ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies

Samya Chakravorty and Madhuri Hegde

Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322; email: mhegde@emory.edu

Annu. Rev. Genom. Hum. Genet. 2017. 18:229–56

First published as a Review in Advance on April 17, 2017

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083115-022545

Keywords

gene annotation, Mendelian disorders, next-generation sequencing, NGS, massively parallel sequencing, MPS, genome, exome, targeted gene panel, variant, clinical genomics

Abstract

Comprehensive annotations of genetic and noncoding regions and corresponding accurate variant classification for Mendelian diseases are the next big challenge in the new genomic era of personalized medicine. Progress in the development of faster and more accurate pipelines for genome annotation and variant classification will lead to the discovery of more novel disease associations and candidate therapeutic targets. This ultimately will facilitate better patient recruitment in clinical trials. In this review, we describe the trends in research at the intersection of basic and clinical genomics that aims to increase understanding of overall genomic complexity, complex inheritance patterns of disease, and patient-phenotype-specific genomic associations. We describe the emerging field of translational functional genomics, which integrates other functional "-omics" approaches that support next-generation sequencing genomic data in order to facilitate personalized diagnostics, disease management, biomarker discovery, and medicine. We also discuss the utility of this integrated approach for diagnostic clinics and medical databases and its role in the future of personalized medicine.

INTRODUCTION

The last decade has brought unprecedented technological advances in all areas of genomics. The increased ability to understand DNA and its downstream products has opened up new areas of investigation, especially in understanding the basic functional unit of DNA: the gene. Traditionally, basic and clinical sciences have followed parallel paths, with the latter branching off from the former once a discovery is made. Basic science is more free to delve into newer concepts that clinical science typically approaches with extreme caution; however, new technologies are beginning to blur the line between basic and clinical sciences. Moreover, fundamental knowledge of how a single gene affects a single phenotype or disorder is no longer in its infancy. The understanding of genotype-phenotype and gene-disease associations has evolved as research has uncovered evidence of genetic pleiotropy and multigenic effects of disease (to name just a few examples). Knowledge of the complexities of genomic association has vastly increased compared with what was known only a decade ago (**Figures 1** and **2**).

Accurate annotation of genes is critical to understand the locations of the coding and noncoding regions of the genes in the genome and their functional associations with pathways in normal and disease states. Accurate annotation of variants in coding, noncoding, or intergenic regions is important to understand their functional effects, whether disease-causing pathogenic or random genetic drift. It is also important to understand the structural and functional effects of a variant on genomic regions or downstream products of the gene or genes that it regulates. This review attempts to explain gene and variant annotation, to define the categories of annotation, and to

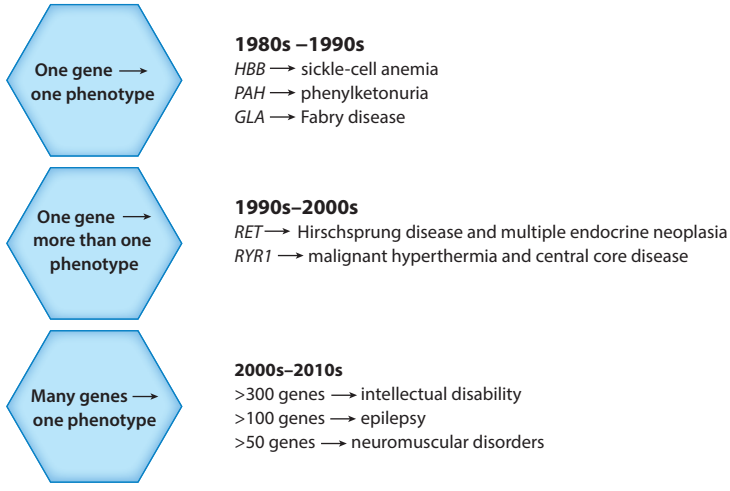

Figure 1

Increasing complexity in genomes: Knowledge of gene-phenotype associations has been increasing, leading to a better understanding of genomic associations of diseases, genomic complexities, and genome annotation. In the 1980s and 1990s, the basic understanding was limited to how a single gene affects a single phenotype—for example, the *HBB* gene in sickle-cell anemia, the *PAH* gene in phenylketonuria, and the *GLA* gene in Fabry disease. In the 1990s and 2000s, the concept of pleiotropy began to emerge, in which a single gene can be associated with multiple disease phenotypes—for example, the *RET* gene can cause both Hirschsprung disease and multiple endocrine neoplasia, and the *RYR1* gene is associated with both malignant hyperthermia and central core disease. This paradigm has shifted even further in the 2010s, and we now know of sets of genes that are associated with various diseases that have heterogeneous phenotypic representations—for example, more than 300 genes are associated with intellectual disability, more than 100 genes with epilepsy, and more than 50 genes with neuromuscular disorders.

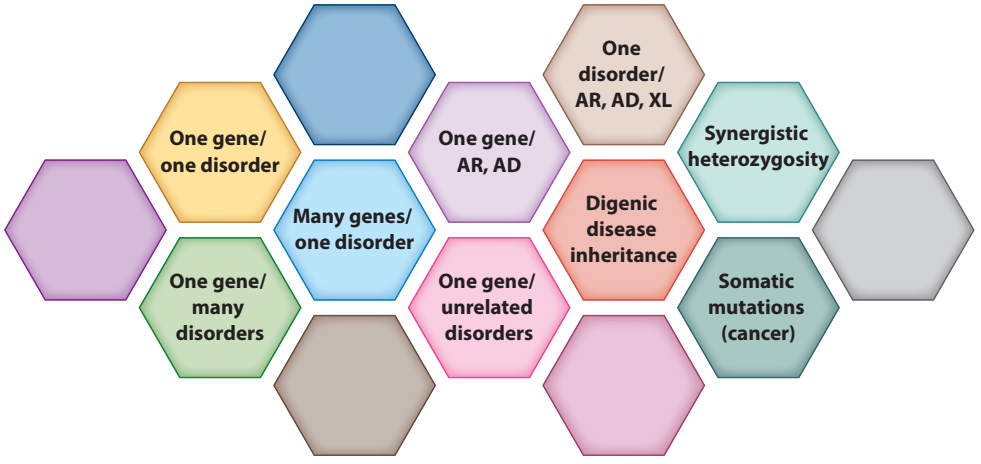

Figure 2

Growing genomic complexity changing the fundamentals: Fundamental genomic understanding is evolving as knowledge of the complexities of the human genome increases. The initial concept that one gene causes one disorder has given way to the idea that one gene can affect many disorders and that one disorder can result from multigenic effects. We now know that gene-gene interactions in the same or different pathways can cause a gene to be associated with an otherwise unrelated disease. Inheritance patterns are also much more complex than originally thought. For example, one disorder can be inherited autosomal dominantly (AD) or autosomal recessively (AR) or can be X linked (XL). Studies have uncovered digenic and multigenic inheritance of various diseases, and new evidence is pointing toward synergistic effects of heterozygous variants in different genes that may affect disease phenotypes. In addition, next-generation sequencing, exome sequencing, and genome sequencing are revealing an increasing number of de novo mutations that are associated with diseases such as cancer, broadening our understanding of genomic complexities and the opportunities to discover biomarkers and therapeutic targets.

describe new concepts that have led to a merger of basic and clinical genomics, especially regarding the new functional modalities of understanding genomic variants and their annotation. The specific areas discussed in this review are the assembly of human reference genome sequences, the impacts of research on gene structure and function since the completion of the human genome sequence, next-generation sequencing (NGS), DNA sequence variation and annotation, Mendelian and complex diseases (and those that fall between these categories), the expansion of knowledge of the phenotypic spectrum of diseases caused by individual genes, the discovery of new disease genes, gene-gene global networks, and the emerging area of integrated functional genomics for diagnostics, biomarkers, and clinical trials.

HUMAN REFERENCE GENOME SEQUENCE ASSEMBLY

Accurately and efficiently comparing next-generation genomic and transcriptomic data from different consortiums, identifying truly functional variants, and interpreting the data require proper reference material, which should encompass known variation in human genomic regions. For this purpose, many efforts have been made to create comprehensive human reference genome sequences.

The first sequencing of a human genome using NGS technology [known at the time as massively parallel sequencing (MPS)] was performed in 2008 using James Watson's DNA (196). This was the first step toward the development of technologies that were faster, cheaper, and more efficient

than shotgun sequencing methods. NGS avoided the loss of genomic regions that resulted from shotgun sequencing and bacterial cloning, providing further insights into the single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and copy number variants (CNVs) in the human genome. This revolutionary work has also provided functional insights into human genome variability among individuals, susceptibility to disease, and ultimately clinical molecular diagnostics and therapy, leading to the emergence of personalized medicine.

In December 2013, the UCSC Genome Browser (**http://genome.ucsc.edu**) released its hg38 (or GRCh38) human genome reference assembly, which introduces significant changes compared with the previous hg19 (or GRCh37) assembly. [UCSC Genome Browser assembly IDs were previously numbered sequentially from hg1 to h19; this latest release is numbered hg38 in order to match the IDs used by the Genome Reference Consortium (GRC).] These new features are as follows:

- Most notably, because the significant variability of several human chromosomes prevents accurate representation by a single reference sequence, the hg38 assembly provides alternate sequences for variant regions by including alternate locus scaffolds at 261 loci. These loci are concentrated in the leukocyte receptor complex/killer immunoglobulin-like receptor region of chromosome 19 and the major histocompatibility locus region of chromosome 6.
- The gaps in the centromeric regions in the hg19 assembly have been filled using centromere databases, which will be useful for read mapping and variation analyses.
- The hg38 assembly provides an updated mitochondrial reference sequence.
- Erroneous and misassembled regions of the hg19 assembly have been corrected, and gaps have been filled using data from other genome sequencing (GS) projects, such as the 1000 Genomes Project (**http://www.1000genomes.org**).
- The hg38 assembly provides better analysis sets that meet the needs of NGS alignment pipelines, with several regions removed to facilitate better alignment and mapping.

Correctly annotating variants from different sequencing data sets worldwide requires highly accurate genotype sets across the human genome. To accomplish this goal, the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIAB) Consortium (**http://genomeinabottle.org**) was established to develop the technology and reference standards, methods, and data to translate human GS into clinical practice by providing well-curated, annotated reference genome sequences. The consortium's main goals are to characterize the human genome in order to validate genomic variants and to develop optimization technology in order to create better reference genotypes.

The consortium used its pilot genome (NA12878) to validate methods to accurately call SNPs, indels, and homozygous reference genotypes. Fourteen data sets across five sequencing technology pipelines were used to validate the methods for reducing bias for variant calls and to estimate the confidence of the reported characteristics (210). The consortium uses Genome Analysis Toolkit methods to integrate multiple data sets from different technologies and platforms on the same genome sequence and uses variant quality score recalibration to identify possible calling biases and create a consensus among discordant data sets. Using multiple platforms not only assesses the platforms' efficiencies, but also will allow annotation of low-coverage areas and enable accurate genotype calls that might be missed when using only one platform. All of the resulting methods, reference materials, and genotype calls from the integrated approach are publicly available from the GIAB website (**http://genomeinabottle.org**), allowing investigators to evaluate the performance of specific sequencing platforms.

Even for data from platforms with high sensitivity and specificity, the GIAB Consortium cautions that alignment around a subset of discordant genotype calls must be examined by using,

for example, the new Genetic Testing Reference Materials Coordination Program (GeT-RM) browser for the NA12878 genome (**https://www.ncbi.nlm.nih.gov/variation/tools/get-rm**), and manual inspection of these regions is always recommended. Structural variants (SVs) are large (generally >1 kb) inversions, translocations, and indels, together known as CNVs, which enhance the complexity of the human genome. To characterize SVs and assess larger indels in order to create SV benchmarks, Parikh et al. (141) developed an integrated approach that combines multiple methods, which they called svclassify. They found that, in this integrated approach, SVs with high scores from multiple technologies agreed well with polymerase chain reaction (PCR) validation and an orthogonal consensus method (MetaSV) with 99.7% concordance, whereas SVs with low scores did not, which gives confidence in the approach and the reference material. More recently, using as many as 12 validated NGS, library preparation, and analysis pipelines, Zook et al. (209) began developing well-annotated human reference genomes as benchmarks from two family trios of different ancestries (Ashkenazi Jewish and Chinese) and a pilot genome of a European ancestry individual. This work is part of the GIAB Consortium's effort to validate and create authentic benchmark reference materials; it is unique in that it includes genomes from different populations and therefore is expected to improve not only sequencing technologies, but also variant calling of SNPs, indels, and SVs as well as de novo assembly.

Using 769 individual genomes from 250 Dutch families, Hehir-Kwa et al. (69) recently created a high-quality human reference genome panel that yielded novel, underreported, and complex midsize SVs (between 21 and 100 base pairs)—in particular, complex indels and retrotransposition-mediated insertions of mobile elements and processed RNAs—as well as their distribution across the genome. The authors focused on comparing large family-based sequences with sequences from substantially unrelated individuals and on having sufficient coverage (14.5× median base coverage and 38.4× median physical coverage) for genotyping and phasing the full spectrum of SVs in order to create a high-quality reference panel. Interestingly, in this study, downstream variant analyses predicted that the distribution and functional impacts of rare and common variants are significantly different, suggesting the importance of population frequency in understanding the clinical significance of variants, at least as a first step in the annotation process. Global efforts are ongoing to create benchmark reference materials for better unbiased calls of variants of any size, not only by using a large number of individual genomes from different global populations and ethnicities, but also by including more methods in an integrated approach to further reduce biases and increase confidence in annotation.

The enormous amount of NGS data worldwide has made it imperative to bioinformatically annotate genomic information into well-represented uniform tracks that will be flexible enough for continuous genomic development and simplify data parsing (62, 164). Even with such annotation, it is difficult to represent a reference genome using a single sequence because specific sequences may contain an individual's unique genomic regions that remain unmapped. As discussed above, to overcome this problem, the hg38 assembly includes annotated alternate sequence loci in order to represent regions that are too complex for a single sequence (see 56). The alignment software is also being updated and new software created to tolerate the alternate loci and facilitate better, more flexible, and more comprehensive mapping, especially in high-complexity genomic regions, such as the immunoglobulin heavy-chain locus (105, 175, 195). But the real challenge of an unbiased reference assembly approach will be the move toward a reference-free assembly. Such a move may be feasible by using multitrack reference information, such as graph-based representations of reference assemblies, string graphs, de Bruijn graphs, and information from the Global Alliance for Genomics and Health (**http://genomicsandhealth.org**) (29, 34, 122).

NEXT-GENERATION SEQUENCING AS A DISRUPTIVE TECHNOLOGY FOR GENE ANNOTATION

Genomics has evolved rapidly since the publication of the first human genome sequence in 2001 (75, 96, 186). Modern scientific development (both research and clinical) is highly dependent on technological advancement. Several new technologies have emerged since 2001; among these, NGS has enabled human genomes, exomes, and gene panels to be sequenced much more quickly and cost-effectively. To facilitate this, approaches to bioinformatic analysis have developed along with NGS and are an integral part of the pipeline. These developments have revolutionized genomics as well as personalized diagnostics and medicine—so much so that Church et al. (30) have proposed that DNA could be a highly efficient data storage medium in the future. Throughout the last decade, the capacity of NGS technology has increased by a factor of 100–1,000 (91); at the same time, its costs have come down considerably, to approximately US$1,000 per genome, facilitating the translation of sequencing from a research technology to a clinical tool for diagnostics and therapeutic management (125, 187, 188). With the new NGS technologies emerging and constantly evolving, the original term, MPS, is coming to the forefront and is often used interchangeably with NGS.

Several comprehensive reviews have described technical advances in NGS (59, 104, 118, 121, 158). Sequencing technology platforms differ in their extent of coverage of the human genome and are classified into three major types. The first type comprises the detection of clonally amplified target DNA (used by Illumina and Ion Torrent platforms) and single-molecule detection per reaction (used by Pacific Biosciences and Oxford Nanopore platforms). The second type comprises sequencing by synthesis (used by the Applied Biosystems SOLiD platform for polymerase-based synthesis and the Polonator platform for ligation-based synthesis) and direct measurement of DNA (used by Illumina, Ion Torrent, and Pacific Biosciences platforms). The third type comprises base read calls through either optical detection (used by Illumina and Pacific Biosciences platforms) or nonoptical detection (used by Ion Torrent and Oxford Nanopore platforms). Recently, different platforms have also been used in hybrid setups that can take advantage of the strengths of each platform (93). However, along with these technological advancements and enormous data collection pipelines come the disadvantages of high error rates (0.1–15%) and shorter read lengths (35–700 base pairs) (109). NGS technologies also compete with more targeted (but also potentially biased) and cost-effective but time-consuming technologies, such as DNA microarrays, NanoString tools, quantitative PCR, optical mapping, and even Sanger sequencing.

Short-read NGS generates clonal template DNA populations using bead-based, solid-state (44, 90, 98, 169), and DNA nanoball generation methods (45). In this approach, after template enrichment, parallel sequencing is performed using (*a*) sequencing by synthesis, either through a Roche 454 or Ion Torrent single-nucleotide addition (SNA) platform or through an Illumina or Qiagen cycle reversible termination (CRT) platform, or (*b*) sequencing by ligation, through a SOLiD or Complete Genomics (a BGI subsidiary company) platform. SOLiD uses the detection of a dinucleotide utilizing cleavable two-base-encoded fluorescent probe signals to achieve genome-wide mapping (184). By contrast, Ion Torrent SNA platforms represent the first NGS technology to use non-optical sensing (163); this platform utilizes a massively parallel semiconductor device to monitor H^+ ion release during DNA synthesis for bacterial and human GS. The Complete Genomics platform uses combinatorial probe-anchor ligation or combinatorial probe-anchor synthesis for human GS (6, 19, 45). Similarly to Sanger sequencing, CRT uses terminator molecules that block the ribose 3′-OH, thereby preventing elongation (64, 86). The Illumina CRT system has the largest market share of all NGS commercial platforms to date, in large part because of its versatility; it includes platforms ranging from low-throughput small

benchtops to ultra-high-throughput units for population-wide GS (180). In 2015, Qiagen acquired and launched its Intelligent Bio-Systems CRT platform under the name GeneReader; this is the first all-in-one NGS platform that can perform all steps from sample preparation to analysis and final data generation (88). Qiagen achieved this by combining the QIAcube sample preparation system and the Qiagen Clinical Insight platform for variant calling and analyses.

All of these platforms vary in their error rates, costs, throughput, and read structure (for detailed comparative reviews, see 59, 104). Overall, Illumina instruments are used more widely than any other NGS technology (180), but such broad use of a single technology may introduce systemic bias, especially in variant identification (131, 160, 210). Illumina platforms range from the low-throughput MiniSeq (25 million reads per run with a maximum output of 7.5 Gb) to the ultra-high-throughput HiSeq X (6 billion reads per run with a maximum output of 1,800 Gb), which can sequence ~18,000 human genomes with 30× coverage in a year. The HiSeq X is the highest-throughput instrument currently available, but its use is restricted mainly to GS and bisulfite mapping.

Different platforms have their own advantages and limitations. For example, the SOLiD and Complete Genomics platforms are highly sensitive (~99.99%) (45, 109) but are not as specific as other platforms, which leads to false positive and false negative variant calls (26, 168, 191) and underrepresented AT- or GC-rich genomic areas (65, 160). The most prominent disadvantage of these two platforms is their very short read length—approximately 75 base pairs for SOLiD and 28–100 base pairs for Complete Genomics (18)—which limits their use in calling SVs. Because the Illumina platform uses CRT, it is much less susceptible to homopolymer errors (99.5% accuracy) (17), demonstrating that reversible dye-terminator chemistry can be used in human GS. The Illumina platform is allowing a groundbreaking range of sequencing applications, including GS and exome sequencing (ES), epigenomic sequencing by ChIP-seq (sequencing after chromatin immunoprecipitation), ATAC-seq (assay for transposase-accessible chromatin using sequencing to identify enhancers), methyl-seq (DNA methylation sequencing), and transcriptomic high-throughput sequencing by RNA-seq (RNA sequencing) (22, 23, 142, 193). Ion Torrent has launched customized chips and instruments for researchers and clinicians that are faster than other platforms and yield a throughput range from approximately 50 Mb to 50 Gb.

It is clear that these new technologies are blurring the lines between basic genomic research, technological advancement, and clinical genomics. They are facilitating the use of targeted approaches for focused clinical and research applications, such as gene-panel sequencing, targeted transcriptome profiling, and splice-site identification (106, 114). In fact, Ion Torrent is moving forward in clinical sequencing with the launch of its Ion Personal Genome Machine (PGM) Dx and Ion S5 series diagnostic instruments, which aim to provide simpler and more user-friendly platforms. Interestingly, the Ion PGM Dx sequencer supports paired-end reads (76), but the high-throughput S5 devices lack that feature, making it difficult to use them for long-range genomic and transcriptomic structural analyses (25).

One of the ongoing challenges of NGS is to develop an efficient technology for long-read sequencing that will be high throughput and sensitive enough to capture SNPs, indels, and SVs, including disease-related repetitive sequences and CNVs, in order to more fully elucidate the complexities of the human genome (116, 120, 174). This will also be most critical for understanding the transcriptome landscape at a functional level and providing a better picture of exon usage or isoform patterns and gene expression. There are two major approaches to long-read sequencing: single-molecule real-time sequencing and a synthetic approach that relies on short reads and computationally constructs long reads. The single-molecule approach is currently offered by Pacific Biosciences and Oxford Nanopore. Unlike other platforms, the MinION from Oxford Nanopore, rather than relying on a secondary detection method, detects DNA sequences

directly using single-molecule nanopore DNA sequencing in a high-throughput manner; it comprises 512 individual channels capable of sequencing approximately 70–500 base pairs per second in an application-specific integrated circuit chip (32). Although it sequences longer reads, the promising new PromethION platform from Oxford Nanopore may challenge the throughput of the Illumina HiSeq X. Currently, however, the Pacific Biosciences RS II instrument is the most efficient at sequencing long reads, generating single-polymerase reads of more than 50 kb with average read lengths of 10–15 kb for long-insert libraries, which is ideal for use in clinical de novo genome assembly applications, studies of long-range genomic structures, and full-length transcriptome profiling (47, 166).

Ultimately, NGS platforms are being used more widely to sequence whole genomes in order to discover variants, especially rare variants in human diseases and their associated biological functions (31); one important example of this is the 1000 Genomes Project, but there are many others (1, 2, 61, 154, 178, 182). In pediatric medicine, GS has a diagnostic yield that is four times that of chromosomal microarrays and twice that of conventional targeted gene sequencing, allowing the identification of multigenic variants and SVs and better clinical diagnostics. In terms of the merging of NGS technology and clinical genomics, however, it is ES and targeted sequencing that are increasing the number of samples being sequenced by looking at focused, disease-specific genomic regions, which ultimately increases the spectrum of research and clinical genomic studies (73).

NEXT-GENERATION SEQUENCING GUIDELINES: QUALITY CONTROL, SEQUENCE ALIGMENT, REFERENCE SEQUENCES, AND GENE AND VARIANT ANNOTATION

Quality Control

The American College of Medical Genetics and Genomics (ACMG) and the US Centers for Disease Control and Prevention (CDC) have laid down guidelines for NGS technology and informatics pipelines so that they can be used reliably and efficiently in clinical work. Here, we discuss a few major recommendations made by the CDC-organized Next-Generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) workgroup in 2012 (52) and the ACMG in 2013 (155). These guidelines address test validation, quality control, proficiency testing, and reference materials based on Clinical Laboratory Improvement Amendments (CLIA) requirements. Comprehensive guidelines have also been published for the use of NGS in clinical microbiology and public health laboratories (54), which are largely similar to the CDC and ACMG guidelines for the use of NGS in clinical molecular genetic disease diagnostics. NGS has diverse applications in clinical microbiology and public health, including GS, microbiome analysis and metagenomics, transcriptome profiling, infectious disease diagnosis, pathogen discovery, and public health surveillance. In this context, use of the appropriate reference material (for example, a reference bacterial strain) is critical for test development, validation, quality control, and proficiency testing and should resemble patient samples as closely as possible. In 2015, the Nex-StoCT workgroup published additional recommendations for NGS informatics pipelines (53).

The CDC and ACMG guidelines recommend that the platform, the particular test, the informatics pipelines, and (if required) alternative methods such as Sanger sequencing be validated by laboratories. Combinations of quality control materials that explain genomic complexity should be used. Quality metrics such as scores, depth, coverage uniformity, mapping quality, and GC bias should be compared with those obtained during validation. Proficiency testing should be performed using both disease-associated and naturally occurring genomic variations targeted by

the test in order to measure sequencing reliability. The ACMG recommends additional merging and collaboration between research and clinical genomics laboratories in order to facilitate the discovery of novel disease-causing candidate genes (155), especially in the case of ES that is used to investigate both known disease-causing and new genetic associations. Test limitations, including low sequence coverage, absent data, and ambiguous variant calls, should be tracked and defined in reports. Because targeted NGS panels analyze focused disease-related genomic regions and therefore have a higher read depth, analytical sensitivity, and analytical specificity, testing should be initiated with an NGS panel prior to ES or GS.

ES and GS can enable a broader, more discovery-driven approach to understanding patient phenotypes that requires more collaboration between research and clinical laboratories and end-point healthcare providers for variant interpretation and diagnosis. ES and GS provide a higher definitive diagnostic yield when both a child and that child's parents are sequenced (trio sequencing) than they do when only the child is sequenced (singleton sequencing) because of the ability to identify segregating parental alleles and the higher accuracy in variant calling, including in low-coverage genomic regions; this leads to more definitive results of molecular diagnoses, especially for newborns and children. Moreover, incidental findings such as carrier status should be evaluated carefully for potential relevance to the patient phenotype before such findings are reported to the patient. Physicians should provide detailed clinical notes and phenotypes so that the laboratory can perform context-dependent interpretations of the relevant variants. In a targeted NGS panel, only genes for which there is sufficient scientific evidence of a particular disease association should be interpreted, and the efficiency and limitations of the targeted capture method should be reported. For diseases with high genetic heterogeneity, ES or GS may be more efficient than targeted gene-panel testing, but the limitations of gene inclusions or coverage should be reported. The use of confirmatory or supplementary technologies (such as Sanger sequencing) is often recommended, especially for ES when causative variants are in noncoding regions or are SVs.

To ensure that a sequencing run is of sufficient quality, analysis at intermediate points during and after the sequencing run can be performed to evaluate real-time errors, target capture and aligned read percentage, and duplicated read percentage and to estimate the coverage depth. The performance parameters of the analysis pipeline should include the analytical sensitivity of false positive and false negative rates, predicted clinical sensitivity and assay robustness, and reproducibility. Importantly, to evaluate the clinical sensitivity and diagnostic yield of the test, the success rates of the test across different disease areas should be tracked and shared. Vendor-supplied indexes should be used to index NGS samples, but for indexes that are prepared in-house, the Nex-StoCT workgroup recommends using design parameters such as index length and composition in order to reduce read misassignments to incorrect samples. The size of the sequenced region, required coverage depth, sample volume, turnaround time, and costs should be considered when choosing a sequencer and corresponding platform.

Sequence Alignment

More than one read alignment tool should be used, depending on the type of variations expected based on the patient disease phenotype, and the alignment should be to the whole reference genome for all types of NGS in order to reduce mismapping of reads caused by off-target captures in ES or gene-panel tests. To accurately identify indel variants, a local realignment should be performed after a global alignment (40). In terms of reporting data, NGS data as sequence reads and the read alignments should be in the .fastq and .bam formats, respectively, and variant calls should be in the .vcf format, as was done by the 1000 Genomes Project. Similarly to the ACMG, the CDC recommends including local realignments in the initial analysis to remove PCR duplicates

in ES or GS and recalibrating base call quality scores to improve variant calling (53). For this purpose, the CDC strongly recommends using multiple variant caller and/or parameter settings. The reference values from the recent NIST-released reference material (RM 8398, obtainable from **http://www.nist.gov/srm**) can be used to assess the performance of variant calling from GS.

Reference Sequences

Suitable reference materials should be established and updated as technology develops and should reflect annotated regions of high and low sequence reliability. Because reference materials are constantly being updated, when aligning the reads to the reference genome assembly, the Nex-StoCT workgroup recommends noting the accession date and data version of the reference material used for each alignment so that variant positions can be tracked back in the reference (53). Analysis tools and software in the pipeline should similarly be updated or new ones developed.

Sequencing is often used to identify the source of infections, since it is faster and more sensitive than traditional culture methods. To address the need to develop reference materials for a variety of pathogenic organisms relevant to public health and clinical laboratory settings, NIST has chosen strains relevant to food safety and clinical microbiology NGS applications that represent diverse genome sizes and a range of plasmid and GC contents (136). Successful pathogen identification and discovery of novel genes and variants require well-curated public databases that contain accurately annotated reference materials from relevant organisms (bacteria, fungi, virus, yeast, and parasites), providing a benchmark for true diversity of both new and old strains. Reference materials can still be biased because many organisms are rare, and some pathogens, such as Ebola and Zika virus, could be a higher priority than others. To address this issue, the US National Center for Biotechnology Information developed and maintains the Reference Sequence (RefSeq) database (**http://www.ncbi.nlm.nih.gov/refseq**), a collection of well-annotated, taxonomically diverse genetic and protein sequence records constructed from sequence data from the International Nucleotide Sequence Database Collaboration. For infectious diseases, the US Food and Drug Administration's Database for Regulatory Grade Microbial Sequences contains regulatory-grade microbial genomic sequences that encompass the diversity of clinically and environmentally relevant circulating microbe strains.

Gene and Variant Annotation

In targeted NGS panels, as new candidate genes are added, it is important to develop automated classification tools to differentiate common benign variants from rare deleterious variants, possibly by using databases of population frequencies such as dbSNP (**https://www.ncbi.nlm.nih.gov/projects/SNP**); the US National Heart, Lung, and Blood Institute's Exome Sequencing Project (**http://evs.gs.washington.edu/EVS**); the Broad Institute's Exome Aggregation Consortium (ExAC; **http://exac.broadinstitute.org**); and the 1000 Genomes Project. Laboratories should employ a frequency cutoff that is higher than the theoretical maximum in order to account for population-specific statistical variance of undocumented and reduced penetrance and undiagnosed nonphenotyped patients in the populations. Investigators should be cautious when carrying out this process because many databases include misclassified variants, particularly benign variants classified as pathogenic (16, 43).

In variant interpretation of ES or GS, one can assume that variants of Mendelian diseases are rare and highly penetrant (113), but further strategies for variant and gene filtering should be employed. A step-wise approach should be used, and the filtering criteria should be flexible enough

to reduce variant filtering bias or errors so that causative variants are not missed. In fact, variant calling performance should be evaluated by analyzing sequences with known variants of different types and sequences from samples of different sources. Even with paired-end reads, problems in sequencing homology and repetitive regions can be addressed by using local realignment after global alignment in order to map the regions correctly. When identifying and interpreting clinically relevant variants, variants in genes that are not relevant to the patient's clinical phenotypes should be filtered but not entirely removed from final reporting. The caveat to such variant filtering is potential incidental findings, which should be reported based on ACMG and Association for Molecular Pathology (AMP) guidelines (60, 68). In the filtering process, one can determine which variants to remove by using population-wide minor allele frequency, predicted effect on protein function or splicing, and disease-variant databases such as the Human Gene Mutation Database (**http://www.hgmd.cf.ac.uk**), ClinVar (**http://www.clinvar.com**), and Online Mendelian Inheritance of Man (**https://www.ncbi.nlm.nih.gov/omim**). However, the Nex-StoCT workgroup also recommends caution when using such databases and other tools because some variants called can be false positives or false negatives for disease associations owing to insufficient curation. During this process, one must ask whether the variant disrupts gene function consistent with disease mechanism, whether it leads to or predisposes the patient to any health issue, and whether this health issue is relevant to the patient's clinical phenotype and the NGS test result interpretations.

Clinical laboratories should share their NGS variant data sets to determine the consistency of variant calls in order to integrate data into medical databases and ultimately into patient health records. To make this sharing more effective and consistent, the Nex-StoCT workgroup recommends the development of a new clinical-grade variant file format that will be compatible with the changing health technology information framework. In clinical microbiology and public health, variant calling methods are similar to human genetic testing, and best practice guidelines for both microbial and human genomes variant calling are being documented (136).

Figure 3 shows our proposed structural framework for an integrated approach that uses not only the relevant database information and proper curation but also a broad "-omics" approach to decipher the functional effects and associations of genes in order to correctly annotate them for disease. We predict that this framework will further facilitate the merging of basic and clinical genomics and ultimately will be of high clinical utility in a diagnostic setting.

NEXT-GENERATION SEQUENCING: LOOKING AT THE FUTURE

NGS technology is not only providing researchers and clinicians with deep information about the human genome, but also going beyond genomics. For example, using NGS technologies such as ChIP-seq, ATAC-seq, and methyl-seq in epigenomic studies is enabling investigators to find genomic regulatory mechanisms, snapshots of protein-DNA interactions, enhancer regulation mechanisms, methyl modifications of the genome, and variants of all of these mechanisms in disease states (23, 77, 117, 134, 142, 153). In transcriptomics, researchers continue to use the power of NGS to deep sequence RNA down to the single-transcript level, which is relevant in clinical genomics—e.g., in identifying variant cryptic splice sites, insertion of pseudoexon sequences, downstream frameshifts, emergence of premature stop codons, allele-specific expression, nonsense-mediated decay, differential exon usage, or transcript abundance. For this purpose, new approaches are being developed for single-cell RNA sequencing in order to characterize different cellular populations, with the functional goal of discovering specific biomarkers (181). One challenge is to understand the specific transcript abundance using long-read sequencing technology, but this technology can certainly provide a picture of the transcriptome structure and differential and/or novel isoform patterns as well as a global picture of exon usage (167).

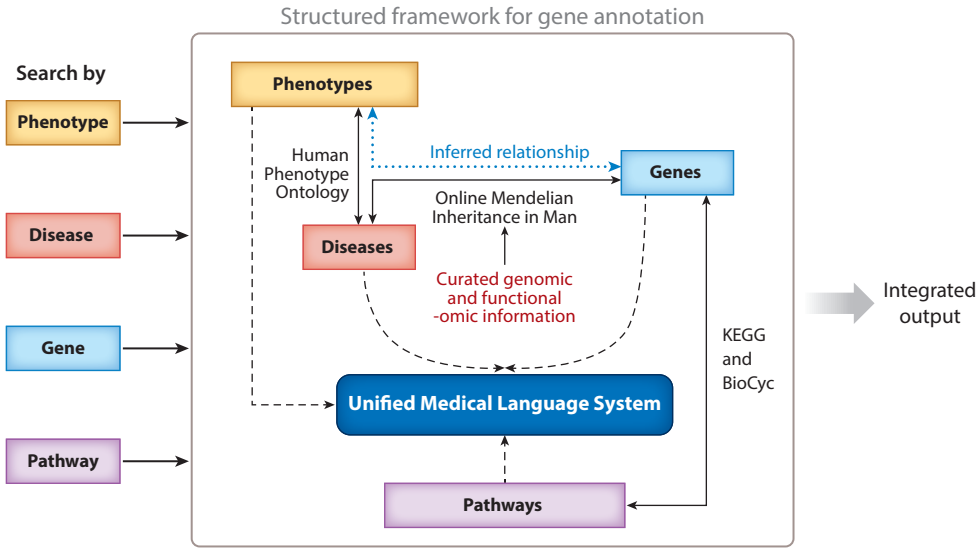

Figure 3

Gene annotation in a structured framework: Dealing with increasing knowledge of genomic complexity requires a structured framework of gene annotation. This representation shows a simplified structure by which genes can be annotated using an integrated approach and their disease associations and functions (*outer gray box*) can be searched by the medical community using patient phenotype, the disease diagnosed, the gene itself, or the possible pathways that might be affected. Genetic and phenotypic relationships can be inferred (*blue dotted arrow*) but are not confirmatory. Curated genomic information, functional "-omics" studies (transcriptomics, proteomics, and metabolomics), and the Online Mendelian Inheritance in Man database at the downstream functional levels should give specific and sensitive information about genetic associations with diseases, from which predicted and observed phenotypes can be matched using the Human Phenotype Ontology database (*solid arrows*). Moreover, pathway analyses using databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioCyc should also be used for information on genes that may be involved in a disease phenotype (*solid arrow*). All of this information should be compiled in the Unified Medical Language System to generate consistent data and terminologies for gene annotations (*dashed arrows*), which will lead to an integrated output for clinical diagnostics and genomic medicine.

NGS tools have led to groundbreaking results in translational human cancer genomics. Both genome and targeted approaches have been used to detect molecular biomarkers, allowing a merging of research and clinical cancer therapeutics (4, 5, 70; for reviews, see 135, 190, 197). Combining the vast, robust NGS data on cancer from numerous publications and consortiums with the numerous cancer research model tools, such as cancer cell lines, should help in categorizing cancer patients for enrollment in clinical treatment and trials (55). NGS can be used to screen large populations of children for cancer susceptibility by using data from GS and ES to catalog germline mutations in genes that may be involved in cancer predisposition, allowing better disease management and precision medicine for those individuals (206). In particular, targeted approaches (such as NGS of disease gene panels) and ES are becoming highly efficient and informative for precise molecular diagnosis of rare human diseases (11, 28, 127, 128, 198). The clinical diagnostic yield of ES is only 25% of selected cases, but its exceptional rate of finding de novo mutations by trio sequencing across both rare diseases and some more common diseases, such as autism, is changing our understanding of these diseases and enabling discovery of new disease genes, ultimately pushing toward functional studies to test multigenic effects on disease pathogenicity and patient phenotype (95, 165, 200, 201).

We are moving into a genomic era in which the type of NGS technology to be used—be it GS, ES, or gene-panel sequencing—will depend on a clinician's diagnosis of a patient's possible disease(s). Narrowing down the best option for molecular diagnosis and genomic medicine will then require the combined efforts of researchers, medical geneticists, and clinicians. Several studies have found that GS has a better diagnostic yield than ES based on overall variant calling sensitivity and efficiency (lower coverage depth required for similar sensitivity), lack of bias, uniformity of coverage, and reduced bias in detecting nonreference alleles (101, 119). However, one study compared the performance of four commercially available exome capture tools with an augmented exome strategy that provided enhanced coverage of a set of 56 medically relevant genes and found that the latter has a higher variant calling sensitivity compared with traditional GS or ES, pointing toward the utility of targeted approaches (144). Importantly, Dewey et al. (42) reported that GS provided incomplete coverage of inherited disease genes, with low reproducibility in detecting pathogenic variants with the largest clinical effects, as determined from the clinical literature. This result further strongly suggests that researchers, medical geneticists, and clinicians should collaborate more. At the same time, in order to reduce the burden of false positive and false negatives, collaborators should be cautious about carefully evaluating the technical performance of the technologies and capture tools and the analytical performance of the platforms before deciding on a particular clinical diagnostic assay. More importantly, the use of NGS at a functional level (RNA-seq) in conjunction with an integrated method that can detect epigenomic patterns of disease and -omics technologies that use mass spectrometry should be critically considered to ensure that the approach is appropriate for the genomic architecture of the patient (57). This should also be done while keeping in mind the available treatment options and the potential for novel therapeutic discoveries that may accelerate precision medicine screening and treatment.

More emphasis on using NGS at the functional and regulatory levels (transcriptomics and epigenomics, respectively) will allow a better understanding of the complexity of the human genome, its modifications, and its downstream products (**Figure 3**), especially genotype-phenotype associations at the single-cell level with high resolution. Angermueller et al. (8) recently reported a new method called scM&T-seq (single-cell genome-wide methylome and transcriptome sequencing) that can yield important insights into regulatory epigenomic mechanisms and gene expression patterns at the same time in a single cell. Using mouse embryonic stem cells, the authors discovered previously undetected associations between heterogeneously methylated regulatory elements and the gene expression of important pluripotency genes, enhancing the spectrum of effects and our understanding of the epigenome.

Macaulay et al. (112) developed a method called G&T-seq (genome and transcriptome sequencing) that enables parallel sequencing of the genome and transcriptome in a single cell in order to elucidate genotype-phenotype relationships. This is an important development because, unlike other techniques for parallel DNA and RNA sequencing, G&T-seq obtains bead-based physical separation of the cell's DNA and RNA without using a bespoke microfluidics platform, and the process can be automated for high throughput. The sequencing is done in the Illumina HiSeq X platform. It is important to note that the coverage is not uniformly distributed across the genome and shows GC bias, which indicates that the technologies for parallel high-throughput single-cell genomics, although advancing quickly, are still in their infancy.

EVOLVING CONCEPTS IN GENE STRUCTURE AND FUNCTION

Because of its advantage in overall genomic coverage, as NGS moves from targeted approaches such as ES and gene-panel tests to GS, it is important to understand the structures of genes and the regulatory elements that can lie in both intra- or intergenic regions. Most nonmicrobial genes

contain both exons and introns, but some exceptions, such as *SOX3*, have a single-exon structure (15). In addition, 5′ untranslated regions (UTRs) can have open reading frames and are translated to regulate coding sequence expression (9). Indeed, pathogenic variants that cause disease are present in 5′ UTRs, such as *FMR1* CCG expansions that cause fragile X syndrome and an *HTR3A* upstream variant that causes bipolar affective disorder (130, 185). Similarly, pathogenic variants can be present in regions of a gene other than the coding sequence. For example, an *SCN1A* 3′ UTR variant leads to reduced mRNA stability, which in turn causes Dravet syndrome (204). Variants at the microRNA sites in the *PAX6* 3′ UTR lead to rolandic epilepsy (139), and microRNAs generally have an important role in regulating protein levels in epilepsy, which suggests potential therapeutic targets (71, 157). Promoter region variants can disrupt normal gene expression and are associated with autism spectrum disorder (27). Splicing variants in exonic or intronic region can cause alternative transcripts in which either exons are skipped or longer or shorter exons are formed that can misbalance the isoform pattern of genes, which in many instances leads to pathogenicity, such as X-linked intellectual disability, epilepsy, primary microcephaly, breast and ovarian cancer caused by the *BRCA1* locus, cardiac abnormality, and variable myopathy caused by the *LMNA* locus (33, 48, 49, 152, 162).

Variants in other structural aspects of genes can also be important for gene function and have regulatory roles in expression. Examples include the poly(A) variant in *ARSA* that causes metachromatic leukodystrophy (46) and the posttranslational modification variant in the SH2 domain of *PTPN11* that inhibits SH2 phosphorylation, which disrupts the auto-inhibitory structural loop and causes Noonan syndrome (156). Moreover, because of splice variations, downstream frameshifts can also occur in the coding sequence of a gene that may introduce a premature stop codon that is toxic for the system if translated. Importantly, these erroneous transcripts are often degraded by the cellular surveillance mechanism known as nonsense-mediated decay (123), which is also an active regulator of transcription (21). Variants in nonsense-mediated decay factors can be pathogenic, causing loss of function or a poison exon effect (129).

Pseudogenes that have lost the ability to encode functional proteins are typically difficult to identify from NGS data and can result in false positives if they are homologous to a disease-associated gene, especially when identifying transcribed and single-exon pseudogenes (66). But the finding that many pseudogenes are transcribed may indicate their functional potential and needs to be studied further (146). In fact, there is evidence that pseudogenes can play a regulatory role. Studies have found that transcripts derived from the pseudogene locus *PTENP1* regulate the expression of the parent gene *PTEN*, as indicated by *PTEN* downregulation by *PTENP1* deletion in breast and colon cancer (148, 149) and by methylation of the *PTENP1* promoter sequence in clear-cell renal cell carcinoma (202).

Another set of important genetic elements are the long (>200 base pairs) noncoding RNAs (lncRNAs). lncRNAs are generally smaller than protein-coding genes and are hard to identify, but they are important to consider in understanding gene-regulatory functions, especially because RNA-seq assays predict that there are more lncRNAs than protein-coding genes (183). lncRNAs are gaining interest because of their potential associations with disease (207). In fact, a few variants of lncRNA loci have been shown to disrupt neighboring gene function, causing disorders such as spinocerebellar ataxia type 8 (151).

Two studies, one using mini-gene-trapping insertional mutagenesis and one using CRISPR/Cas9, recently screened and identified the most critical human genes needed for the viability of different human cell lines (20, 192), which opens up opportunities to identify in further detail the essential tissue- and disease-specific genes and new therapeutic targets (for a review, see 205). Detailed annotation of gene structure and function will help identify potential therapies for diseases. NGS technologies greatly aid in annotating gene structure and function, especially

with combinatorial use of short and long reads (189). Although current RNA-seq technology is unable to sequence by long reads, which makes the assembly of full-length long transcripts difficult, new technologies such as the Pacific Biosciences platforms and SLR-seq (synthetic long read sequencing) could potentially sequence a complete transcript in a single read (13, 179). GS is expected to be able to elucidate all genetic elements using longer reads, especially SVs. Moreover, transcript data sets from technologies such as cap analysis gene expression (CAGE), RNA annotation and mapping of promoters for analysis of gene expression (RAMPAGE), and polyA-seq [poly(A) sequencing] greatly improve the accurate identification of the 5′ and 3′ ends of transcripts (14, 41, 170), further helping in gene annotation. Considerable efforts are being made to increase the catalog of new genes associated with diseases such as neurological and developmental disorders (38). As described above, however, there could be many more unidentified pathogenic variants in noncoding regulatory regions of the genome that affect or are associated with diseases, which represents the next challenge for genetic annotation. New data sets from technologies such as Hi-C and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) are needed to identify the physical interactions between the regulatory regions affected by variants in them and the regulated genes (51, 107).

NEW CONCEPTS OF HUMAN DISEASE-ASSOCIATED GENES AND VARIANTS

Functionally classifying genes and their downstream products is important in order to understand gene-disease associations and the general modes of these associations. Efforts to catalog all human disease-related genes are ongoing. A functional classification of 1,000 known disease-associated genes found that features of the associated diseases such as age of onset and mode of inheritance were well correlated with the causative gene product function (81). These results indicate that the discovery of new disease-associated genes can help determine which diseases would benefit from better functional genome annotation. We currently know of approximately 19,000 human protein-coding genes; this number continues to change as genome annotations improve. A comprehensive list of human genes will facilitate faster targeted NGS assays to help increase our understanding of variants related to the disease context.

Gene deletion and duplication events, collectively known as CNVs, make major contributions to genomic variations that can cause pathogenicity (137). A comprehensive CNV map of the human genome has been created using high-quality genomic data from healthy individuals of different ethnic populations; interestingly, it revealed that almost 100 genes can be completely deleted without any predicted phenotypic effect (203). This finding not only suggests the genetic redundancy in the human genome, but also indicates that a combination of subsets of genes controls phenotypic effects in health and disease. It also may lead to more targeted genomic approaches to elucidate genome-phenome relationships.

Mandelker et al. (115) recently created an exome-wide resource that catalogs highly homologous genomic regions in order to aid molecular diagnostic applications. This resource will be helpful in genome annotation and the identification of pathogenic and benign variants from NGS applications because it ranks homologous regions based on their degree of affectedness and medical relevance and classifies them by the nature of their homology. It will also be helpful in NGS applications by reducing the false positive and false negative calls that can result from high homology with pseudogenes.

Much attention is being given to noncoding regions and the corresponding variants, which are mostly regulatory. For example, SVs that cause changes in the three-dimensional structure of topologically associating domains or genomic neighbors or in the chromatin structure at

specific locations may relocate enhancer or repressor elements that affect gene expression, which is associated with developmental disorders and cancer (for a review, see 173). Genomic structural mechanisms can also be intertwined with disease-associated variations. Using high-coverage GS data from 1,400 individuals from the 1000 Genomes Project and CARTaGENE, Hussin et al. (74) showed that the exons in genomic regions of low recombination belong to highly conserved essential genes and are significantly enriched in putatively deleterious disease-causing variants. Disease-related genome-scale models can be helpful in understanding these genomic complexities and in identifying disease genes (132). Oberhardt et al. (133) created such a model for metabolic disorders using genome-scale metabolic reconstructions, generating a predictive network model of several thousand metabolic disorders and their associated genes and protein products.

THE EXPANDING PHENOTYPIC SPECTRUM OF GENES: THE POWER OF EXOMES AND GENE PANELS

ES and targeted gene-panel sequencing provide increased flexibility and control by enabling focused genomic analyses from the perspective of specific disease phenotypes. Ultimately, such targeted analyses are expected to lead to the development of focused gene-gene networks for different penetrances and expressivities of disease phenotypes, including multigenic heterogeneous disorders, and to provide clearer molecular diagnostics for individual patients. This, in turn, will also help in the discovery of new gene-disease associations and lead to further merging of genomic research and the corresponding clinical work, particularly personalized medicine.

There are thousands of genetic disorders, and most of them are rare and clinically heterogeneous. The diagnostic yield for identifying balanced translocations by common genetic tests such as karyotyping and array comparative genomic hybridization is low, ranging from 5% to 20% at most. Specialized, focused testing is therefore needed to confirm a diagnosis. ES is now used as a specialized diagnostic screen for candidate genes that may influence a patient's disease phenotype because each ES can provide roughly 30,000 variants in the coding regions that can then be filtered to help identify the most relevant genetic etiology (58).

Several important guidelines have been put forward regarding the use of ES as a diagnostic tool in clinical laboratories (97), some of which are reviewed below. ES can sometimes lead to incidental findings of pathogenic variants that are not related to the initial patient diagnosis and may cause late-onset treatable or untreatable disease or indicate a predisposition to cancer or other diseases. The ACMG has suggested that a list of 59 actionable gene-disease pairs be evaluated in genomic tests when patients consent (87). An estimated 2–3% of patients might have such actionable incidental findings, although this varies by ancestry (7, 43, 82). ES can also detect carrier status of recessive diseases and variants that may affect the patient's response to various pharmaceutical drugs (60, 68).

From the technological point of view, almost all CLIA-certified laboratories currently use the Illumina platform for ES; the preferred choice for the target capture library is the Agilent SureSelect Clinical Research Exome, which includes 80% of all exome targets with a minimum 20× coverage and an additional 10% coverage of disease-associated genes (143). As discussed above, for ES in particular, laboratories should share the specific technologies they use, including the library and bioinformatics pipeline, to help make NGS clinical procedures more consistent. Interpretations of disease associations for rare and novel variants from ES data should include the medical and family histories of the patients.

Because of the rapid decrease in NGS cost per base, ES has become a standardized platform for both research and clinical diagnostics and is often used to bridge the gap between the two. Because of its enhanced coverage and capacity, it is being used in many studies worldwide, leading

to the creation of the ExAC data set, which comprises data from more than 60,000 individuals and is expected to grow further (50, 110). As an example, Posey et al. (150) recently used ES in conjunction with observed clinical phenotypes to understand the clinical utility of ES in the molecular diagnosis of adult patients. The authors curated the phenotypic compositions of the ES data using the Human Phenotype Ontology database. Interestingly, they found that de novo mutations contributed to approximately 61% of autosomal dominant diagnoses and that, overall, the diagnostic rate was higher (approximately 24%) for patients between 18 and 30 years of age and lower for older patients. These results indicate that an individual's age should be considered by both physicians and clinics when deciding on the type of NGS test to perform. They also suggest the power of ES in adult molecular diagnostics and, potentially, the importance of de novo mutations in the Mendelian basis of genetic disease in the adult population. Moreover, the molecular diagnoses revealed that 7% of cases were a combination of Mendelian disorders, indicating blended phenotypes and a broader spectrum of associated genes.

ES has been able to reveal the expanding genetic and phenotypic spectra of various disease types, leading to deeper molecular understanding, better diagnoses, and the discovery of biomarkers. For example, ES revealed the broad genetic and corresponding phenotypic spectrum of kidney diseases: *COL4A3-5* genes that were classically associated with Alport syndrome were also found to affect focal segmental glomerulosclerosis, and the *DGKE* gene, which is involved in nephrotic syndrome, was found to harbor variants that are associated with atypical hemolytic uremic syndrome (for a review of the utility of NGS platforms, particularly ES, in studies of kidney disease, see 177). ES supplemented with split-read mapping was also used to discover pathogenic CNVs and the wider spectrum of *AHI1* and *TMEM237* mutations that are involved Meckel-Gruber syndrome and Joubert syndrome (194), demonstrating the enhanced diagnostic yield of ES, especially in discovering new genes, multigenic associations, and variant pathogenicity.

Leslie et al. (103) used ES in combination with Sanger sequencing to elucidate the broader genotypic and phenotypic spectra of popliteal pterygium disorders, which are highly heterogeneous in their phenotypic representations. They found that multiple genes contribute to the different phenotypic gradations and showed that these genes are potentially linked in interconnected pathways that are involved in epidermal and craniofacial development. These findings clearly indicate the clinical relevance of using ES in helping clinicians not only in diagnostics, but also in disease management and counseling for patient families.

A recent case study that used muscle biopsies to perform ES reported that two individuals from the same family had a novel heterozygous mutation in exon 3 of the *NKX2-1* transcription factor gene that causes mitochondrial dysfunction (35). This study's identification of a new pathogenic effect arising from a previously unreported mutation shows the versatility of ES in finding broad spectra of genes and disease—in this case, *NKX2-1*-related mitochondrial and immunological dysfunction.

Many studies have shown that disease phenotypic spectra are broad and highly heterogeneous. Examples include congenital disorders of glycosylation (78); X-linked *SLC9A6* gene mutations, which have a wide range of effects, including global developmental delay and intellectual disability (172); mutations of the *CHD7* gene, which lead to the broad phenotypic effects of CHARGE syndrome (83); and glucose transporter 1 deficiency syndromes (100). Some of these studies were carried out using automated Sanger sequencing. We predict that ES will enhance the coverage of the coding genome in these heterogeneous, phenotypically diverse disorders and will be able to precisely pinpoint de novo and inherited variations and their disease associations.

As technology continues to advance, the increasing use of ES will generate large amounts of data from different cohorts, which creates new challenges in bioinformatics concerning how to store, analyze, and interpret these data (176). Work in this area is ongoing, and new developments should

help researchers and clinicians cope with this problem (102). With regard to variant identification from ES data, current exome capture kits can capture 95% of the coding regions with a minimum coverage of 20× and a median coverage of 100× (101). One drawback of ES is its low coverage to identify CNVs, for which normalized read counts in a genomic region of an individual are generally compared with those in other exomes; numerous algorithms have been developed for this purpose, including CODEX (Copy Number Detection by Exome Sequencing), Convex, Conifer, and XHMM (Exome Hidden Markov Model) (80, 94). The population frequencies of variants are highly informative for variant interpretation, but there is a need for population-specific databases of variants; however, interpretations should be done carefully to avoid false positives, missing disease-associated founder mutations, and somatic or tissue-specific variants (3, 63, 111). The CADD (Combined Annotation-Dependent Depletion) score is widely used to evaluate variant pathogenicity by computing a functional meta-score that integrates a variety of genome-wide annotations (92). This approach is more sensitive and specific than tools such as Polyphen2, SIFT (Sorting Intolerant from Tolerant), and PhyloP, which depend heavily on the evolutionary conservation of the protein-coding variant. For noncoding variants that are potentially regulatory, DeepSEA (a deep-learning-based algorithm framework for predicting the chromatin effects of sequence alterations with single-nucleotide sensitivity) and DeltaSVM (Delta Support Vector Machine) are widely used tools that are also trained by a deep-learning algorithm on a variety of noncoding annotations, primarily from the Encyclopedia of DNA Elements (ENCODE) Project (99, 208). For splice variants, SPANR (Splicing-Based Analysis of Variants) predicts the effect of a variant on mRNA splicing using a deep-learning computational model scoring and is quite effective (199).

It is essential to understand how a gene that harbors a particular variant is relevant to a particular disease. Algorithms such as PHIVE (Phenotypic Interpretation of Variants in Exomes) try to compare cross-species phenotypic similarity in order to prioritize genes in the exome data for a given disease (161). Human disease genes are much less tolerant to genetic variations than other genes are (89, 147), which opens up a new approach to understanding the deleterious nature of genetic variations based on the use of population variance and the tolerance level of a given variation. Screening for recurrent mutations in heterogeneous disorders has relied on ES in large cohorts of patients for different variants in the same candidate gene in order to determine variant pathogenicity (36), which in turn has relied heavily on statistical approaches such as the use of genome-wide mutation rates to identify genes enriched in de novo mutations (126). It is important to note that various data-sharing platforms for rare diseases have facilitated understanding of common phenotypes and genotypes and patterns across populations that help enable variant classification.

Targeted NGS gene-panel tests have also been giving the field of genomics an edge in identifying causative genes and variants, providing a faster pipeline with a high diagnostic yield in clinical settings. It can be more biased than ES or GS, but for heterogeneous disorders, it provides higher coverage for known disease-associated genes and can be flexible in including suspected genes in the targeted library based on patient phenotype. Au et al. (12) recently evaluated the clinical utility of an NGS 54-gene-panel test for acute myeloid leukemia. They used the Illumina MiSeq platform for 50 patient samples, and comparing the panel result with conventional molecular testing revealed more than 95% similarity with sufficiently high coverage and diagnostic yield. Numerous deleterious variants were identified, especially in the *TP53* gene, which causes the disease, suggesting an overall high sensitivity and specific detections of mutations in disease-relevant genes. Panel tests are already becoming standard for adult disorders and are likely to have a large impact in future research and clinical genomics, especially as a first-stage diagnostic test.

THE BURDEN OF VARIANTS OF UNKNOWN SIGNIFICANCE

Clinical diagnostic laboratories follow ACMG guidelines (159) when interpreting and annotating variants from NGS data in order to classify them as benign, likely benign, pathogenic, or likely pathogenic or as a variant of unknown significance (VUS). As genome, exome, and targeted gene-panel NGS continue to produce massive amounts of data, the number of VUS findings is also increasing, especially in cases of heterogeneous disorders in which multigenic effects are predicted. The increased number of VUSs creates a burden on variant classification, disease association, and clinical reporting for follow-up diagnosis and therapy options.

To help address the burden of the rising number of VUSs, Narravula et al. (124) attempted to reclassify the VUSs present in ClinVar (a database of all discovered variants in the human genome, with classifications by laboratories across the world) NGS data from newborn screening, focusing on variants associated with three monogenic metabolic disorders (variants of the *ACADM*, *GALT*, and *PAH* genes). To do so, they used the population variance of disease, ExAC curation of population data, and published literature surveys, including functional data. By using a reclassification strategy, they were able to reannotate multiple VUSs as benign (or likely benign) or pathogenic (or likely pathogenic). The VUS burden is significant in clinical diagnostic settings, and this work suggests the importance of standardized curation of all information on variants, especially functional data, across all clinical laboratories. We predict that functional genomics, including gene- or protein-domain-specific functional assays and the integration of genomics with downstream functional -omics (proteomics, epigenomics, and metabolomics), will ultimately help elucidate genotype-phenotype relationships for variant classification (similar to gene annotation; see **Figure 3**).

CONCLUSION: MERGING BASIC AND CLINICAL GENOMICS

The genomics era is moving toward an understanding of the functionality of genomic complexities by using an integrated -omics approach that is blurring the distinctions between basic and clinical genomics, albeit with cautionary recommendations and guidelines. For example, because of the increasing importance of epigenomics in neurological diseases that emerged from basic research, such as epilepsy (72, 79), several medical industries are including analysis of epigenetic signatures, such as detection of the DNA methylome, in their diagnostic pipeline priority in both discovery and clinical settings (138). For this purpose, genome interpretation companies such as Congenica, Sophia Genetics, and Omicia are providing rapid turnaround times. Yet the high failure rate of drugs in clinical trials and subsequent financial losses (10, 145) suggest the need to delve more deeply into identifying personalized and tissue-specific -omics data and merging basic and clinical genomic research—for example, by identifying tissue-specific exons and transcripts through better annotation. Simultaneously, improved methods for generating not only genomic data but also data from other -omics approaches and their integrative methodologies should be developed in order to facilitate the translation of -omics techniques into the clinic in the form of genomic medicine.

Emerging efforts to use companion diagnostics in personalized medicine are driving the integration of combinatorial -omics approaches while simultaneously developing personalized genomic and other -omic databases that will likely be used in clinical settings in the future. These developments have resulted largely from the meteoric rise of NGS technologies, which have facilitated the merging of genotypic assays such as genome, exome, and targeted gene-panel sequencing with phenotypic approaches such as epigenetic analyses, RNA-seq, and proteomic and metabolomic profiling using mass spectrometry. This merging drives personalized biomarker discovery and medicine and, in turn, facilitates recruitment in clinical trials, which have so far

centered mainly on cancer diagnostics and therapy (for detailed reviews, including recommendations and guidelines for clinical genomic medicine, see 24, 37, 39, 67, 84, 85, 108, 140, 171). Practical, ethical, and regulatory aspects of personalized medicine should also be considered in great detail, because this field will eventually help in building new technologies for merging genomic and phenomic data that will have a large impact both on basic research and in the clinic.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank all of our colleagues at the Department of Human Genetics at Emory University for helpful comments on earlier versions of this review.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73
2. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
3. Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, et al. 2015. Post-zygotic point mutations are an underrecognized source of de novo genomic variation. *Am. J. Hum. Genet.* 97:67–74
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. 2013. Signatures of mutational processes in human cancer. *Nature* 500:415–21
5. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3:246–59
6. AllSeq. 2017. *Complete Genomics*. **http://allseq.com/knowledge-bank/sequencing-platforms/complete-genomics**
7. Amendola LM, Dorschner MO, Robertson PD, Salama JS, Hart R, et al. 2015. Actionable exonic incidental findings in 6503 participants: challenges of variant classifications. *Genome Res.* 25:305–15
8. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13:229–32
9. Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, et al. 2012. Before it gets started: regulating translation at the 5′ UTR. *Comp. Funct. Genom.* 2012:475731
10. Arrowsmith J, Miller P. 2013. Trial watch: phase II and phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* 12:569
11. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. 2010. Clinical assessment incorporating a personal genome. *Lancet* 375:1525–35
12. Au CH, Wa A, Ho DN, Chan TL, Ma ES. 2016. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagn. Pathol.* 11:11
13. Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. *PLOS ONE* 7:e46679
14. Batut P, Gingeras TR. 2013. RAMPAGE: promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. *Curr. Protoc. Mol. Biol.* 104:25B.11.1–16
15. Bauters M, Frints SG, Van Esch H, Spruijt L, Baldewijns MM, et al. 2014. Evidence for increased *SOX3* dosage as a risk factor for X-linked hypopituitarism and neural tube defects. *Am. J. Med. Genet. A* 164A:1947–52
16. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, et al. 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* 3:65ra4

17. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
18. BGI. 2015. *Revolocity*^{TM} *whole genome sequencing service*. **http://www.bgi.com/wp-content/uploads/2015/10/Global-WGSRevolocity-ENG-10-15.pdf**
19. BGI. 2017. *BGISEQ-500: a BGI sequencer*. **http://seq500.com/en/portal/Sequencer.shtml**
20. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, et al. 2015. Gene essentiality and synthetic lethality in haploid human cells. *Science* 350:1092–96
21. Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev*. 29:63–80
22. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, et al. 2009. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res*. 19:1044–56
23. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10:1213–18
24. Caberlotto L, Lauria M. 2015. Systems biology meets -omic technologies: novel approaches to biomarker discovery and companion diagnostic development. *Expert Rev. Mol. Diagn.* 15:255–65
25. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* 40:722–29
26. Chan M, Ji SM, Yeo ZX, Gan L, Yap E, et al. 2012. Development of a next-generation sequencing method for *BRCA* mutation screening: a comparison between a high-throughput and a benchtop platform. *J. Mol. Diagn.* 14:602–12
27. Chiocchetti AG, Kopp M, Waltes R, Haslinger D, Duketis E, et al. 2015. Variants of the CNTNAP2 5′ promoter as risk factors for autism spectrum disorders: a genetic and functional approach. *Mol. Psychiatr.* 20:839–49
28. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *PNAS* 106:19096–101
29. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, et al. 2015. Extending reference assembly models. *Genome Biol*. 16:13
30. Church GM, Gao Y, Kosuri S. 2012. Next-generation digital information storage in DNA. *Science* 337:1628
31. Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11:415–25
32. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4:265–70
33. Colombo M, Blok MJ, Whiley P, Santamariña M, Gutiérrez-Enriquez S, et al. 2014. Comprehensive annotation of splice junctions support pervasive alternative splicing at the *BRCA1* locus: a report from the ENIGMA consortium. *Hum. Mol. Genet.* 23:3666–80
34. Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29:987–91
35. Coon EA, Ahlskog JE, Patterson MC, Niu Z, Milone M. 2016. Expanding phenotypic spectrum of *NKX2-1*-related disorders—mitochondrial and immunologic dysfunction. *JAMA Neurol.* 73:237–38
36. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, et al. 2012. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* 367:1921–29
37. Deans ZC, Costa JL, Cree I, Dequeker E, Edsjo A, et al. 2017. Integration of next-generation sequencing in clinical diagnostic molecular pathology laboratories for analysis of solid tumours; an expert opinion on behalf of IQN Path ASBL. *Virchows Arch*. 470:5–20
38. Deciphering Dev. Disord. Study. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519:223–28
39. Delaney SK, Hultner ML, Jacob HJ, Ledbetter DH, McCarthy JJ, et al. 2016. Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert Rev. Mol. Diagn.* 16:521–32

40. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491–98
41. Derti A, Garrett-Engele P, MacIsaac KD, Stevens RC, Sriram S, et al. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22:1173–83
42. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, et al. 2014. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311:1035–45
43. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, et al. 2013. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum Genet.* 93:631–40
44. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *PNAS* 100:8817–22
45. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81
46. Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* 14:496–506
47. English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, et al. 2015. Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genom.* 16:286
48. Esposito MV, Nunziato M, Starnone F, Telese A, Calabrese A, et al. 2016. A novel pathogenic *BRCA1* splicing variant produces partial intron retention in the mature messenger RNA. *Int. J. Mol. Sci.* 17:2145
49. Farooq M, Fatima A, Mang Y, Hansen L, Kjaer KW, et al. 2016. A novel splice site mutation in CEP135 is associated with primary microcephaly in a Pakistani family. *J. Hum. Genet.* 61:271–73
50. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–20
51. Fullwood MJ, Ruan Y. 2009. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.* 107:30–39
52. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, et al. 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat. Biotechnol.* 30:1033–36
53. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, et al. 2015. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat. Biotechnol.* 33:689–93
54. Gargis AS, Kalman L, Lubin IM. 2016. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J. Clin. Microbiol.* 54:2857–65
55. Garnett MJ, McDermott U. 2014. The evolving role of cancer cell line-based screens to define the impact of cancer genomes on drug response. *Curr. Opin. Genet. Dev.* 24:114–19
56. Genome Ref. Consort. 2017. *Human genome overview*. **https://www.ncbi.nlm.nih.gov/grc/human**
57. GenomeWeb Staff Report. 2016. White House announces efforts to accelerate Precision Medicine Initiative. *GenomeWeb*, Feb. 25. **https://www.genomeweb.com/molecular-diagnostics/white-house-announces-efforts-accelerate-precision-medicine-initiative**
58. Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* 20:490–97
59. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17:333–51
60. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, et al. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15:565–74
61. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* 47:435–44
62. Gundersen S, Kalas M, Abul O, Frigessi A, Hovig E, Sandve GK. 2011. Identifying elemental genomic track types and representing them uniformly. *BMC Bioinform.* 12:494
63. Gunel M, Awad IA, Finberg K, Anson JA, Steinberg GK, et al. 1996. A founder mutation as a cause of cerebral cavernous malformation in Hispanic Americans. *N. Engl. J. Med.* 334:946–51
64. Guo J, Xu N, Li Z, Zhang S, Wu J, et al. 2008. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *PNAS* 105:9145–50

65. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10:R32
66. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. 2012. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* 22:1760–74
67. Hayes DF, Schott AF. 2015. Personalized medicine: genomics trials in oncology. *Trans. Am. Clin. Climatol. Assoc.* 126:133–43
68. Hegde M, Bale S, Bayrak-Toydemir P, Gibson J, Bone Jeng LJ, et al. 2015. Reporting incidental findings in genomic scale clinical sequencing—a clinical laboratory perspective: a report of the Association for Molecular Pathology. *J. Mol. Diagn.* 17:107–17
69. Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* 7:12989
70. Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15:585–98
71. Henshall DC. 2014. MicroRNA and epilepsy: profiling, functions and potential clinical applications. *Curr. Opin. Neurol.* 27:199–205
72. Henshall DC, Kobow K. 2015. Epigenetics and epilepsy. *Cold Spring Harb. Perspect. Med.* 5:a022731
73. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39:1522–27
74. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, et al. 2015. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat. Genet.* 47:400–4
75. Int. Hum. Genome Seq. Consort. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
76. Ion Torrent. 2011. *Ion semiconductor sequencing uniquely enables both accurate long reads and paired-end sequencing*. **https://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_098680.pdf**
77. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, et al. 2008. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* 18:780–90
78. Jaeken J, Matthijs G. 2007. Congenital disorders of glycosylation: a rapidly expanding disease family. *Annu. Rev. Genom. Hum. Genet.* 8:261–78
79. Jakovcevski M, Akbarian S. 2012. Epigenetic mechanisms in neurological disease. *Nat. Med.* 18:1194–204
80. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. 2015. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* 43:e39
81. Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. *Nature* 409:853–55
82. Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, et al. 2012. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am. J. Hum. Genet.* 91:97–108
83. Jongmans MC, Admiraal RJ, van der Donk KP, Vissers LE, Baas AF, et al. 2006. CHARGE syndrome: the phenotypic spectrum of mutations in the *CHD7* gene. *J. Med. Genet.* 43:306–14
84. Jørgensen JT. 2015. Clinical application of companion diagnostics. *Trends Mol. Med.* 21:405–7
85. Jørgensen JT. 2015. Companion diagnostics: the key to personalized medicine. *Expert Rev. Mol. Diagn.* 15:153–56
86. Ju J, Kim DH, Bi L, Meng Q, Bai X, et al. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *PNAS* 103:19635–40
87. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, et al. 2017. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* 19:249–55
88. Karow J. 2015. Qiagen launches GeneReader NGS system at AMP; presents performance evaluation by Broad. *GenomeWeb*, Nov. 4. **https://www.genomeweb.com/molecular-diagnostics/qiagen-launches-genereader-ngs-system-amp-presents-performance-evaluation**
89. Khurana E, Fu Y, Chen J, Gerstein M. 2013. Interpretation of genomic variants using a unified biological network approach. *PLOS Comput. Biol.* 9:e1002886

90. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, et al. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* 316:1481–84
91. Kircher M, Kelso J. 2010. High-throughput DNA sequencing—concepts and limitations. *BioEssays* 32:524–36
92. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46:310–15
93. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30:693–700
94. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22:1525–32
95. Ku CS, Polychronakos C, Tan EK, Naidoo N, Pawitan Y, et al. 2013. A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol. Psychiatr.* 18:141–53
96. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
97. Lapin V, Mighion LC, da Silva CP, Cuperus Y, Bean LJ, Hegde MR. 2016. Regulating whole exome sequencing as a diagnostic test. *Hum. Genet.* 135:655–73
98. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, et al. 2003. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* 24:3769–77
99. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, et al. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47:955–61
100. Leen WG, Klepper J, Verbeek MM, Leferink M, Hofste T, et al. 2010. Glucose transporter-1 deficiency syndrome: the expanding clinical and genetic spectrum of a treatable disorder. *Brain* 133:655–70
101. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. 2015. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum. Mutat.* 36:815–22
102. Lelieveld SH, Veltman JA, Gilissen C. 2016. Novel bioinformatic developments for exome sequencing. *Hum. Genet.* 135:603–14
103. Leslie EJ, O'Sullivan J, Cunningham ML, Singh A, Goudy SL, et al. 2015. Expanding the genetic and phenotypic spectrum of popliteal pterygium disorders. *Am. J. Med. Genet. A* 167A:545–52
104. Levy SE, Myers RM. 2016. Advancements in next-generation sequencing. *Annu. Rev. Genom. Hum. Genet.* 17:95–115
105. Li H. 2014. On the graphical representation of sequences. *Heng Li's Blog*, July 25. **http://lh3.github.io/2014/07/25/on-the-graphical-representation-of-sequences**
106. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, et al. 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* 32:915–25
107. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93
108. Lin E, Chien J, Ong FS, Fan JB. 2015. Challenges and opportunities for next-generation sequencing in companion diagnostics. *Expert Rev. Mol. Diagn.* 15:193–209
109. Liu L, Li Y, Li S, Hu N, He Y, et al. 2012. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364
110. Lohmueller KE, Sparsø T, Li Q, Andersson E, Korneliussen T, et al. 2013. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.* 93:1072–86
111. MacArthur DG, Tyler-Smith C. 2010. Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* 19:R125–30
112. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. 2016. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protoc.* 11:2081–103
113. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. 2011. What can exome sequencing do for you? *J. Med. Genet.* 48:580–89
114. Malapelle U, Vigliar E, Sgariglia R, Bellevicine C, Colarossi L, et al. 2015. Ion Torrent next-generation sequencing for routine identification of clinically relevant mutations in colorectal cancer patients. *J. Clin. Pathol.* 68:64–68

115. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, et al. 2016. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* 18:1282–89
116. McCarroll SA, Altshuler DM. 2007. Copy-number variation and association studies of human disease. *Nat. Genet.* 39:S37–42
117. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33:5868–77
118. Metzker ML. 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11:31–46
119. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinform.* 15:247
120. Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* 447:932–40
121. Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML, Cocho JA. 2013. A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* 110:3–24
122. Myers EW. 2005. The fragment assembly string graph. *Bioinformatics* 21(Suppl. 2):ii79–85
123. Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* 23:198–99
124. Narravula A, Garber KB, Askree SH, Hegde M, Hall PL. 2017. Variants of uncertain significance in newborn screening disorders: implications for large-scale genomic sequencing. *Genet. Med.* 19:77–82
125. Natl. Hum. Genome Res. Inst. 2016. *The cost of sequencing a human genome*. Updated July 16. **https://www.genome.gov/sequencingcosts**
126. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485:242–45
127. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, et al. 2010. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42:790–93
128. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. 2010. Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42:30–35
129. Nguyen LS, Jolly L, Shoubridge C, Chan WK, Huang L, et al. 2012. Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability. *Mol. Psychiatr.* 17:1103–15
130. Niesler B, Flohr T, Nöthen MM, Fischer C, Rietschel M, et al. 2001. Association between the 5′ UTR variant C178T of the serotonin receptor gene HTR3A and bipolar affective disorder. *Pharmacogenetics* 11:471–75
131. Nothnagel M, Herrmann A, Wolf A, Schreiber S, Platzer M, et al. 2011. Technology-specific error signatures in the 1000 Genomes Project data. *Hum. Genet.* 130:505–16
132. Oberhardt MA, Gianchandani EP. 2014. Genome-scale modeling and human disease: an overview. *Front. Physiol.* 5:527
133. Oberhardt MA, Palsson BØ, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5:320
134. Oda M, Glass JL, Thompson RF, Mo Y, Olivier EN, et al. 2009. High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res.* 37:3829–39
135. Offit K. 2014. Decade in review—genomics: a decade of discovery in cancer genomics. *Nat. Rev. Clin. Oncol.* 11:632–34
136. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, et al. 2015. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet.* 6:235
137. Oskoui M, Gazzellone MJ, Thiruvahindrapuram B, Zarrei M, Andersen J, et al. 2015. Clinically relevant copy number variations detected in cerebral palsy. *Nat. Commun.* 6:7949
138. Pac. Biosci. 2012. *Detecting DNA base modifications using single molecule, real-time sequencing*. **http://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf**
139. Panjwani N, Wilson MD, Addis L, Crosbie J, Wirrell E, et al. 2016. A microRNA-328 binding site in PAX6 is associated with centrotemporal spikes of rolandic epilepsy. *Ann. Clin. Transl. Neurol.* 3:512–22

140. Pant S, Weiner R, Marton MJ. 2014. Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics. *Front. Oncol.* 4:78
141. Parikh H, Mohiyuddin M, Lam HY, Iyer H, Chen D, et al. 2016. svclassify: a method to establish benchmark structural variant calls. *BMC Genom*. 17:64
142. Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10:669–80
143. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. 2011. A comparative analysis of exome capture. *Genome Biol.* 12:R97
144. Patwardhan A, Harris J, Leng N, Bartha G, Church DM, et al. 2015. Achieving high-sensitivity for clinical applications using augmented exome sequencing. *Genome Med*. 7:71
145. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, et al. 2010. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9:203–14
146. Pei B, Sisu C, Frankish A, Howald C, Habegger L, et al. 2012. The GENCODE pseudogene resource. *Genome Biol*. 13:R51
147. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet*. 9:e1003709
148. Poliseno L, Haimovic A, Christos PJ, Vega Y Saenz de Miera EC, Shapiro R, et al. 2011. Deletion of *PTENP1* pseudogene in human melanoma. *J. Investig. Dermatol.* 131:2497–500
149. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465:1033–38
150. Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, et al. 2016. Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet. Med.* 18:678–85
151. Qureshi IA, Mehler MF. 2012. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat. Rev. Neurosci.* 13:528–41
152. Ramser J, Abidi FE, Burckle CA, Lenski C, Toriello H, et al. 2005. A unique exonic splice enhancer mutation in a family with X-linked mental retardation and epilepsy points to a novel role of the renin receptor. *Hum. Mol. Genet.* 14:1019–27
153. Rauch C, Trieb M, Wibowo FR, Wellenzohn B, Mayer E, Liedl KR. 2005. Towards an understanding of DNA recognition by the methyl-CpG binding domain 1. *J. Biomol. Struct. Dyn.* 22:695–706
154. Regalado A. 2015. U.S. to develop DNA study of one million people. *MIT Technology Review*, Jan. 30. **https://www.technologyreview.com/s/534591/us-to-develop-dna-study-of-one-million-people**
155. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, et al. 2013. ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* 15:733–47
156. Reimand J, Wagih O, Bader GD. 2015. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLOS Genet*. 11:e1004919
157. Ren L, Zhu R, Li X. 2016. Silencing miR-181a produces neuroprotection against hippocampus neuron cell apoptosis post-status epilepticus in a rat model and in children with temporal lobe epilepsy. *Genet. Mol. Res.* 15:gmr.15017798
158. Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Mol. Cell* 58:586–97
159. Richards S, Aziz N, Bale S, Bick D, Das S, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17:405–24
160. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, et al. 2013. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLOS ONE* 8:e66621
161. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genet. Proj., Wang K, et al. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* 24:340–48
162. Rogozhina Y, Mironovich S, Shestak A, Adyan T, Polyakov A, et al. 2016. New intronic splicing mutation in the *LMNA* gene causing progressive cardiac conduction defects and variable myopathy. *Gene* 595:202–6
163. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348–52
164. Rydbeck H, Sandve GK, Ferkingstad E, Simovski B, Rye M, Hovig E. 2015. ClusTrack: feature extraction and similarity measures for clustering of genome-wide data sets. *PLOS ONE* 10:e0123261

165. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46:944–50
166. Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res.* 20:1165–73
167. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31:1009–14
168. Shen Y, Sarin S, Liu Y, Hobert O, Pe'er I. 2008. Comparing platforms for *C. elegans* mutant identification using high-throughput whole-genome sequencing. *PLOS ONE* 3:e4012
169. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–32
170. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *PNAS* 100:15776–81
171. Simon R, Roychowdhury S. 2013. Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* 12:358–69
172. Sinajon P, Verbaan D, So J. 2016. The expanding phenotypic spectrum of female *SLC9A6* mutation carriers: a case series and review of the literature. *Hum. Genet.* 135:841–50
173. Spielmann M, Mundlos S. 2016. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* 25:R157–65
174. Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* 61:437–55
175. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* 24:2066–76
176. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. 2015. Big data: astronomical or genomical? *PLOS Biol.* 13:e1002195
177. Stokman MF, Renkema KY, Giles RH, Schaefer F, Knoers NV, van Eerde AM. 2016. The expanding phenotypic spectra of kidney diseases: insights from genetic studies. *Nat. Rev. Nephrol.* 12:472–83
178. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81
179. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, et al. 2015. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33:736–42
180. Timmerman L. 2015. DNA sequencing market will exceed $20 billion, says Illumina CEO Jay Flatley. *Forbes*, Apr. 29. **http://www.forbes.com/sites/luketimmerman/2015/04/29/qa-with-jay-flatley-ceo-of-illumina-the-genomics-company-pursuing-a-20b-market**
181. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, et al. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509:371–75
182. UK10K Consort. 2015. The UK10K project identifies rare variants in health and disease. *Nature* 526:82–90
183. Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* 154:26–46
184. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18:1051–63
185. Vengoechea J, Parikh AS, Zhang S, Tassone F. 2012. De novo microduplication of the *FMR1* gene in a patient with developmental delay, epilepsy and hyperactivity. *Eur. J. Hum. Genet.* 20:1197–200
186. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
187. VeritasGenetics. 2015. *VeritasGenetics breaks $1,000 whole genome barrier*. Press Release, Sept. 29. **https://www.veritasgenetics.com/documents/VG-PGP-Announcement-Final.pdf**
188. VeritasGenetics. 2016. *VeritasGenetics launches $999 whole genome and sets new standard for genetic testing*. Press Release, Mar. 4. **https://www.veritasgenetics.com/documents/veritas-mygenome-final3-mar-9-2016.pdf**
189. Voelkerding KV, Dames SA, Durtschi JD. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55:641–58

190. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. 2013. Cancer genome landscapes. *Science* 339:1546–58
191. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, et al. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res.* 24:1734–39
192. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, et al. 2015. Identification and characterization of essential genes in the human genome. *Science* 350:1096–101
193. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57–63
194. Watson CM, Crinnion LA, Berry IR, Harrison SM, Lascelles C, et al. 2016. Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. *BMC Med. Genet.* 17:1
195. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* 92:530–46
196. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–76
197. Wheeler DA, Wang L. 2013. From human genome to cancer genome: the first decade. *Genome Res.* 23:1054–62
198. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* 13:255–62
199. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806
200. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, et al. 2011. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* 43:864–68
201. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. 2013. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* 369:1502–11
202. Yu G, Yao W, Gumireddy K, Li A, Wang J, et al. 2014. Pseudogene PTENP1 functions as a competing endogenous RNA to suppress clear-cell renal cell carcinoma progression. *Mol. Cancer Ther.* 13:3086–97
203. Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat. Rev. Genet.* 16:172–83
204. Zeng T, Dong ZF, Liu SJ, Wan RP, Tang LJ, et al. 2014. A novel variant in the 3′ UTR of human *SCN1A* gene from a patient with Dravet syndrome decreases mRNA stability mediated by GAPDH's binding. *Hum. Genet.* 133:801–11
205. Zhan T, Boutros M. 2016. Towards a compendium of essential genes—from model organisms to synthetic lethality in cancer cells. *Crit. Rev. Biochem. Mol. Biol.* 51:74–85
206. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, et al. 2015. Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* 373:2336–46
207. Zhang X, Weissman SM, Newburger PE. 2014. Long intergenic non-coding RNA HOTAIRM1 regulates cell cycle progression during myeloid maturation in NB4 human promyelocytic leukemia cells. *RNA Biol.* 11:777–87
208. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12:931–34
209. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3:160025
210. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, et al. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32:246–51