms (SNPs) from germline cells have been detected and typed using new analysis algorithms. Various statistical tests and strategies have been developed to examine the rare variations and their contribution to complex phenotypes.

What are *de novo* mutations? What information can be gleaned from identification of these types of mutations? Have any tools been developed that can assist in the identification of *de novo* mutations?

De novo mutations are rare mutations that arise first with an individual. In terms of disease progression, individuals carrying *de novo* mutations likely have the disease phenotype. In addition to disease biology, identification of these mutation types can assist researchers with defining mutation rates. Multiple computational tools have been developed that can assist researchers with identification of *de novo* mutations. These tools generally examine pedigrees, relationships, siblings, and phenotypes; sequence coverage; error rate; and expected rates of *de novo* mutations.

Studying rare Mendelian disorders has helped identify genes that play a role in disease biology in humans. However, the studies often fail in their attempts to search for Mendelian disease genes. Why?

Even though the sample sizes are sufficiently large, often the identified variants are deemed nonpathogenic. In addition, some disease variations are not located within conserved regions or coding regions and could potentially be missed due to insufficient coverage. Also, functional analysis of variants needs improvement. Expected genotypic patterns based on incorrect diagnoses or assumed inheritance types could lead to errors in identifying variants associated with Mendelian disease genes.

How might comparing an individual's cancer genome to a normal genome be beneficial? Are there any somatic variations that are particularly difficult to identify?

When a cancer genome is compared with an unaffected genome, the somatic changes that occurred during the transition period can be identified. Structural variants, which are highly present in cancer genomes, are difficult to predict and must use a different analysis type.

The genomic applications for next-generation sequencing techniques and applications are numerous. In addition to those listed in the preceding questions, how else might this technology be used?

This technology and subsequent analyses could potentially be used as a noninvasive method for prenatal testing. Fetal DNA circulates in maternal blood plasma and can potentially be used to examine single nucleotide polymorphisms. There are still limitations to this technology, including cost-prohibitive sequencing technologies.

The possibilities for next-generation sequencing technologies are immense and range from cancer biology, disease detection, mutation rates, and prenatal testing applications. Some limitations of these technologies with regards to applications are due to such factors including cost and sequence coverage. There are also privacy concerns and ethical aspects to consider as we now have the ability to decode individual genomes and identify mutations associated with diseases.

Cell

The Next-Generation Sequencing Revolution and Its Impact on Genomics

Daniel C. Koboldt,[1] Karyn Meltz Steinberg,[1] David E. Larson,[1] Richard K. Wilson,[1] and Elaine R. Mardis[1,*]
[1]The Genome Institute, School of Medicine, Washington University, St. Louis, MO 63108, USA
*Correspondence: emardis@wustl.edu
http://dx.doi.org/10.1016/j.cell.2013.09.006

Genomics is a relatively new scientific discipline, having DNA sequencing as its core technology. As technology has improved the cost and scale of genome characterization over sequencing's 40-year history, the scope of inquiry has commensurately broadened. Massively parallel sequencing has proven revolutionary, shifting the paradigm of genomics to address biological questions at a genome-wide scale. Sequencing now empowers clinical diagnostics and other aspects of medical care, including disease risk, therapeutic identification, and prenatal testing. This Review explores the current state of genomics in the massively parallel sequencing era.

Prior to the advent of next-generation sequencing (NGS) technology, genomics initially was concerned with studying genomes that were tractable from the standpoint of size and repetitive content (e.g., viruses and bacteria) and with characterization of single genes associated with disease (e.g., Cystic Fibrosis, Huntington disease, and cancer). As the ability to construct large clone-based physical maps improved, the subcloned fragments of the genome contributing to physical map construction could be sequenced as individual projects, and their finished sequences melded together to represent entire chromosomes. Hence, important large genomes, including model organisms and the human genome, were decoded. Indeed, in the era of NGS, the short reads obtained from most platforms absolutely require these reference genomes as a substrate for read alignment prior to variation discovery. The impact of these technologies on genomic variant discovery has been profound, as we will describe. Although we limit the scope of this Review to genomics, an accompanying Review explores the disruptive impact of NGS on studies of the epigenome to further highlight the profound transformation brought on by NGS technology (Rivera and Ren, 2013 [this issue of *Cell*]).

Genomic Techniques

Although NGS technology initially was used to study whole genomes, a variety of approaches that address defined regions of the genome have emerged. There are essentially two technical preparatory approaches to explore selected regions of the genome with NGS. The first is by PCR, typically involving multiple primer pairs in a mixture that are combined with genomic DNA of interest in a multiplex approach to preserve precious DNA. The use of multiplex primer pairs couples the high throughput of NGS platforms and the fact that each sequence read represents a single DNA product in the mixture due to the nature of the sequencing platforms (Mardis, 2013). Following the PCR, the resulting fragments have platform-specific adapters ligated to their ends to form a library that is suitable for sequencing. The second approach involves hybrid capture, which has been developed by several groups and commercialized (Albert et al., 2007; Gnirke et al., 2009; Hodges et al., 2007). Essentially, hybrid capture takes advantage of the hybridization of DNA fragments from a whole-genome library to complementary sequences that were synthesized and combined into a mixture of probes designed with high specificity for the matching regions in the genome. These probes typically have covalently linked biotin moieties, enabling a secondary "capture" by mixing the probe:library complexes with streptavidin-coated magnetic beads. Hence, the targeted regions of the genome can be selectively captured from solution by applying a magnetic field, whereas most of the remainder of the genome is washed away in the supernatant. Subsequent denaturation releases the captured library fragments from the beads into solution, ready for postcapture amplification, quantitation, and sequencing. When the probes are designed to capture essentially all of the known coding exons in a genome, the capture approach is referred to as "exome sequencing." Additional probes may be designed, synthesized, and added to an exome reagent, typically referred to as "exome plus." When only a subset of the exome or of the genome outside of the exome is targeted, this is called a "targeted panel."

Genomic Analysis

As important as techniques to produce the NGS data that address biological questions are, analytical approaches are equally critical for successful interpretation of those data. Many analytical approaches depend on the digital nature of NGS data, a consequence of the fact that individual DNA fragments of the library are amplified either on beads or on flat surfaces (platform specific) prior to the sequencing reaction. Hence, each sequence read is equivalent to a single DNA fragment. What follows are selected data analysis techniques from a dizzying number of advances published in just the last 18 months. The pace of innovation in analytical approaches to genome-wide data analysis continues to engage and excite the computational biology community as the number of technical applications continues.

Technological advances have often driven the methods for discovering new disease genes. Early studies leveraged families in which a disease was segregating to identify the genetic causes of the phenotype. These linkage analysis studies were successful for highly penetrant, monogenic diseases such as cystic fibrosis. Standard parametric linkage studies of some complex traits were successful, particularly when sampling from extreme ends of the phenotypic distribution. For example, analyzing families segregating early onset Alzheimer's disease led to the discovery of multiple genes that contribute significantly to the phenotype and shed light on the biological mechanisms (e.g., plaque formation) of disease progression (Goate et al., 1991; Harrington et al., 1995; Pericak-Vance et al., 1991).

Yet, for many complex diseases and traits, this model was not as successful because the genetic predispositions to complex traits are, as their name implies, more difficult to elucidate and require larger numbers of samples to discern signal from noise. Theoretically, it was determined that comparing allele frequencies across the genome between large numbers of cases and controls would be able to capture common disease susceptibility alleles (Risch and Merikangas, 1996), and this ushered in the era of genome-wide association studies (GWAS). It was economically practical to screen thousands of individuals by genotyping hundreds of thousands of common single-nucleotide polymorphisms (SNPs) on microarrays. GWAS are well suited too and have been successful in studying population structure (Price et al., 2010b), anthropomorphic traits (Berndt et al., 2013), targets of natural selection such as variants associated with high-altitude adaptation (Bigham et al., 2009, 2010; Scheinfeldt et al., 2012), and some complex diseases such as Crohn's disease (Yamazaki et al., 2005) and age-related macular degeneration (Klein et al., 2005). These studies led to hundreds of replicable associated loci that cannot be fully enumerated in this Review. GWAS has perhaps had the most impact in the area of pharmacogenomics, where robust, highly replicable associations have impacted clinical actions. For example, warfarin dose is routinely adjusted based upon *VKORC1*, *CYP2C9*, and *CYP4F2* genotypes confirmed by GWAS (Takeuchi et al., 2009), which has significantly improved patient outcomes. Yet, most early GWAS yielded few variants with large effect sizes; this was perhaps to be expected, given the heterogeneity of the phenotypes and sample sizes needed to statistically detect signals of association.

The exponentially decreasing cost of next-generation sequencing data generation has put large-scale investigation of rare variation within reach, and there has been a resultant shift in the field of complex disease genetics over the past 5 years. GWAS data strongly suggest that the vast majority of the heritability of complex traits will not be due to a few common variants with low to moderate effects (Schork et al., 2009). Rare variation with large effect sizes is likely contributing a significant proportion to the "missing heritability" of complex traits and disease (Cohen et al., 2006; Manolio, 2009; Zhu et al., 2010). The common disease-common variant versus common disease-rare variant debate remains unresolved. There are still questions that remain as to whether the genetic contribution to common traits can be attributed to an infinite number of common alleles with small effect, a large number of rare alleles with large effects, or some combination of genes and environment (Gibson, 2011). But the evaluation of rare variants in common disease is ongoing.

Variant Detection

The advent of NGS has enabled the inquiry of nearly every base in the genome, and thus techniques to reliably interpret and identify millions of variants are being developed. As will be described below, the advantage of sequencing in this regard is that most variants, common and rare, can be discovered with the appropriate sequencing read coverage, algorithmic methods to identify the variants, and a sufficient careful orthogonal validation to confirm true from false positives. The exception to this discovery potential is due to the reliance on alignment to the Human Genome Reference sequence, which is the first step to analysis of NGS data, as this reference does not contain the entirety of novel genome content across all humans. Numerous variant-calling algorithms have been developed for the detection and genotyping of germline SNPs (DePristo et al., 2011; Koboldt et al., 2009; Li et al., 2008; McKenna et al., 2010; Shen et al., 2010) and small indels (Emde et al., 2012; Leone et al., 2013; Ye et al., 2009) in high-throughput sequencing data. Once detected, these variants can be analyzed in case-control studies using the same methods that have been developed for GWAS.

Rare Variation and Burden Testing

However, unlike GWAS (which examines common mutations), sequencing facilitates the discovery of rare mutations that, combined with the continuing unexplained genetic contributions to complex phenotypes from GWAS (Manolio et al., 2009), has sparked intense interest in measuring their association with complex phenotypes. This interest has given rise to a variety of statistical tests with varying strategies for detecting association of rare variation with phenotype (Chen et al., 2013; Han and Pan, 2010; Ionita-Laza et al., 2013; Lee et al., 2012a, 2012b; Li and Leal, 2008; Liu and Leal, 2010; Madsen and Browning, 2009; Neale et al., 2011; Oualkacha et al., 2013; Price et al., 2010a; Wu et al., 2011; Zhang et al., 2011). In any single gene, there are a large number of rare variants due to recent human population growth (Coventry et al., 2010; Nelson et al., 2012; Tennessen et al., 2012), and there may be many nonassociated variants in a gene. Furthermore, even in large cohorts, there may not be enough individuals with a given variant to achieve statistical significance.

To deal with the aforementioned challenge, all of these types of tests share the common feature that they group or collapse rare variation, usually by gene, in order to increase statistical power (see Wu et al., 2013 for a recent review). Early tests (such as the cohort allelic sums test [Morgenthaler and Thilly, 2007] and the combined multivariate collapsing method [Li and Leal, 2008]) assumed that each variant had the same direction of effect and, in addition, required a fixed minor allele frequency cutoff to define which variants to include; but these assumptions are not always valid or optimal. Further innovations have allowed for weighting of individual variants (for example, by variant frequency in the weighted sum statistic [Madsen and Browning, 2009] or the data [Han and Pan, 2010; Lin and Tang, 2011; Wu et al., 2011; Zhang et al., 2011]), variants with heterogeneous direction of effect (Han and Pan, 2010; Lin and Tang, 2011; Neale et al., 2011; Wu et al., 2011; Zhang et al., 2011), and selection of the ideal frequency cutoff for rare variants (Price et al., 2010a).

Though this remains an active area of research, the SKAT family of tests (Chen et al., 2013; Ionita-Laza et al., 2013; Lee et al., 2012a, 2012b; Oualkacha et al., 2013; Wu et al., 2011) has emerged as one of the most popular. SKAT and its variants allow for inclusion of covariates for managing both case-control and quantitative data and family or unrelated data, and they are computationally undemanding. Although the initial version of SKAT lost power in cases in which all variants in a gene have the same direction of effect, the newer SKAT-O (Lee et al., 2012a) test combines a test handling bidirectional effects and a test handling unidirectional effects to achieve excellent power in either case.

Identifying De Novo Mutations

The rarest of variants are de novo mutations: those variants that arise first in an individual. They have tremendous relevance for disease biology, as they are more likely to have functional consequences in rare diseases. Characterizing these mutations also allows for the estimation of the baseline human mutation rate as well as its correlation to parental age (Abecasis et al., 2010; Kong et al., 2012). An entire class of computational tools has arisen that utilize both sequencing data and pedigree information to identify de novo mutations genome wide. Most of these tools currently deal with trios (mother, father, and child) only and can identify de novo variants arising in the children (Cartwright et al., 2012; http://sourceforge.net/p/denovogear/wiki/Home/; Li et al., 2012; Li, 2011). Because sequencing reads have a higher error rate than traditional genotyping, these tools incorporate information about coverage, the sequencing error rate, the expected de novo mutation rate, and family relationships.

Although all of these tools identify potential de novo mutations, there remain significant feature differences between them, and no single tool has yet emerged as the frontrunner. In addition, only Samtools, DeNovoGear (DNG), and GATK can also predict de novo indels. Both DNG and Polymutt can handle larger pedigrees, with DNG able to handle multiple siblings and Polymutt able to handle arbitrarily large pedigrees.

Studying Rare Mendelian Disorders

Rare monogenic disorders have provided unique opportunities to identify disease genes in humans. Traditionally, such disorders were studied by positional cloning or candidate gene approaches. Determining their molecular basis, however, was often hindered by small kindred sizes, genetic heterogeneity, and diagnostic classifications that may not reflect molecular pathogenesis. However, high-throughput sequencing of the full set of protein-coding genes—the exome—helps to overcome these obstacles by screening thousands of genes in a single experiment. Although this limits the types of mutations that can be discovered, rare coding variants that are predicted to have significant functional consequences can be discovered (Bamshad et al., 2011). In fact, it is estimated that, in ~60% of projects, exome sequencing will identify new Mendelian disease genes (Gilissen et al., 2012), and it is likely this approach also will contribute to complex disease genetics. Hence, the exome represents an enriched target space to identify rare variants with large effect sizes, as opposed to GWAS, wherein variants have low effect sizes.

The analytical approach applied to most exome sequencing studies of rare disorders is relatively straightforward. First, genetic variants shared by affected individuals (or segregating with a phenotype, in family studies) are collected. Hundreds or thousands of variants might be in this initial set. These are filtered using information from public databases (e.g., dbSNP [Sherry et al., 2001]) to remove common polymorphisms, based on the expectation that causal mutations will be extremely rare in human populations. Next, annotation with gene structure information and bioinformatics programs (e.g., SIFT, Polyphen, CONDEL) further restricts the list of candidates to those most likely to affect an encoded protein. Ideally, these sequential filtering steps reduce the list to a handful of candidate causal variants, which can be further evaluated with mutation screening (in other family members or unrelated, affected individuals), pathway analysis, and functional validation.

Somatic Variant Detection

The comparison of an individual's cancer genome to the normal genome (derived from an unaffected tissue DNA) provides a comprehensive description of the somatic changes that have occurred in the transition from normal to cancerous cells. WGS approaches to somatic variant detection are more challenging due to the size of the data and the numerous types of variants that can be discovered by different algorithmic predictors, relative to exome sequencing. However, structural variants, which are most difficult to predict accurately and with a reasonable false positive rate, occur frequently in cancer genomes and only can be discovered from WGS data. With an increasing focus on characterizing cancer heterogeneity, discussed below, the ability of somatic variant detection algorithms to predict low-frequency single-nucleotide variants (SNVs) in cancer cell populations is becoming critically important. There are several new algorithms with this capability, including Strelka (Saunders et al., 2012), VarScan 2 (Koboldt et al., 2012), and MuTect (Cibulskis et al., 2013). Strelka implements a Bayesian approach that treats the tumor and normal allele frequencies as continuous variables. In particular, the normal sample is represented as a mixture of diploid germline variation with noise, and the tumor samples are represented as a mixture of the normal sample with somatic variation. This approach is meant to provide robust call sensitivity on low-purity samples and, as such, provides the same robust sensitivity for low-level variants. Accuracy around indel detection is achieved by Strelka by jointly performing indel search and read realignment in the context of both samples. VarScan 2 is a somatic variation version of the original VarScan algorithm that applies heuristic methods and a statistical test to detect SNVs and indels and their somatic status by simultaneously analyzing the tumor and normal data. In addition, VarScan 2 can identify both LOH and somatic copy number alterations as deviations from the log ratio of sequence coverage depth within the pair that are quantified statistically. MuTect takes input data from matched tumor and normal DNA alignments and removes low-quality sequence data. Variant detection is performed in the tumor data by a Bayesian classifier, filters to remove false positives due to sequencing artifacts that are not captured by the prior error-model-based filters, and designates variants as somatic or germline using a second Bayesian classifier. The step to remove rare sources of false positives uses a panel of normal samples filter that represents rare error modes only detectable from the comparison to additional samples.

Table 1. OMIM Phenotypes for which the Molecular Basis Is Known, 2007 and 2013

Inheritance Pattern	January 2007	July 2013
Autosomal	1,851	3,525
X Linked	169	277
Y Linked	2	4
Mitochondrial	26	28
Total	2,048	3,834

Exciting Biological Insights from Recent Studies

Rare Inherited Disorders

Although next-generation sequencing has impacted the human genetics field as a whole, few areas have benefited more than the study of rare genetic diseases. Some of the earliest applications of NGS to Mendelian disorders (Table 1) demonstrated that it was possible to identify disease-causing genes by sequencing the exomes of a few unrelated individuals (Gilissen et al., 2010; Hoischen et al., 2010; Lalonde et al., 2010; Ng et al., 2010a, 2010b) or affected family members (Bilgüvar et al., 2010; Bolze et al., 2010; Johnson et al., 2010; Krawitz et al., 2010; Musunuru et al., 2010; Walsh et al., 2010; Wang et al., 2010). Even the exome sequence of a single index case proved sufficient for genetic diagnosis for some disorders when information about the molecular underpinnings of the disease was known. For example, prioritization of mitochondrial proteins helped to identify *ACAD9* in a case with complex I deficiency (Haack et al., 2010), whereas prior evidence linking *STIM1* to recessive immunodeficiency helped to implicate this gene in a pediatric case with classic Kaposi sarcoma associated with human herpesvirus 8 infection (Byun et al., 2010).

The impact of NGS technologies on rare genetic diseases is further evidenced by the growth of the Online Mendelian Inheritance in Man (OMIM) database (McKusick, 2007), in which the number of inherited phenotypes for which the molecular basis is known has nearly doubled since 2007 (Table 2). The number of genes associated with rare diseases, too, has grown at an impressive rate. Yet for many disorders, elucidation of the genetic basis has outstripped an understanding of the molecular and pathological mechanisms of disease. More work will be required to determine the precise relationship between genotype and phenotype.

Lessons from Mendelian Disease Studies

Although NGS offers a powerful strategy to search for Mendelian disease genes, it is important to realize that many such studies fail despite sufficient numbers of samples. One failure occurs when the causal variant is found but is deemed nonpathogenic. While the majority of known disease-causing mutations affect highly conserved protein residues, other pathogenic mechanisms—such as synonymous changes of rare codons that affect the rate of cotranslational folding (Kimchi-Sarfaty et al., 2007)—may be responsible but not ascribed importance. This emphasizes the need for better functional assays of discovered variants.

It is also possible to miss a causal variant. Even with NGS and hybrid capture, ~5% of target coding bases do not achieve sufficient coverage for reliable variant detection. Furthermore, with adequate coverage, certain types of mutations (e.g., inversions, duplications, and other structural variants) remain challenging to detect. Causal mutations also may reside outside of the regions targeted for exome sequencing. Nearly half of familial ALS in Finnish populations, for example, is caused by a hexanucleotide repeat expansion in the intron of the *C9orf72* gene (Renton et al., 2011).

Failure also can result when one of the underlying assumptions was incorrect. Genetic and phenotypic heterogeneity can hinder correct diagnosis of cases, or the assumed mode of inheritance (and therefore expected genotype pattern) could be incorrect. In retinitis pigmentosa (RP), for example, around 8.5% of families provisionally diagnosed with autosomal dominant disease truly have X-linked RP (Churchill et al., 2013).

In addition, the reliance on public databases such as dbSNP may confound some analyses of NGS data. The number of known variants in the human genome has risen dramatically over the past decade (Figure 1), fueled in large part by the advent of NGS technologies. Intriguingly, although the submissions from 2007 to 2012 grew almost exponentially, the number of unique reference variants (RefSNPs) followed a more linear growth. Further, a comparison of the global minor allele frequency (GMAF) distribution between dbSNP builds 135 (October 2011) and 137 (June 2012) demonstrates that most of the recent growth came from variants that were rare (GMAF < 0.01) or extremely rare (GMAF < 0.001) in human populations (Figure 2).

These trends suggest that the majority of common sequence variants in humans have already been reported, and those that remain undiscovered tend to be rare, perhaps specific to an individual or population. This has important implications for studies of rare genetic diseases. The ponderous size of dbSNP certainly makes it a powerful discriminatory tool for common variation. However, it also suggests that dbSNP filtering approaches must be applied with caution because dbSNP entries are associated with disease—variants from Online OMIM (McKusick, 2007) or mutations from the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2010, 2011) —and a growing number are too rare to exclude from consideration.

Sequencing under GWAS Peaks

One way to leverage the results from GWAS and linkage studies to identify rare variation is to perform targeted sequencing of the regions identified under significant peaks. This strategy has been used to identify a rare variant in a gene under a linkage peak where common SNPs could not explain the variance in the phenotype (Bowden et al., 2010). In this study, common polymorphisms in the *ADIPOQ* gene that are highly associated with circulating plasma adiponectin levels in European populations were minimally associated with plasma adiponectin levels in Hispanic families; however, a rare coding mutation was identified that contributes up to 17% of the observed variance in Hispanic plasma adiponectin levels. Additionally, Wang et al. (2013) sequenced exons of >1,000 genes identified from GWAS linkage peaks that impact human stature. Using a pooled sample strategy of individuals who were significantly shorter than the average population but were not diagnosed with any known syndrome affecting height or with any endocrinological deficiency, the researchers identified unique rare nonsynonymous and splicing

Table 2. Disease-Causing Genes Identified by Exome Sequencing Studies, 2009–2010

Gene	Disorder	Individuals	Citation
DHODH	Miller syndrome	four affected from three kindreds	(Ng et al., 2010b)
FLVCR2	Fowler syndrome	two unrelated	(Lalonde et al., 2010)
GPSM2	Nonsyndromic hearing loss	one proband	(Walsh et al., 2010)
MLL2	Kabuki syndrome	ten unrelated	(Ng et al., 2010a)
WDR62	Severe brain malformations	one proband	(Bilgüvar et al., 2010)
PIGV	Hyperphosphatasia mental retardation	three siblings	(Krawitz et al., 2010)
WDR35	Sensenbrenner syndrome	two unrelated	(Gilissen et al., 2010)
STIM1	Kaposi sarcoma	one patient	(Byun et al., 2010)
ANGPTL3	Familial combined hypolipidemia	two family members	(Musunuru et al., 2010)
ACAD9	Complex I deficiency	one patient	(Haack et al., 2010)
SETBP1	Schinzel-Giedion syndrome	four unrelated	(Hoischen et al., 2010)
TGM6	Spinocerebellar ataxia	four family members	(Wang et al., 2010)
FADD	Autoimmune lymphoproliferative syndrome	one proband	(Bolze et al., 2010)
VCP	Familial ALS	two family members	(Johnson et al., 2010)

mutations. In a similar study design, researchers were able to narrow a large 288 Kbp region identified from GWAS of multiple sclerosis to an 86.5 Kbp haplotype block containing 42 SNPs, using targeted capture and NGS (Cortes et al., 2013).

Family Studies of Complex Disease

There has been a return to family-based experimental designs for complex disease genetics recently, as it is expected that many members of the same family will carry a particular rare variant; hence, the number of individuals needed for rare variant discovery is much smaller than in cohorts of unrelated individuals (Bailey-Wilson and Wilson, 2011). Using a combination of exome and whole-genome sequencing of affected individuals in consanguineous families, researchers can use homozygosity mapping to identify and characterize the variants contributing to genetically heterogeneous disorders. Nonconsanguineous, large multigenerational, and multiplex pedigrees can also be used to identify rare inherited variants. For example, Weedon et al. (2011) identified a novel heterozygous mutation in *DYNC1H1*, segregating in a four-generation family affected with Charcot-Marie-Tooth disease by using whole-exome sequencing (WES). Similarly, WES was performed on a three-generation family with multiple individuals affected with pulmonary arterial hypertension who did not carry the canonical *TGF*-β mutation (Austin et al., 2012). WES revealed a frameshift mutation in caveolin-1 (*CAV1*) that reduced the normal caveolin-1 in the endothelial cell layer of the small arteries. In many cases, the variants identified in these studies can also be independently validated in other cohorts. For example, WES of a large pedigree identified a missense mutation in the affected individuals that also segregated with other families suffering from late-onset Parkinson disease (Vilariño-Güell et al., 2011; Zimprich et al., 2011), thus bolstering the significance of the association. In some studies, WES results provide insights into the biological pathways involved in disease susceptibility and/or pathogenicity. For example, Timms et al. (2013) analyzed the exomes of multiplex families with schizophrenia and identified rare coding variants in N-methyl-D-aspartate (NMDA) receptor genes in all of the families. Although the variants were dispersed over many genes, this pathway was significantly enriched for rare, deleterious mutations and suggested a possible role for glutamate signaling in the pathogenesis of schizophrenia.

De Novo Mutation Studies

Although genomic research in the past decade has largely emphasized inherited variation, NGS technologies also allow us to study, at base-pair resolution, the mutational processes that occur in humans from one generation to the next. Family-based WGS studies have shown that each individual's genome harbors ~74 germline de novo mutations (DNMs) (Conrad et al., 2011). These mutations are potentially more deleterious because they have not been subject to natural selection and therefore are of considerable interest for sporadic diseases.

Neurological and developmental disorders in particular highlight the impact of DNMs on disease risk. Exome sequencing revealed rare de novo protein-altering mutations in seven of ten individuals with idiopathic intellectual disability (ID) affecting nine different genes (Vissers et al., 2010). Four large-scale studies (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) evaluated the impact of DNMs in autism spectrum disorder (ASD) via exome sequencing of family quartets (patient, parents, and an unaffected sibling). Each study included >100 families and found that DNM rates were consistently higher in patients than in their unaffected siblings. Similar WES approaches have implicated genes expressed in the developing heart for sporadic congenital heart disease (Zaidi et al., 2013) and genes encoding chromatin regulators for sporadic ALS (Chesi et al., 2013). De novo mutational paradigms have also been suggested by exome sequencing in sporadic psychiatric disorders, such as schizophrenia (Girard et al., 2011; Xu et al., 2012). These findings collectively support a role for de novo mutational processes in sporadic disorders and highlight the extraordinary locus heterogeneity underlying susceptibility to complex diseases.

The application of NGS to both rare and common genetic diseases has offered many insights into disease etiology that undoubtedly merit deeper investigation. Taken together, these studies have also served to highlight our incomplete

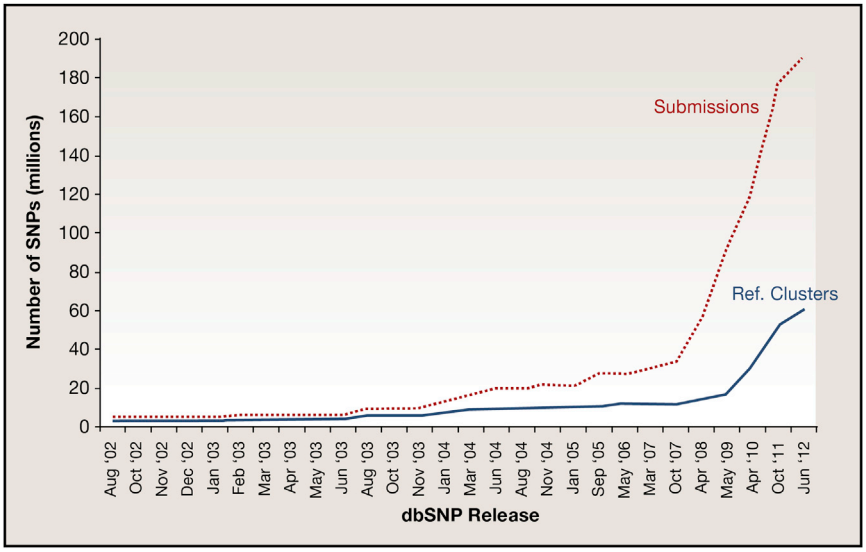

Figure 1. Growth in the Numbers of dbSNP Variants in the Human Genome

Increases in the numbers of SNPs submitted (dotted line) and cataloged as unique reference variants (solid line) in dbSNP are charted over the periodic database releases from August 2002 until the most recent release in June 2012. As indicated by the two lines, while overall submissions have increased exponentially since 2008 (when large projects such as the 1,000 Genomes Project began), the number of unique variants has not increased at a comparable rate.

understanding of the molecular mechanisms by which mutations cause disease. Nevertheless, it seems likely that applying NGS to uncover the genetic underpinnings of disease will help us to better understand the complex relationship between genotype and phenotype.

Cancer Genomics Discovery

Over the past two years, the growth in cancer genomics discovery due to NGS is unprecedented, with multiple examples of large-scale WGS- or WES-based studies published in the literature for both adult and pediatric cancer types. The growth in our knowledge of the genes frequently mutated in cancer genomes is illustrated in Figure 3, based on the number of new mutations deposited in COSMIC (Forbes et al., 2010, 2011). Here, the number of unique variants identified in tumor genomes stands in stark contrast to those in germline DNA shown in Figure 1. Namely, in dbSNP, there is a clear saturation effect because the majority of variants in any individual genome are shared with other members of the population (and thus already in dbSNP). In COSMIC, however, the number of unique variants closely mirrors the number of mutations submitted, reflecting the fact that most mutations in a tumor genome are private to that tumor.

Cancer Genome Heterogeneity

For >100 years, the view of cancer cells through the pathologist's microscope has indicated that not all cancer cells in a tissue block are entirely similar. Several groups, using the digital nature of NGS data, now have proven this "heterogeneity" of cancer cells at the genomic level. Initially, genomic heterogeneity was demonstrated by copy number comparisons between primary and metastatic disease (Campbell et al., 2010) and by whole-genome amplification and low-coverage sequencing of amplified genomic DNA from single breast cancer cells (Navin et al., 2011). Within the past year, published studies using either WES or WGS have demonstrated the changes in genomic heterogeneity in cancers over the primary-to-relapse/metastatic transition or have characterized heterogeneity with primary tumor specimens. Specifically, these changes are determined by comparing the associated changes in the percentage of tumor cells carrying specific mutations detected by deep coverage NGS data during disease progression. These studies evoke an evolutionary aspect to cancer's response to survival pressures, including therapy, and have fueled interest in better understanding the genomes of patients who are likely to recur in their disease.

Early in 2012, Ding et al. described changes in heterogeneity and subclonal architecture of primary acute myeloid leukemia (AML) samples compared to their matched first relapse samples for eight patients (Ding et al., 2012). Using WGS coupled with secondary deep hybrid capture-based NGS data on variant sites, clusters of mutations defining the genotypes of a founding clone and derived subclones were identified. In each case studied, the primary AML sample was either mono- or multiclonal, whereas the relapse sample was monoclonal and carried the somatic profile of one of the primary subclones, as well as new mutations that were acquired during chemotherapy. An analysis of transversion and transition mutations indicated that all types of transversions were elevated in the relapse samples, a DNA damage phenomenon that is attributable to the use of DNA-damaging chemotherapy agents.

In genomic analyses of renal cell cancers, Gerlinger and colleagues (Gerlinger et al., 2012) studied regional heterogeneity in four advanced tumors and metastases from a clinical trial of everolimus (an inhibitor of mTOR) to evaluate the similarities and differences in the genomic landscapes. Their approach included WES, SNP arrays, and gene expression arrays. Their results indicated a branching evolution of the primary and metastatic tumors studied, with a combination of universally shared and primary region-specific or metastasis-specific private mutations. Unlike the previous study, everolimus was shown to not impact the number and types of new mutations in posttreatment samples studied. A case was made for phenotypic convergent evolution due to spatially separated, distinct mutations in *SETD2*, *KDM5C*, and *PTEN*.

A study in breast cancer heterogeneity utilized data from 20 breast cancers selected across the different molecular subtypes, one of which was sequenced to 188-fold depth to provide sufficient sensitivity for heterogeneity analysis (Nik-Zainal et al., 2012). Much like the AML study mentioned above, clustering of mutations sharing similar variant fractions from high-coverage

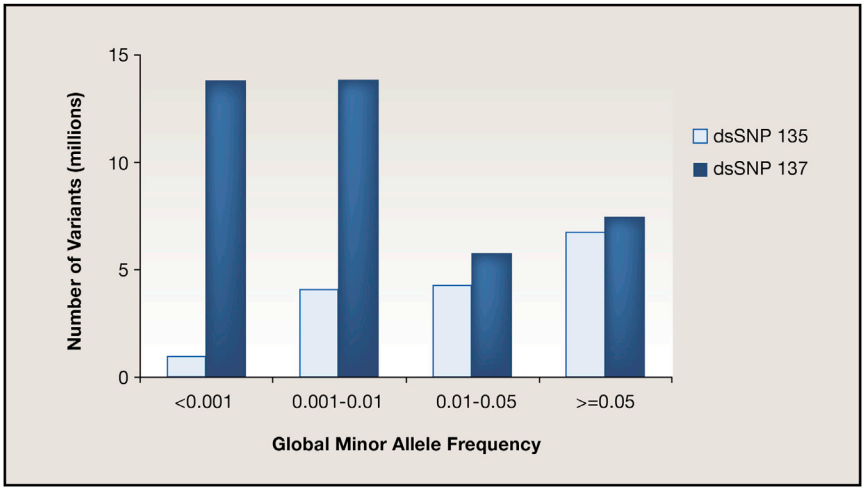

Figure 2. dbSNP Growth due to Rare Variant Discovery

This graphic illustrates the amount of rare and extremely rare variant discovery in two recent releases of the NCBI dbSNP database, where a global minor allele frequency of >0.05 is considered a common variant. As indicated, rare variant discovery has increased dramatically in the most recent build of dbSNP (137).

data was performed to identify the subclones. The clusters were further refined by application of a Bayesian Dirichlet process, and further associations were made to identify a hierarchy of mutational events in the natural history of the cancer's development.

Prediction of Targeted Therapy/Actionable Mutations

Since the earliest descriptions of specific mutations in *EGFR* predicting response to small-molecule inhibitors such as tyrosine kinase inhibitors (Lynch et al., 2004; Paez et al., 2004; Pao et al., 2004), the association of somatic mutations to drug response has been of increasing interest. The use of NGS technologies in this regard has several advantages over the original methods (PCR and Sanger fluorescent sequencing) used to acquire these data. Namely, the NGS-based inquiries required for discovering the gene-therapy association can be less hypothesis driven and examine all genes, the associated cost to generate the data for each patient sample is both less expensive and more rapidly obtained, and the ability to detect specific types of mutations such as insertions or deletions of one or several nucleotides is facilitated by NGS. The first aspect is important because most small-molecule therapies target a range of mutated proteins, so multiple genes must be tested in each patient. The second aspect is important because these queries are now approaching clinical usage wherein identification of appropriate therapy(ies) must happen in a 2–3 week period to be applicable to patient care. Lastly, although small insertion/deletion mutations are rarer than single-nucleotide substitutions, their impact on the resulting protein may be more profound. Because Sanger sequencing typically fails to detect these variants, it is both likely that the frequency of these mutations is underestimated and certain that their response to therapy is less well understood as a result.

One downside of the use of targeted small-molecule inhibitors is that many patients experience an initial complete pathologic response or at least stable disease but then acquire resistance to the therapy and progress (Engelman et al., 2007). This phenomenon has mainly been studied at the protein level (Girotti et al., 2013; Prahallad et al., 2012) or by focused sequencing (Sequist et al., 2011). Here, results often demonstrate that the cellular pathway blockade affected by targeted therapy is circumvented by new mutations and/or overexpression either of the targeted gene or of another gene in the same pathway. Given these discoveries, it remains to be demonstrated by deep NGS or single-cell sequencing of progression disease biopsies whether the mutations that enable circumvention of the blockade are pre-existing in a minor proportion of tumor cells or are new mutations that arise in response to the pathway blockade.

Circulating Tumor DNA Analysis

Many solid tumors shed cells and/or DNA into the blood stream at very low levels that are thought to fluctuate with increases or decreases in the disease burden of the patient. Hence, the ability to detect these changes with high sensitivity poses an interesting and potentially powerful disease-monitoring capability that likely would complement imaging modalities such as CT or MRI but at much lower cost and with lower associated morbidities (Diehl et al., 2005; Diehl et al., 2008; Swisher et al., 2005). In this regard, several groups have recently published manuscripts describing the selective capture of circulating tumor cells (CTCs) or the amplification and sequencing of circulating tumor DNA or RNA. This so-called "liquid biopsy" approach using plasma can detect the predominant somatic mutations for that tumor type (Forshew et al., 2012), or if chromosomal translocations or structural variants already are known from prior characterization of the cancer genome, PCR primers can be designed to amplify the tumor-specific products for NGS and analysis (Dawson et al., 2013; Leary et al., 2012). Another recently published example of this type of detection by NGS involved the detection of ovarian or endometrial cancer by gene-specific assays of PAP test samples (Kinde et al., 2013).

Noninvasive Prenatal Testing

As mentioned, clinical use of NGS in cancer diagnosis, therapeutic decision making, and progression monitoring is poised for introduction. Several large academic centers and a handful of commercial entities are offering NGS-based assays in the CLIA-regulated environment. An NGS-based clinical assay that already has received widespread adoption is noninvasive prenatal testing for chromosomal abnormality diagnosis using samples such as maternal blood. In 1997, Lo et al. demonstrated that male sex could be determined from circulating fetal DNA in maternal plasma and serum samples and that the level of circulating fetal DNA increases with gestational age (Lo et al., 1997). However, achieving high sensitivity and specificity of fetal genotype was difficult, given the low levels of fetal DNA and the cost of high-depth sequencing.

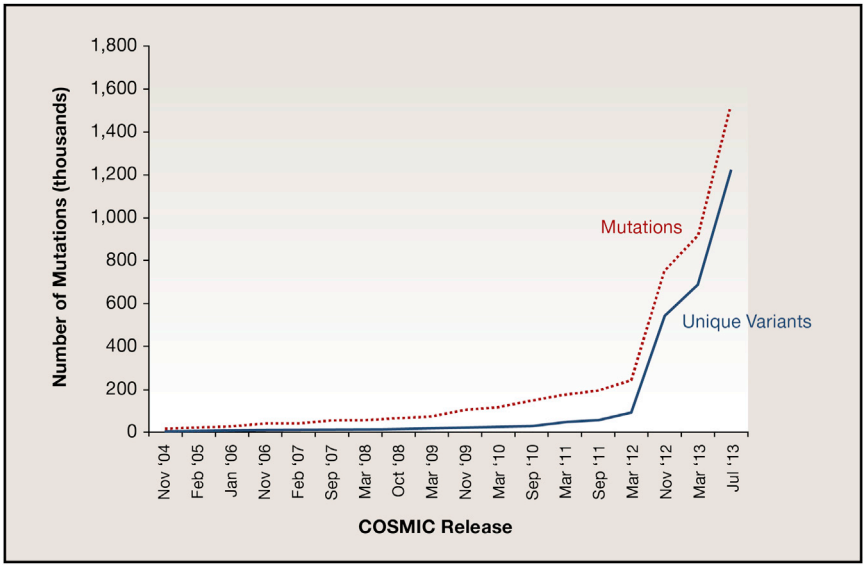

Figure 3. Growth in COSMIC Database Reports of Identified and Unique Mutations
Increases in the numbers of mutations and unique variants identified from DNA sequencing of cancer samples as cataloged in the COSMIC database, from November 2004 until the most recent release in July 2013. Note that the numbers of unique variants identified are increasing at a rate equal to the numbers of mutations discovered.

With the advent of NGS, resolving the whole genome of a fetus from maternal blood sources became possible. In 2010, Lo et al. sequenced maternal plasma genomic DNA to 65× coverage and then used the parental SNP genotypes (from SNP array data) to distinguish fetal versus maternal sequencing reads (Lo et al., 2010). This elegant proof-of-concept study demonstrated that the entire fetal genome is represented in the maternal plasma. Yet, this approach was limited by the use of a chorionic villus sample and the somewhat circular logic by which parental haplotypes were inferred from common heterozygous SNP genotypes and then used to predict the fetal haplotype, thereby missing a large proportion of the rare variation. In addition, the authors were unable to detect de novo mutations. To overcome these obstacles, Kitzman et al. used WGS with maternal plasma as well as fosmid clone pooling to resolve long haplotype blocks in the mother (i.e., "phasing"; Kitzman et al., 2012). The paternal genome was sequenced but not phased. This approach achieved >99% genotype accuracy at maternal heterozygous sites when predicting the fetal genotype. In addition, de novo mutations and recombination switch breakpoints were detected using a Hidden Markov model. The results were confirmed by WGS from cord blood after birth. Similarly, Fan et al. (2012) performed WGS and WES with maternal plasma and maternal haplotype resolution via direct deterministic phasing using single cells. The paternal genome was inferred using detection of paternal-specific alleles and imputation, and the fetal genome was resolved to >99% accuracy using molecular counting of parental genotypes in the maternal plasma.

These studies demonstrate the feasibility of prenatal testing at single-nucleotide resolution, but major limitations likely will hinder clinical translation. For example, sequencing to sufficient depth to detect fetal DNA genotypes is still quite expensive. In addition, it is prohibitively expensive and time consuming to routinely create and sequence maternal fosmid pools. As single-molecule sequencing technologies improve, it may be realistic to routinely resolve extended parental haplotypes to assist in fetal genotyping. For the time being, commercial noninvasive NGS prenatal tests are offered, but these only detect common chromosomal aneuploidies such as Trisomy 21.

Concluding Remarks

In summary, next-generation sequencing technologies have had an incredible impact on our knowledge of human genetic diseases over a very short time frame. Whether this trend will continue rests on a variety of issues, some quite complex. For example, the size of whole-human-genome data sets remains large, and this poses significant challenges for data download and storage and for computational infrastructure. Data privacy of human subjects is paramount but is increasingly difficult to control, raising concerns in the public that may inhibit consent by individuals to participate in genetic studies. Ethical aspects overshadow the return of information to study participants and individuals seeking genetic diagnosis due to our remaining ignorance about the pathologic and functional consequences of variation in the human genome. The next few years will determine which applications of NGS are incorporated into the clinical diagnostic setting, many of which may benefit patients but yet may not be covered by insurers. Even as this scenario plays out, it is undoubtedly the case that NGS will continue to be a revolutionary force in basic biomedical and biological genomics inquiry for some time to come.

REFERENCES

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods *4*, 903–905.

Austin, E.D., Ma, L., LeDuc, C., Berman Rosenzweig, E., Borczuk, A., Phillips, J.A., 3rd, Palomero, T., Sumazin, P., Kim, H.R., Talati, M.H., et al. (2012). Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. Circ. Cardiovasc. Genet *5*, 336–343.

Bailey-Wilson, J.E., and Wilson, A.F. (2011). Linkage analysis in the next-generation sequencing era. Hum. Hered. *72*, 228–236.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Berndt, S.I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. *45*, 501–512.

Bigham, A.W., Mao, X., Mei, R., Brutsaert, T., Wilson, M.J., Julian, C.G., Parra, E.J., Akey, J.M., Moore, L.G., and Shriver, M.D. (2009). Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. Hum. Genomics *4*, 79–90.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. *6*, e1001116.

Bilgüvar, K., Oztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A.O., Gökben, S., et al. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature *467*, 207–210.

Bolze, A., Byun, M., McDonald, D., Morgan, N.V., Abhyankar, A., Premkumar, L., Puel, A., Bacon, C.M., Rieux-Laucat, F., Pang, K., et al. (2010). Whole-exome-sequencing-based discovery of human FADD deficiency. Am. J. Hum. Genet. *87*, 873–881.

Bowden, D.W., An, S.S., Palmer, N.D., Brown, W.M., Norris, J.M., Haffner, S.M., Hawkins, G.A., Guo, X., Rotter, J.I., Chen, Y.D., et al. (2010). Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. Hum. Mol. Genet. *19*, 4112–4120.

Byun, M., Abhyankar, A., Lelarge, V., Plancoulaine, S., Palanduz, A., Telhan, L., Boisson, B., Picard, C., Dewell, S., Zhao, C., et al. (2010). Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J. Exp. Med. *207*, 2307–2312.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Cartwright, R.A., Hussin, J., Keebler, J.E., Stone, E.A., and Awadalla, P. (2012). A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. Stat. Appl. Genet. Mol. Biol. *11*, 11.

Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. *37*, 196–204.

Chesi, A., Staahl, B.T., Jovičić, A., Couthouis, J., Fasolino, M., Raphael, A.R., Yamazaki, T., Elias, L., Polak, M., Kelly, C., et al. (2013). Exome sequencing to identify de novo mutations in sporadic ALS trios. Nat. Neurosci. *16*, 851–855.

Churchill, J.D., Bowne, S.J., Sullivan, L.S., Lewis, R.A., Wheaton, D.K., Birch, D.G., Branham, K.E., Heckenlively, J.R., and Daiger, S.P. (2013). Mutations in the X-linked retinitis pigmentosa genes RPGR and RP2 found in 8.5% of families with a provisional diagnosis of autosomal dominant retinitis pigmentosa. Invest. Ophthalmol. Vis. Sci. *54*, 1411–1416.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Cohen, J.C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G.L., Grundy, S.M., and Hobbs, H.H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc. Natl. Acad. Sci. USA *103*, 1810–1815.

Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project. (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.

Cortes, A., Field, J., Glazov, E.A., Hadler, J., Stankovich, J., Brown, M.A., and Brown, M.A.; ANZgene Consortium. (2013). Resequencing and fine-mapping of the chromosome 12q13-14 locus associated with multiple sclerosis refines the number of implicated genes. Hum. Mol. Genet. *22*, 2283–2292.

Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat. Commun. *1*, 131.

Dawson, S.J., Tsui, D.W., Murtaza, M., Biggs, H., Rueda, O.M., Chin, S.F., Dunning, M.J., Gale, D., Forshew, T., Mahler-Araujo, B., et al. (2013). Analysis of circulating tumor DNA to monitor metastatic breast cancer. N. Engl. J. Med. *368*, 1199–1209.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., Diaz, L.A., Jr., Goodman, S.N., David, K.A., Juhl, H., et al. (2005). Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc. Natl. Acad. Sci. USA *102*, 16368–16373.

Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S.A., et al. (2008). Circulating mutant DNA to assess tumor dynamics. Nat. Med. *14*, 985–990.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Emde, A.K., Schulz, M.H., Weese, D., Sun, R., Vingron, M., Kalscheuer, V.M., Haas, S.A., and Reinert, K. (2012). Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. Bioinformatics *28*, 619–627.

Engelman, J.A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J.O., Lindeman, N., Gale, C.M., Zhao, X., Christensen, J., et al. (2007). MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science *316*, 1039–1043.

Fan, H.C., Gu, W., Wang, J., Blumenfeld, Y.J., El-Sayed, Y.Y., and Quake, S.R. (2012). Non-invasive prenatal measurement of the fetal genome. Nature *487*, 320–324.

Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. *38*(Database issue), D652–D657.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*(Database issue), D945–D950.

Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D.W., Kaper, F., Dawson, S.J., Piskorz, A.M., Jimenez-Linan, M., Bentley, D., et al. (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci. Transl. Med. *4*, 136ra168.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Gibson, G. (2011). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145.

Gilissen, C., Arts, H.H., Hoischen, A., Spruijt, L., Mans, D.A., Arts, P., van Lier, B., Steehouwer, M., van Reeuwijk, J., Kant, S.G., et al. (2010). Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am. J. Hum. Genet. *87*, 418–423.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. *20*, 490–497.

Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., et al. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. Nat. Genet. *43*, 860–863.

Girotti, M.R., Pedersen, M., Sanchez-Laorden, B., Viros, A., Turajlic, S., Niculescu-Duvaz, D., Zambon, A., Sinclair, J., Hayes, A., Gore, M., et al. (2013). Inhibiting EGF receptor or SRC family kinase signaling overcomes BRAF inhibitor resistance in melanoma. Cancer Discov. *3*, 158–167.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182–189.

Goate, A., Chartier-Harlin, M.C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature *349*, 704–706.

Haack, T.B., Danhauser, K., Haberberger, B., Hoser, J., Strecker, V., Boehm, D., Uziel, G., Lamantea, E., Invernizzi, F., Poulton, J., et al. (2010). Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. Nat. Genet. *42*, 1131–1134.

Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. *70*, 42–54.

Harrington, C.R., Anderson, J.R., and Chan, K.K. (1995). Apolipoprotein E type epsilon 4 allele frequency is not increased in patients with sporadic inclusion-body myositis. Neurosci. Lett. *183*, 35–38.

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

Hoischen, A., van Bon, B.W., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G., et al. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. *42*, 483–485.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. Am. J. Hum. Genet. Published online May 14, 2013. http://dx.doi.org/10.1016/j.ajhg.2013.04.015.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., et al.; ITALSGEN Consortium. (2010). Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron *68*, 857–864.

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science *315*, 525–528.

Kinde, I., Bettegowda, C., Wang, Y., Wu, J., Agrawal, N., Shih Ie, M., Kurman, R., Dao, F., Levine, D.A., Giuntoli, R., et al. (2013). Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. Sci. Transl. Med. *5*, 167ra164.

Kitzman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C., et al. (2012). Noninvasive whole-genome sequencing of a human fetus. Sci. Transl. Med. *4*, 137ra176.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics *25*, 2283–2285.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.

Krawitz, P.M., Schweiger, M.R., Rödelsperger, C., Marcelis, C., Kölsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat. Genet. *42*, 827–829.

Lalonde, E., Albrecht, S., Ha, K.C., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J., and Jabado, N. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum. Mutat. *31*, 918–923.

Leary, R.J., Sausen, M., Kinde, I., Papadopoulos, N., Carpten, J.D., Craig, D., O'Shaughnessy, J., Kinzler, K.W., Parmigiani, G., Vogelstein, B., et al. (2012). Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci. Transl. Med. *4*, 162ra154.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

Lee, S., Wu, M.C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.

Leone, M.A., Barizzone, N., Esposito, F., Lucenti, A., Harbo, H.F., Goris, A., Kockum, I., Oturai, A.B., Celius, E.G., Mero, I.L., et al.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium 2; PROGEMUS Group; PROGRESSO Group. (2013). Association of genetic markers with CSF oligoclonal bands in multiple sclerosis patients. PLoS ONE *8*, e64408.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M., and Abecasis, G.R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. *8*, e1002944.

Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. *89*, 354–367.

Liu, D.J., and Leal, S.M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. *6*, e1001156.

Lo, Y.M., Corbetta, N., Chamberlain, P.F., Rai, V., Sargent, I.L., Redman, C.W., and Wainscoat, J.S. (1997). Presence of fetal DNA in maternal plasma and serum. Lancet *350*, 485–487.

Lo, Y.M., Chan, K.C., Sun, H., Chen, E.Z., Jiang, P., Lun, F.M., Zheng, Y.W., Leung, T.Y., Lau, T.K., Cantor, C.R., and Chiu, R.W. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci. Transl. Med. *2*, 61ra91.

Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N. Engl. J. Med. *350*, 2129–2139.

Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.

Manolio, T.A. (2009). Cohort studies and the genetics of complex disease. Nat. Genet. *41*, 5–6.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Mardis, E.R. (2013). Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto Calif.) *6*, 287–303.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. *80*, 588–604.

Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.

Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J., et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N. Engl. J. Med. *363*, 2220–2227.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.F., Stevens, C., Wang, L.S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010b). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium. (2012). The life history of 21 breast cancers. Cell *149*, 994–1007.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P.E., Spector, T.D., Hammond, C.J., Richards, J.B., Ciampi, A., and Greenwood, C.M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genet. Epidemiol. *37*, 366–376.

Paez, J.G., Jänne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science *304*, 1497–1500.

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., et al. (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc. Natl. Acad. Sci. USA *101*, 13306–13311.

Pericak-Vance, M.A., Bebout, J.L., Gaskell, P.C., Jr., Yamaoka, L.H., Hung, W.Y., Alberts, M.J., Walker, A.P., Bartlett, R.J., Haynes, C.A., Welsh, K.A., et al. (1991). Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am. J. Hum. Genet. *48*, 1034–1050.

Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R.L., Bardelli, A., and Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature *483*, 100–103.

Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010a). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.

Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010b). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.

Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al.; ITALSGEN Consortium. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron *72*, 257–268.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell *155*, this issue, 39–55.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Scheinfeldt, L.B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., Lambert, C., Jarvis, J.P., Abate, D., Belay, G., and Tishkoff, S.A. (2012). Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. *13*, R1.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. *19*, 212–219.

Sequist, L.V., Waltman, B.A., Dias-Santagata, D., Digumarthy, S., Turke, A.B., Fidias, P., Bergethon, K., Shaw, A.T., Gettinger, S., Cosper, A.K., et al. (2011). Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. Sci. Transl. Med. *3*, 75ra26.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A., and Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. *20*, 273–280.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Swisher, E.M., Wollan, M., Mahtani, S.M., Willner, J.B., Garcia, R., Goff, B.A., and King, M.C. (2005). Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. Am. J. Obstet. Gynecol. *193*, 662–667.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., Whittaker, P., Ranganath, V., Kumanduri, V., McLaren, W., et al. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet. *5*, e1000433.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Timms, A.E., Dorschner, M.O., Wechsler, J., Choi, K.Y., Kirkwood, R., Girirajan, S., Baker, C., Eichler, E.E., Korvatska, O., Roche, K.W., et al. (2013).

Support for the N -methyl-d-aspartate receptor hypofunction hypothesis of schizophrenia from exome sequencing in multiplex families. JAMA Psychiatry *70*, 582–590.

Vilariño-Güell, C., Wider, C., Ross, O.A., Dachsel, J.C., Kachergus, J.M., Lincoln, S.J., Soto-Ortolaza, A.I., Cobb, S.A., Wilhoite, G.J., Bacon, J.A., et al. (2011). VPS35 mutations in Parkinson disease. Am. J. Hum. Genet. *89*, 162–167.

Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.

Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M.K., Thornton, A.M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K.B., King, M.C., and Kanaan, M. (2010). Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am. J. Hum. Genet. *87*, 90–94.

Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., Jin, X., Jiang, H., Zhang, P., Shen, L., Guo, J.F., et al. (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain *133*, 3510–3518.

Wang, S.R., Carmichael, H., Andrew, S.F., Miller, T.C., Moon, J.E., Derr, M.A., Hwa, V., Hirschhorn, J.N., and Dauber, A. (2013). Large Scale Pooled Next-Generation sequencing of 1077 genes to identify genetic causes of short stature. J. Clin. Endocrinol. Metab. *98*, E1428–E1437.

Weedon, M.N., Hastings, R., Caswell, R., Xie, W., Paszkiewicz, K., Antoniadi, T., Williams, M., King, C., Greenhalgh, L., Newbury-Ecob, R., and Ellard, S. (2011). Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease. Am. J. Hum. Genet. *89*, 308–312.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

Wu, X., Wang, L., Ye, Y., Aakre, J.A., Pu, X., Chang, G.C., Yang, P.C., Roth, J.A., Marks, R.S., Lippman, S.M., et al. (2013). Genome-wide association study of genetic predictors of overall survival for non-small cell lung cancer in never smokers. Cancer Res. *73*, 4028–4038.

Xu, B., Ionita-Laza, I., Roos, J.L., Boone, B., Woodrick, S., Sun, Y., Levy, S., Gogos, J.A., and Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat. Genet. *44*, 1365–1369.

Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T., et al. (2005). Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum. Mol. Genet. *14*, 3499–3506.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J.D., Romano-Adesman, A., Bjornson, R.D., Breitbart, R.E., Brown, K.K., et al. (2013). De novo mutations in histone-modifying genes in congenital heart disease. Nature *498*, 220–223.

Zhang, Q., Irvin, M.R., Arnett, D.K., Province, M.A., and Borecki, I. (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. Genet. Epidemiol. *35*, 679–685.

Zhu, X., Feng, T., Li, Y., Lu, Q., and Elston, R.C. (2010). Detecting rare variants for complex traits using family and unrelated data. Genet. Epidemiol. *34*, 171–187.

Zimprich, A., Benet-Pagès, A., Struhal, W., Graf, E., Eck, S.H., Offman, M.N., Haubenberger, D., Spielberger, S., Schulte, E.C., Lichtner, P., et al. (2011). A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am. J. Hum. Genet. *89*, 168–175.