Hindawi Publishing Corporation
Journal of Healthcare Engineering
Volume 2016, Article ID 3617572, 14 pages
http://dx.doi.org/10.1155/2016/3617572

Review Article

Challenges of Identifying Clinically Actionable Genetic Variants for Precision Medicine

Tonia C. Carter[1] and Max M. He[1,2,3]

[1]*Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA*
[2]*Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA*
[3]*Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA*

Correspondence should be addressed to Max M. He; he.max@mcrf.mfldclin.edu

Received 5 August 2015; Revised 16 March 2016; Accepted 17 March 2016

Academic Editor: Saverio Affatato

Advances in genomic medicine have the potential to change the way we treat human disease, but translating these advances into reality for improving healthcare outcomes depends essentially on our ability to discover disease- and/or drug-associated clinically actionable genetic mutations. Integration and manipulation of diverse genomic data and comprehensive electronic health records (EHRs) on a big data infrastructure can provide an efficient and effective way to identify clinically actionable genetic variants for personalized treatments and reduce healthcare costs. We review bioinformatics processing of next-generation sequencing (NGS) data, bioinformatics infrastructures for implementing precision medicine, and bioinformatics approaches for identifying clinically actionable genetic variants using high-throughput NGS data and EHRs.

1. Introduction

High-throughput genomics technology has made possible the era of precision medicine, an approach to healthcare that involves integrating a patient's genetic, lifestyle, and environmental data and then comparing these data to similar data collected for thousands of other individuals to predict illness and determine the best treatments. Precision medicine aims to tailor healthcare to patients by using clinically actionable genomic mutations to guide preventive interventions and clinical decision making [1]. In the past 25 years, more than 4,000 Mendelian disorders have been studied at the genetic level [2]. In addition, more than 80 million genetic variants have been uncovered in the human genome [3, 4]. Clinical pharmacology research using electronic health record (EHR) systems has recently become feasible as EHRs have been implemented more widely [5]. Also, studies such as the Electronic Medical Records and Genomics-Pharmacogenomics (eMERGE-PGx) project [6], GANI_MED project [7], SCAN-B initiative [8], and Cancer 2015 study [9] have been designed to assess the value of next-generation sequencing (NGS) in healthcare.

Combining the functional characterization of identified genomic mutations with comprehensive clinical data available in EHRs has the potential to provide compelling evidence to implicate novel disease- and/or drug-associated mutations in phenotypically well-characterized patients. NGS is increasingly used in biomedical research and clinical practice. NGS technological advances in clinical genome sequencing and adoption of EHRs will pave the way to create patient-centered precision medicine in clinical practice. NGS technology is an essential component supporting genomic medicine but the volume and complexity of the data pose challenges for its use in clinical practice [10]. Sequencing a single human genome generates megabytes of data; therefore, investment in a bioinformatics infrastructure is required to implement NGS in clinical practice.

The term "big data" is defined differently by different people [11]. Gartner defines big data as "high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (http://www.gartner.com/it-glossary/big-data/) while others define it as the 5 Vs, which are Volume, Velocity,

Table 1: Sequencing assays.

Characteristic	DNA sequencing			RNA-seq	
	Targeted genomic regions	Whole exome	Whole genome	Targeted	Transcriptome profiling
Capture method*	Amplicon-based targeting; hybrid capture; in-solution capture	Hybrid capture; in-solution capture	None	Hybridization only; hybridization and extension; multiplexed PCR	None
Amount of genome/transcriptome sequenced	~150 bp–62 Mb (≤2% of genome)	~30–60 Mb (1-2% of genome)	~3 Gb (≥95% of genome)	Variable: transcripts of ~10–1000 genes	Entire transcriptome
Amplification	Yes	Yes	Not required	Yes	Required for low-quantity RNA samples
Sequencing depth	100–1000x[Ü]	80–100x[Ü]	30–50x[Ü]	0.3–25 million reads[‡]	15–200 million reads[‡]
Amount of sequence data generated per sample	~0.3–5 Gb	~4-5 Gb	~90 Gb	~0.5–3 Gb	~5-6 Gb

bp, base pairs; Mb, megabases; Gb, gigabases; PCR, polymerase chain reaction.
*Method used to select genomic regions for sequencing.
[Ü]Number of times a single base is read during a sequencing run.
[‡]A greater number of reads are needed to detect rare transcripts.

Variety, Verification/Veracity, and Value [12]. In this review, we describe how one source of big data, in the form of genomic data generated by NGS, is processed and being used to improve healthcare and clinical research. We give an overview of NGS technologies, bioinformatics processing of NGS data, bioinformatics approaches for identifying clinically actionable variants in sequence data, guidelines for maintaining high standards when generating genomic data for clinical use, bioinformatics infrastructures of studies aimed at implementing precision medicine, and methods for ensuring the security of genomic data. We also discuss the need for the efficient integration of genomic information into EHRs.

2. Genomic Data Generation

2.1. Approaches to Sequencing. NGS includes DNA sequencing and RNA sequencing (RNA-seq) (Table 1). DNA sequencing approaches include (1) whole-genome sequencing (WGS), (2) whole exome sequencing (WES) of the coding regions of all known genes, and (3) targeted sequencing of genomic regions or genes implicated in a disease [13]. In addition, RNA-seq is used in transcriptome profiling to sequence all RNA transcripts (the transcriptome) in cells at a given time point to measure gene expression, targeted sequencing for measuring the expression of transcripts encoded by a specific genomic region, and sequencing of small RNAs. Targeted DNA sequencing is already being applied in some areas of clinical practice such as pharmacogenomics (e.g., the eMERGE-PGx project [6]), while WGS, and particularly WES, is emerging into the clinic for the evaluation of developmental brain disorders such as intellectual disability [14], autism [15], and seizures [16]. With continuing decreases in the costs of sequencing, it is expected that the use of WES/WGS and RNA sequencing in healthcare will become more common.

2.2. Read Depth. NGS involves breaking DNA into fragments and determining the order of the nucleotide bases in each fragment. The sequence of each fragment is called a "read." Because the distribution of reads across the genome is uneven (due to biases in sample preparation, sequencing-platform chemistry, and bioinformatics methods for genomic alignment and assembly of the reads) [17, 18], some bases are present in more reads and others in fewer reads. Read depth refers to the number of reads that contain a base; for example, a 10x read depth means that each base was present in an average of 10 reads. For RNA-seq, read depth is more often stated in terms of the number of millions of reads. Variant calling is more reliable with increasing read depth, and a greater depth is advantageous for detecting rare genetic variants with confidence. The read depth needed can depend on multiple factors including guidelines from the scientific community, the presence of repetitive genomic regions (these are more difficult to sequence), the error rate of the sequencing platform, the algorithm used for assembling reads into a genomic sequence, and gene expression level (for RNA-seq). Read depth recommendations from the scientific literature include 100x for heterozygous single nucleotide variant detection by WES [19], 35x for genotype detection by WGS [20], 60x for detecting insertions/deletions (INDELs) by WGS [21], 10–25 million reads for differential gene expression profiling by RNA-seq [22], and 50–100 million reads for allele-specific gene expression by RNA-seq [23].

2.3. Sequencing Technologies

2.3.1. Description of Technologies. Commercially available sequencing platforms use a variety of methods to generate sequence data (Table 2). Sequencing-by-synthesis (MiSeq and HiSeq 4000 platforms) is the enzymatic synthesis of a DNA strand complementary to a template DNA strand. For

Table 2: Comparison of sequencing instruments.

Characteristic	MiSeq	PacBio RS II	Ion S5	HiSeq 4000	454 GS FLX Titanium XL+	SOLiD 5500xl W	Sanger Genetic Analyzer 3500xL
Instrument price	~$125 K	~$695 K	~$65 K	~$900 K	~$500 K	~$595 K	~$173 K
Sequencing mechanism	Sequencing-by-synthesis	Single-molecule, real-time sequencing	Semiconductor sequencing	Sequencing-by-synthesis	Pyrosequencing	Oligonucleotide ligation	Dideoxynucleotide chain termination
Sequencing application	Targeted	Targeted; transcriptome profiling	Targeted; whole exome; transcriptome profiling	Whole exome/genome; transcriptome profiling	Whole exome/genome; transcriptome profiling	Whole exome/genome; transcriptome profiling	Next-generation sequencing validation, targeted sequencing of mutations or small insertions/deletions
Maximum read length	300 bp PE	10,000 bp	200 bp	150 bp PE	700 bp	75 bp SE, 50 bp mate-paired	850 bp
Reads per run	15 million	55–900 K	60–80 million	2.5–5 billion	~1 million	100 million–4.8 billion	Not applicable
Output data per run	0.5–15 Gb	0.5–16 Gb	~44 Gb	125–1500 Gb	~0.7 Gb	160–320 Gb	2–100 Kb
Run time	4–55 hours	6 hours	1-2 days	<1–3.5 days	23 hours	2–7 days	0.5–3 hours
Advantages	Low error rate; short run time	Long read length; short run time	Short run time; low start-up cost	Low error rate; high throughput	Long read length	Low error rate	Low error rate; long read length
Disadvantages	Higher cost per base compared to HiSeq instruments	Medium/high cost per base	High error rate for homopolymer tracts and insertions/deletions	Short read length	High error rate for homopolymer tracts	Short read length; long run time	High cost per base; low throughput

bp, base pairs; Gb, gigabases; K, thousand; Kb, kilobases; PE, paired-end; SE, single-end.

NGS, the procedure involves DNA fragmentation, creation of a DNA library by attaching adaptors to each fragment, amplification of the fragments on a solid surface, synthesis of a DNA strand complementary to each template DNA fragment (using DNA polymerase), and fluorescence imaging to identify each newly incorporated nucleotide on the synthesized DNA strands [24]. Single-molecule, real-time sequencing (PacBio RS II platform) is a modification of sequencing-by-synthesis [25]. In this approach, each DNA polymerase molecule is immobilized at the bottom of a nanoscale well called a zero-mode waveguide. A laser light illuminates the well from below and emits a pulse of light when a fluorescent-labelled nucleotide is added to the nascent DNA strand by DNA polymerase (bound to a template DNA fragment), allowing detection of the incorporated nucleotide. Semiconductor sequencing (Ion S5 platform) is another modification of sequencing-by-synthesis that uses a semiconductor-sensing device to detect the addition of unmodified nucleotides during DNA synthesis [26]. Pyrosequencing (454 GS FLX Titanium XL+ platform) is a technique that couples sequencing-by-synthesis to a chemiluminescent enzyme (luciferase) reaction that generates visible light allowing detection of nucleotide incorporation during DNA synthesis [27]. Oligonucleotide ligation (SOLiD 5500xl W platform) involves ligating oligonucleotide probes to template DNA strands to determine the sequence of the template [28]. Sequencing by dideoxynucleotide (ddNTP) chain termination (Sanger Genetic Analyzer 3500xL platform), often called Sanger sequencing, involves incorporation of ddNTPs by DNA polymerase during DNA synthesis [29]. Fluorescence labelling allows identification of each of the ddNTPs added to the synthesized DNA strands.

2.3.2. Comparison of Sequencers. The MiSeq, PacBio RSII, and Ion S5 sequencers were designed for targeted sequencing and sequencing small genomes (e.g., the genomes of microorganisms) whereas the HiSeq 4000, 454 GS FLX Titanium XL+, and SOLiD 5500xl W can be used for WES and WGS of human genomes (Table 2). The instruments most often used in precision medicine programs performing WES/WGS of the human genome in clinical care settings are the HiSeq sequencers [30] that have the advantages of a relatively high sample throughput and a low sequencing error rate. However, all of the NGS technologies are being applied to health research [31–36]. The single-molecule, real-time sequencing technology generates the longest reads (Table 2), making the PacBio RS II instrument well suited for de novo sequencing (by assembly of reads into long contiguous sequences) of the genomes of organisms that do not have a reference genome (e.g., many microbial genomes) [37].

The sequencers that cost the least are the bench-top Ion S5 and MiSeq instruments (Table 2), and for many laboratories it would be feasible to buy more than one of these instruments. While they can be used to perform WES of the human genome, the sequencing cost per base would be much higher compared with WES on the HiSeq instrument. The HiSeq 4000, 454 GS FLX Titanium XL+, and SOLiD 5500xl W instruments are more expensive, costing between $500,000 and $900,000 each, but they are capable of sequencing several human genomes or exomes within a few days to one week. Large laboratories that expect to assay many samples routinely by WES/WGS might consider it cost-efficient to buy more than one of these sequencers to meet assay demand. All six next-generation sequencers in Table 2 produce at least 0.5 gigabases per run and most output several gigabases per run, giving an idea of the volume of data that needs to be considered when planning for the data storage and processing capabilities of bioinformatics pipelines to be used in clinical laboratories that perform NGS assays.

2.3.3. Sequencing Accuracy. With continued refinement in technology, many NGS platforms have demonstrated a low rate of errors in variant detection (1/1000 to 1/50 bases depending on the instrument and read depth) [38, 39]. Previous reports have compared sequencing accuracy among the technologies presented in Table 2. In a comparison of the HiSeq 2000 and SOLiD 5500xl platforms for WGS of human DNA samples, the HiSeq 2000 had higher sensitivity for calling single nucleotide variants but the SOLiD 5500xl had a lower false positive rate [40]. When the Ion PGM, MiSeq, and PacBio RS sequencers were compared by sequencing four microbial genomes, the PacBio RS had the highest sequencing error rate, and Ion PGM data had slightly more variant calls and a higher false positive rate than MiSeq data [41]. Compared with other technologies, the 454 and PacBio RS platforms have demonstrated the most unbiased read distribution in genomic regions with a high GC content [41, 42], an important factor affecting the probability of calling a variant in these regions. However, the 454 platform has a tendency to assess the length of homopolymer tracts incorrectly, resulting in false positive single nucleotide variant calls in these tracts [42].

In comparison with NGS technologies, Sanger sequencing is widely considered the most accurate sequencing method (error rate as low as 1 in 10,000 bases) [43] and remains the gold standard. Genetic variants detected using NGS should always be validated by an independent method if the variants are relevant to clinical care or are associated with health outcomes in research studies. Because of its high accuracy, Sanger sequencing is often used for validation. Other methods of validation, especially for common single nucleotide variants, INDELs, or structural variants, include polymerase chain reaction (PCR) and genotype/copy number variant arrays.

3. Genomic Data Processing and Quality Control

3.1. Data Processing. Data files generated by next-generation sequencers contain raw sequence reads, each with a unique identifier, and their quality scores. Sequence reads need to be evaluated for data quality and to exceed minimum quality thresholds, before being processed for read alignment [44], variant calling [45], and variant annotation [46, 47] in a bioinformatics pipeline (Figure 1). Read alignment involves aligning the sequence reads to a reference sequence [48, 49] of the human genome to allow comparison of sequence data from the patient sample with the reference sequence. Reads

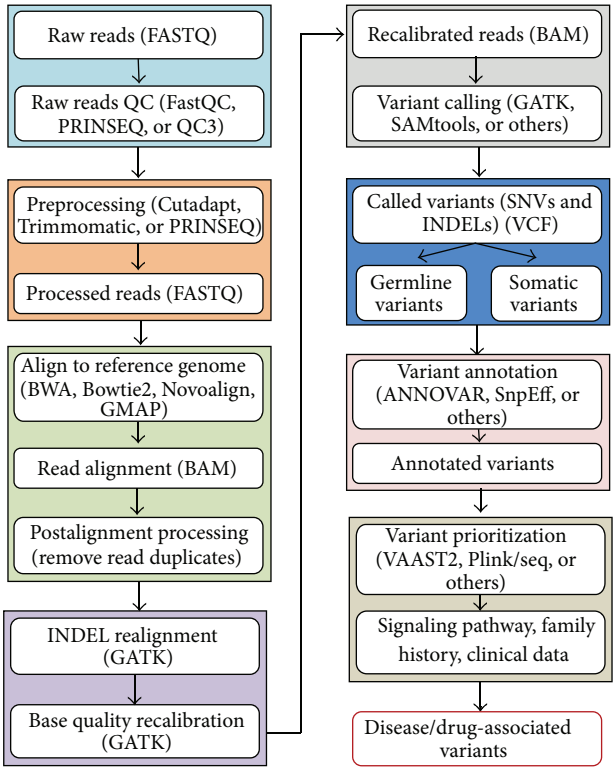

Figure 1: A flow chart of processing next-generation sequencing data.

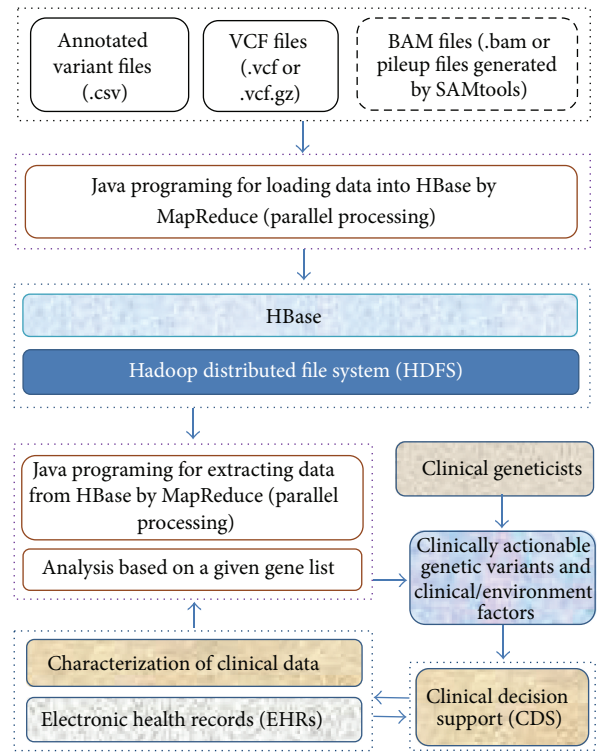

Figure 2: The basic framework of SeqHBase for detecting clinically actionable genetic variants.

with an uncertain alignment location need to be removed before further data processing. Alignment allows a number of quality control measures to be determined, for example, the percentage of all reads that align to a reference sequence, the percentage of unique reads that align to a reference sequence, and the number of reads that align at a specific locus (read depth). These measures influence the reliability of variant calling, the next step in a NGS bioinformatics pipeline. Variant calling tools, such as SAMtools [50], GATK [45], and others, are used to identify differences in sequence between the patient sample and a reference. These differences can include changes of one nucleotide (single nucleotide variants, SNVs), a few nucleotides (small INDELs), or larger regions, such as copy number variants (CNVs) and other structural variations (SVs). These software programs allow users to specify different parameters to adjust for minimizing false positive and false negative variant calls. Variant annotation depends on biological knowledge and provides information on the known or likely impact of variants on gene and protein function [46, 47]. To produce a patient report, annotated variants are interpreted in a disease-specific context and are often classified based on their known or expected clinical impact. For instance, the ClinVar [51] variant database, released on May 4, 2015 (http://www.ncbi.nlm.nih.gov/clinvar/), by the National Center for Biotechnology Information (NCBI), contained more than 110,000 unique genetic variants having clinical interpretations [52].

3.2. Clinically Actionable Variants. In clinical care, the American College of Medical Genetics and Genomics (ACMG) has recommended the identification and return of incidental findings (IFs) for clinically significant variants in a set of 56 "highly medically actionable" genes associated with 24 inherited conditions [53, 54]. Also, the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) has reported actionable exomic IFs from 112 genes in 6,503 participants [55]. The 112 genes included 52 ACMG genes and an additional 60 "actionable" genes. To infer biological insights from massive amounts of NGS data and comprehensive clinical data in a short period of time, we have developed an analysis pipeline within a software framework called SeqHBase [56] (Figure 2) to quickly identify disease- or drug-associated genetic variants. There were more than 27 million unique variants among 300 patients with WGS data that we analyzed using SeqHBase. In addition to identifying variants that are annotated as "pathogenic" or "likely pathogenic" by ClinVar [51], we compiled a list of low frequency or rare variants that are possibly damaging, and novel loss-of-function (LoF) variants that are absent in the ClinVar database, to allow clinical geneticists to review the potential pathogenicity of these variants further. As SeqHBase is a big data-based toolset, it takes only a few minutes to analyze WGS data for 300 individuals and to generate a candidate list of actionable genomic variants. More detailed discoveries from these WGS data will be described in future reports.

SeqHBase is one of several, freely accessible bioinformatics tools for prioritization of variants from WES/WGS data. Daneshjou et al. reported a web-based tool for identifying clinically actionable variants in the 56 ACMG genes [57], and Zhou et al. developed a variant characterization framework for targeted analysis of relevant reads from high-throughput sequencing data [58]. Other tools include PHIVE [59] which prioritizes variants in genes responsible for mouse model phenotypes that are similar to the phenotypes of patients being tested by WES and OVA [60] that performs prioritization by integrating data on human and model organism phenotypes, gene function, and known biological pathways.

Identifying clinically actionable variants remains a challenge despite the availability of variant prioritization tools. A workshop convened by the National Human Genome Research Institute and the Wellcome Trust identified limited evidence of the clinical significance of genetic variants and the lack of a comprehensive database of genetic variant-phenotype associations as barriers to the implementation of precision medicine [61]. It was noted that existing catalogs of clinically actionable variants are not standardized, are maintained by different entities (e.g., laboratories or government organizations), and are not designed to interact with EHRs. To speed the incorporation of genomic data into clinical care, the workshop advocated for a dynamic, centralized database that can be updated with available, reliable evidence on variant pathogenicity. The Clinical Genome Resource (ClinGen) program [52], developed in response to this recommendation, provides resources (e.g., ClinVar [51]) to aid the understanding of genetic variation and the use of genetic variation in clinical practice.

3.3. Quality Control. Best practices for quality control in the bioinformatics processing of NGS data have been reported in the scientific literature [45, 62]. Quality control metrics include total reads, ratio of unique reads to total reads, proportion of bases covered at a specified minimum read depth, mean read depth, raw sequence error rates, ratio of transitions (pyrimidine-to-pyrimidine or purine-to-purine mutation) to transversions (pyrimidine-to-purine mutation or vice versa), missingness (proportion of genomic sites at which a variant could not be called), homozygosity, heterozygosity, and distribution of known and novel variants relative to those contained in the dbSNP database. For targeted or exome sequencing, an additional metric is capture efficiency, the percentage of targeted bases that are covered by one or more reads.

These metrics can be calculated using the PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/) or VCFtools [63] software programs that can readily be incorporated into a bioinformatics pipeline, allowing assessment of NGS data quality in both clinical and research settings. Values for the first four metrics depend on the type of sequencing assay performed but, in general, higher values indicate better data quality. The raw sequence error rates and missingness should be as low as possible. The ratio of transitions to transversions (Ti/Tv ratio) is expected to be ~2.0–2.1 for WGS data overall, 2.10 for known variants in WGS data, 2.07 for new variants in WGS data, ~3.0–3.3 for WES data overall, 3.5 for known variants in WES data, and 3.0 for new variants in WES data [45]. Homozygosity and heterozygosity depend on the type of population: heterozygosity is expected to be more frequent in admixed populations and homozygosity to be more frequent in inbred populations. It is estimated that each person has ~200 novel SNPs not present in the dbSNP database [64]; therefore, a value that is much larger than 200 is indicative of a high false positive rate of single nucleotide variant calls. Capture efficiency is reported to range within ~50–75% [65].

There are no existing, quality control standards that relate to generating clinical interpretations for genetic variants. However, substantial efforts are being made to identify clinically actionable pharmacogenetics variants, and it is instructive to review the approach being used. The Coriell Personalized Medicine Collaborative [66], the Clinical Pharmacogenetics Implementation Consortium [67], the Pharmacogenetics Working Group established by the Royal Dutch Association for the Advancement of Pharmacy [68], and the Evaluation of Genomic Applications in Practice and Prevention initiative sponsored by the Centers for Disease Control and Prevention [69] have independently developed similar processes for selecting candidate drugs, reviewing the published literature to identify drug-gene associations, scoring the evidence supporting associations between genetic variants and drug response, and interpreting the evidence to provide treatment guidelines.

This approach involving review and interpretation of the scientific literature by an expert committee can be considered the gold standard for determining whether a variant is clinically relevant or actionable but also can be expensive and time-consuming. It will not be feasible for experts, either individually or in committees, to review the large number of genetic variants identified in NGS data. Tools such as POLYPHEN-2 [70], VEP [71], MutationAssessor [72], and SIFT [73] can be used to predict variant effects. However, because these tools are sometimes inaccurate [74] and often differ in their predictions for the same variant [75, 76], there will likely be many variants that have no clear predicted, clinical interpretation. Furthermore, an additional problem is that the predictions made by these tools are not specific to a given gene or class of genes. For example, many genes would tolerate the substitution of glycine for another amino acid, but, in a gene that encodes a collagen fibril, loss of a glycine would impair fiber assembly resulting in a significant phenotype [77]. New methods that are both accurate and efficient need to be developed for predicting the pathogenicity of genetic variants found by NGS.

A limitation of using the ClinVar database [51] to identify clinically actionable genomic mutations is that a genetic variant in ClinVar can be described as having a different potential for pathogenicity by different submitters. For example, of the 12,895 unique variants with multiple clinical interpretations that have been submitted by more than one laboratory, 2,229 (17%) were interpreted differently by different submitters, with one- or two-step differences between any of three major levels: "pathogenic or likely pathogenic," "uncertain significance," and "likely benign or benign" [52]. Differences in interpreting the pathogenicity of variants have also been reported by the Clinical Sequencing Exploratory Research

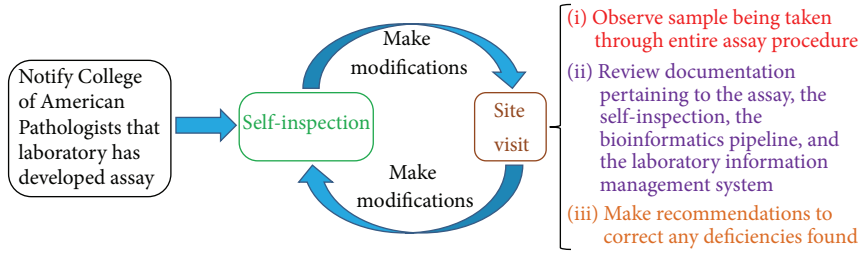

Figure 3: Overview of steps for a laboratory to obtain accreditation by the College of American Pathologists.

(CSER) program [30], an initiative designed to trial the use of WES/WGS data in clinical practice. The program compared CSER laboratories on their clinical interpretations of 98 variants and observed one-step differences in interpretation for 42% of variants and two-step or larger differences for 20% of variants [78]. To estimate and interpret the pathogenicity of new variants that are absent in the ClinVar database and to achieve some level of consensus on the clinical interpretations of variants, evaluations from experts, such as clinical geneticists, and/or further biological functional studies are needed.

4. Guidelines for Bioinformatics Processes

4.1. Summary of Guidelines. Bioinformatics pipelines are constituted of multiple databases and software programs to convert raw sequence reads to a list of clinically actionable or candidate variants. To promote the transparency of pipeline processes and data flow, the ACMG [79], the College of American Pathologists (CAP) [80], Weiss et al. [81], and Gargis et al. [82] have offered guidelines for NGS and the operation of bioinformatics pipelines in a clinical setting. The recommendations of these guidelines include thorough documentation of the pipeline and of deviations from pipeline standard operating procedures (e.g., software updates, changes in software settings, operator error, hardware failure, or other failures in the pipeline), validation of the pipeline, development of a pipeline quality management program, and implementation of policies to ensure secure data storage and data transfer.

The recommendations for written patient reports state that gene names should be provided according to HUGO Gene Nomenclature Committee nomenclature (http://www.genenames.org/) and genetic variants according to the nomenclature guidelines of the Human Genome Variation Society (http://www.hgvs.org/). Laboratories should follow the recommendations of the ACMG [53, 54] for interpreting the clinical significance of variants. Patient reports should also include the genome build and reference sequence used for variant detection, the genomic coordinates of identified variants, and mention of whether clinically significant variants were confirmed by an independent assay method [81]. Laboratories should also report genetic variant data (gene name, zygosity, cDNA nomenclature, protein nomenclatures, exon number, and clinical significance) in a structured format according to HL7 standards (HL7 version 2 Implementation Guide: Clinical Genomics, http://www.hl7.org/implement/standards/). This is aimed at providing sufficient data to facilitate both clinical decision support and the display of genetic information in the EHR.

Challenges to implementing these guidelines include the constantly evolving nature of NGS technologies, bioinformatics tools (necessitating frequent updates of the bioinformatics pipeline), clinical interpretation (necessitating frequent updates of genetic variant annotation), the limited capacity of health care organizations/laboratories to store the voluminous data generated by NGS platforms (data storage options considered must ensure security of the stored genetic data), and the need for personnel trained in bioinformatics and statistics to develop a bioinformatics pipeline and to process and analyze NGS data. However, these challenges are not insurmountable, and it is likely that health care institutions that want to use NGS data in clinical care will attempt to overcome these hurdles and follow the guidelines.

4.2. Accreditation from the College of American Pathologists (CAP). Clinical laboratories that develop Clinical Laboratory Improvement Amendments- (CLIA-) certified NGS assays based on CAP standards [80] can seek accreditation from CAP, an agency that can provide accreditation on behalf of the CLIA program. The accreditation process involves a site visit inspection by a peer institution/laboratory once every two years and a self-inspection in alternate years (Figure 3). For the self-inspection, CAP sends the laboratory a list of items, specific to the NGS assay, that need to be checked by the laboratory. Completing the self-inspection for a NGS assay would allow the laboratory to determine how closely it adheres to the CAP standards for the assay. For the site visit, the inspectors would observe a sample being taken through the entire assay procedure. Any deficiencies found must be corrected, and CAP should be provided with a report describing the corrective measures within 30 days after the site visit. Through the mechanism of CAP accreditation, the laboratory would inform external entities that it provides a CLIA-certified assay that meets CAP standards for the assay.

5. Bioinformatics Infrastructure for Genomic Data

5.1. Separate Databases for Different Types of Data. Welch et al. have proposed an infrastructure comprised of independent, interacting databases for processing and storing genomic data in a clinical setting [83] (Figure 4). These

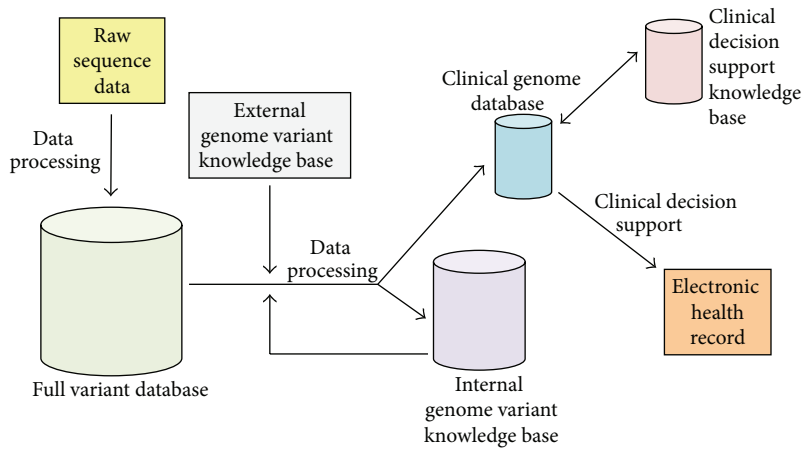

Figure 4: Elements of a proposed infrastructure for bioinformatics processing of sequencing data in clinical laboratories.

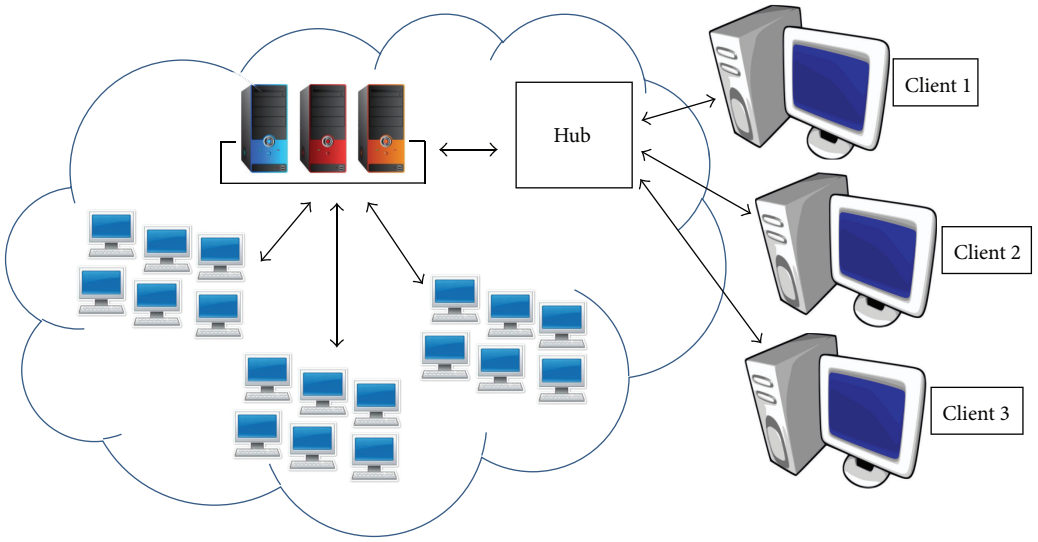

Figure 5: Cloud computing diagram.

databases include a "full variant database" to store all genetic variants for each patient, a clinical genome database to store only the clinically relevant variants for each patient, a clinical decision support knowledge base that integrates decision rules and guidelines for providing care (e.g., drug dosing rules) with genomic and clinical information, and a genome variant knowledge base to store known genetic variants and their clinical interpretations. ClinVar [51] is an example of a freely accessible genome variant knowledge base but clinical laboratories will likely also maintain their own internal genome variant knowledge bases (based on the genomic data of patients they test). The proposed infrastructure can potentially accommodate large amounts of genomic data because it involves warehousing the data external to EHRs. However, it would require investment in data storage capacity external to the EHR database system and the development and maintenance of interfaces between the genomic databases and the EHR database system [84].

5.2. Cloud Computing. Cloud computing, involving the use of remote servers to store and access data and software programs (Figure 5), has also been proposed for genomic data processing and storage. Cloud computing providers offer infrastructure, software, and programming platforms as services and incur the costs for developing and maintaining these services [85]. Because clients pay only for the services they use, cloud computing offers an economical approach to genomic data management compared with investment in the creation and maintenance of databases by healthcare entities to house genomic data. Hadoop is an open-source programming platform that is already being used to develop software for genomic data processing in a cloud computing environment [85]. Hadoop breaks data into small fragments, distributes the fragments over many computers, distributes computation to where the fragments are located so all fragments are processed in parallel, and aggregates the results at the end of computation [85]. The parallel processing of many small pieces of data greatly reduces computation

time. Examples of open-source software developed on the Hadoop platform for processing genomic data are Crossbow [86], GATK [87], and Hadoop-BAM [88]. Challenges to the use of cloud computing for genomic data include the long data transfer times for uploading NGS data files to the cloud, the perceived lack of data security in the cloud computing environment, and the need for advanced programming skills in Java to develop software using Hadoop [85].

5.3. Infrastructure for Data Sharing. The separation of genomic and clinical data repositories facilitates the use of genomic data in research as well as clinical care. To engage in collaborative research, infrastructure for sharing genomic data with researchers internal and external to the institution that generated the data is required. The Global Alliance for Genomics and Health (GA4GH) [89], an international coalition of healthcare and academic centers that aims to advance the sharing of genomic and clinical data to improve health, has launched efforts to create such an infrastructure. The group has developed an application programming interface (API) to support the sharing of data on DNA sequences and genomic variants across organizations and bioinformatics pipelines [90]. GA4GH is also developing APIs for other types of genomics-related data including variant annotations, RNA-seq, and genotype-phenotype associations. These tools will allow genomic data from multiple organizations to be analyzed in aggregate, increasing statistical power to identify genetic variants that have a clinical effect.

5.4. Security of Genomic Data. Genomic data is protected health information; therefore, its privacy and confidentiality should be maintained similarly to other protected health information. Safeguards include the use of data encryption, password protected files, secure data transfer, audits of data transfer processes, and the implementation of institutional policies against data breeches and malicious use of the data [91]. The use of cloud computing presents added security concerns because data storage and/or processing services are provided by an entity external to the healthcare organization. Measures that the cloud service provider can take to address these concerns include logging access to the data, creating a role-based access system (level of access depends on the type of user), complying with third-party certifications for information security (e.g., the International Organization for Standardization/International Electrotechnical Commission 21001:2013 information security standard http://www.iso.org/iso/home/standards/management-standards/iso27001.htm/), protecting the security of the computer network, using notification alarms to track when changes are made to stored data, and guaranteeing the complete removal of data from its servers once the cloud storage service is no longer being used [92].

6. Examples of Implementing Genomic Data in Clinical Care

6.1. Clinical Sequencing Exploratory Research Program. A survey of six health centers participating in the CSER consortium has described how the centers have integrated genomic data into the EHR [30]. Five centers performed sequencing at their own laboratories, and one site used an external laboratory but confirmed variants on-site using Sanger sequencing. Each center created a local bioinformatics pipeline for variant annotation, but all used multiple online catalogs of variants (e.g., ClinVar [51] and dbSNP) for annotation. Each site also built and maintained its own genome variant knowledge base (based on genetic variants ascertained in patients at the site) and created tools to use data from this internal database in variant annotation. Additionally, sites used manual or semiautomated methods to search the scientific literature or online gene-specific databases to determine the clinical significance of variants. EHR software was obtained from commercial providers at four centers and was locally developed at two centers. The laboratories at all six centers generated a human-readable PDF document, containing genetic results, that was designed to be incorporated into the EHR. The two sites with custom-built EHRs, and one site with commercial EHR software, also reported results in a structured, machine-readable format. Active clinical decision support (automated alerts through the EHR) for genetic variants was available at two of the centers. Only one center had an automated system for sending alerts to physicians when new genomic findings resulted in the reclassification of a genetic variant's clinical significance (e.g., a variant initially classified as being of unknown significance was subsequently discovered to have serious clinical consequences).

6.2. eMERGE Network Pharmacogenetics Study. Sites in the eMERGE network are also engaged in pilot efforts to incorporate genomic data, particularly data relevant to pharmacogenetics, into EHRs [6]. At one eMERGE site, separate data repositories were created for unprocessed sequence/genotype data and for variants of known pharmacogenetics relevance [93]. Software that applied approved pharmacogenetics-medication guidelines to patients' genetic data was used to determine a patient's pharmacogenetics phenotype (e.g., predicted poor metabolizer of a specific drug), and the phenotype data were stored as a laboratory result in the EHR. The site developed software that extended its existing, custom-built medication alert system, enabling the system to check for a relevant pharmacogenetics laboratory result when a physician prescribes a pharmacogenetics-related drug. If a patient has a pharmacogenetics phenotype, the system sends an alert to the physician and suggests alternative treatment. Another eMERGE site reported developing similar infrastructure that supported storage of all genetic variants separately from variants with pharmacogenetics relevance, the translation of genetic data into genotype-phenotype associations, and active clinical decision support for physicians prescribing pharmacogenetics-related drugs [94]. Changes in the clinical interpretation of genetic variants (based on new knowledge) that resulted in phenotype reassignment prompted the site to update its genotype-phenotype translation database to reflect the newly determined genotype-phenotype relationships. Because this database was linked to the site's clinical information system, pharmacogenetics data in the EHR was automatically updated.

6.3. Lessons from CSER and eMERGE. The CSER and eMERGE pharmacogenetics programs are in progress and have not yet reported on improvements in patient outcomes as a result of incorporating genomic data into clinical care. Each site in these programs had its own customized bioinformatics pipeline, laboratory information management system, clinical decision support capabilities, and electronic health records that would not be generalizable to other sites. This presents a challenge as a more uniform infrastructure for genomic data processing could be adopted more widely and easily. Based on their experiences, sites in both programs identified a number of factors that need to be addressed to facilitate the integration of genomic data into healthcare: (1) the requirement for active clinical decision support; (2) tools to examine and interpret sequence variants, especially new, undefined variants; (3) approaches to update changes in the clinical significance of sequence variants over time; (4) giving healthcare providers access to consultants trained in genetics; (5) infrastructure for secure and reliable delivery of results to external healthcare providers; and (6) methods for explaining genomic information to patients.

7. Discussion

The ideal, preventive model of patient care is to understand as much about a patient as possible, as early in his/her life as possible, to detect warning signs of serious but preventable illness at an early stage so that preemptive health interventions can be simpler and/or less expensive than treatment implemented at a later stage. Also, knowing a person's individual characteristics is often relevant for providing effective treatment against disease because patients can respond differently to the same treatment. By facilitating precision medicine, advances in genomics have the potential to change the way we prevent and treat diseases. However, the translation of these advances into reality for patient care depends mainly on our ability to discover disease- and/or drug-associated clinically actionable genetic mutations and on our understanding of the roles of these mutations in the disease process.

Healthcare centers that are conducting pilot studies of the integration of genomic data into clinical care have developed a bioinformatics infrastructure for processing NGS data that consists of a group of databases ancillary to the EHR [30, 93, 94]. The infrastructures were, for the most part, locally developed and proprietary, but this is because these centers are among the first healthcare providers to use genomic data in clinical care and there are no established infrastructures to meet their bioinformatics needs. The infrastructures were built along the same general plan: a bioinformatics pipeline for processing NGS data, a database for storing all genetic variants detected in patient samples, a genome variant knowledge base for storing known genetic variants and their clinical interpretation, a database for the subset of variants deemed to be clinically actionable (with variants linked to a specific clinical phenotype), links between databases allowing data transfer, and a method for reporting the results of clinically actionable variants in the EHR. Developing and maintaining a bioinformatics infrastructure for NGS data requires substantial investment in resources and personnel and can be too expensive for small clinical laboratories. However, because genetic variant databases are maintained separately from the EHR, it might be possible for multiple, small laboratories to pool resources to build and share a common bioinformatics infrastructure. The storage and bioinformatics processing of raw NGS data output by sequencing platforms might exceed the infrastructure capacity of even some large healthcare organizations. Therefore, healthcare providers might want to consider cooperatively establishing a cloud computing service designed to store and process genomic data securely for the healthcare community. Clinical laboratories must also consider the cost of sequencing instruments as part of infrastructure costs. Bench-top instruments used for targeted sequencing are less expensive and output less data than instruments that perform WES/WGS. For these reasons, more laboratories are likely to perform targeted sequencing before, or instead of, attempting to build infrastructure to support WES/WGS.

A major challenge to incorporating genomic data into clinical care is the lack of standards for generating NGS data, bioinformatics processing, data storage, and clinical decision support. Standards would promote consistency in data quality, and adherence to standards would facilitate the routine use of genomic data in clinical practice, but it is difficult to create standards when NGS technology and bioinformatics software are constantly evolving. Further, approaches to clinical decision support vary across healthcare institutions [30]. In a survey of 17 health centers participating in the CSER program or the eMERGE network, most centers did not have active clinical decision support for genetic data in the EHR although there were existing mechanisms for clinically actionable information to trigger alerts in the majority of the EHR systems [95]. Centers with active clinical decision support either built their own software locally or customized the clinical decision support capabilities of commercial EHR software [30]. Most centers reported that genetics results were available as a portable document format (PDF) file in the EHR and recommended the development of clinical decision support for disease-defining and pharmacogenetics variants and creation of a clinical decision support knowledge base to advise on appropriate clinical actions (e.g., a change in treatment).

Appropriately integrating EHRs with genomic data for the discovery of clinically actionable variants can generate new insights into disease mechanisms and provide better predictions about effective treatments, all leading to improved targeting of interventions to patients. To generate knowledge on the nature of disease from comprehensive EHR data, new methods such as machine learning, natural language processing, and other artificial intelligence methods are needed. However not all patients are likely to benefit from the use of big data in healthcare due to our current knowledge gaps on how to extract useful information from large-volume genomic and clinical data and how to interpret discovered genetic variants appropriately. At the same time, targeted therapies are not yet available for many important genes, and regulatory issues need to be resolved before some useful bioinformatics tools can be applied in a healthcare setting.

Finally, as EHRs are extremely personal, measures to protect patient data have to be put in place to make certain that patient information is only shared with those who need to see it. Despite this challenge, the potential advantages that genomic data can bring to healthcare far outweigh the potential disadvantages. The growing trend towards integration of genomic data and EHRs will cause concern, but as long as patient privacy and data security can be rigorously maintained, genomic data is certain to play an essential role in precision medicine.

8. Conclusion

To reach the goal of precision medicine, healthcare institutions need to invest in a bioinformatics infrastructure and in personnel trained in bioinformatics and genetics, to develop the capacity to process, store, and interpret genomic data and to link these data with EHRs. In addition, more efforts are needed to distinguish genetic variants that are truly clinically actionable; that is, the variants are useful for guiding clinical decisions regarding interventions to improve health outcomes. Clinical research studies of the implementation of genomic data in healthcare can provide valuable lessons about how genomic data should be managed, and patient privacy protected, when incorporating genomic data into clinical practice on a larger scale. These lessons can alert healthcare institutions to the scientific and technical challenges of using genomic data in precision medicine.

Competing Interests

The authors indicated no potential competing interests.

Acknowledgments

The authors would like to thank Dr. Rachel Stankowski in the Office of Scientific Writing and Publication at the Marshfield Clinic Research Foundation for assistance with editing of this paper. This work was supported by the Clinical and Translational Science Award program through a grant from the National Institutes of Health, National Center for Advancing Translational Sciences [UL1TR000427] and by the Marshfield Clinic Research Foundation.

References

[1] J. L. Vassy, B. R. Korf, and R. C. Green, "How to know when physicians are ready for genomic medicine," *Science Translational Medicine*, vol. 7, no. 287, Article ID 287fs219, 2015.

[2] L. R. Brunham and M. R. Hayden, "Hunting human disease genes: lessons from the past, challenges for the future," *Human Genetics*, vol. 132, no. 6, pp. 603–617, 2013.

[3] G. R. Abecasis, D. Altshuler, A. Auton et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[4] G. R. Abecasis, A. Auton, L. D. Brooks et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[5] O. Gottesman, H. Kuivaniemi, G. Tromp et al., "The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future," *Genetics in Medicine*, vol. 15, no. 10, pp. 761–771, 2013.

[6] L. J. Rasmussen-Torvik, S. C. Stallings, A. S. Gordon et al., "Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems," *Clinical Pharmacology and Therapeutics*, vol. 96, no. 4, pp. 482–488, 2014.

[7] H. J. Grabe, H. Assel, T. Bahls et al., "Cohort profile: greifswald approach to individualized medicine (GANI_MED)," *Journal of Translational Medicine*, vol. 12, article 144, 2014.

[8] L. H. Saal, J. Vallon-Christersson, J. Häkkinen et al., "The Sweden Cancerome Analysis Network—breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine," *Genome Medicine*, vol. 7, no. 1, article 20, 2015.

[9] S. Q. Wong, A. Fellowes, K. Doig et al., "Assessing the clinical value of targeted massively parallel sequencing in a longitudinal, prospective population-based study of cancer patients," *British Journal of Cancer*, vol. 112, no. 8, pp. 1411–1420, 2015.

[10] R. R. Gullapalli, M. Lyons-Weiler, P. Petrosko, R. Dhir, M. J. Becich, and W. A. LaFramboise, "Clinical Integration of Next-Generation Sequencing Technology," *Clinics in Laboratory Medicine*, vol. 32, no. 4, pp. 585–599, 2012.

[11] E. Baro, S. Degoul, R. Beuscart, and E. Chazard, "Toward a literature-driven definition of big data in healthcare," *BioMed Research International*, vol. 2015, Article ID 639021, 9 pages, 2015.

[12] Q. Huang, S. Jing, J. Yi, and W. Zhen, *Innovative Testing and Measurement Solutions for Smart Grid*, John Wiley & Sons, Singapore, 2014.

[13] A. Sboner and O. Elemento, "A primer on precision medicine informatics," *Briefings in Bioinformatics*, vol. 17, no. 1, Article ID bbv032, pp. 145–153, 2016.

[14] C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung et al., "Genome sequencing identifies major causes of severe intellectual disability," *Nature*, vol. 511, no. 7509, pp. 344–347, 2014.

[15] I. Iossifov, B. J. O'Roak, S. J. Sanders et al., "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[16] A. S. Allen, S. F. Berkovic, P. Cossette et al., "De novo mutations in epileptic encephalopathies," *Nature*, vol. 501, no. 7466, pp. 217–221, 2013.

[17] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[18] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[19] M. J. Clark, R. Chen, H. Y. K. Lam et al., "Performance comparison of exome DNA sequencing technologies," *Nature Biotechnology*, vol. 29, no. 10, pp. 908–916, 2011.

[20] S. S. Ajay, S. C. J. Parker, H. O. Abaan, K. V. Fuentes Fajardo, and E. H. Margulies, "Accurate and comprehensive sequencing of personal genomes," *Genome Research*, vol. 21, no. 9, pp. 1498–1505, 2011.

[21] H. Fang, Y. Wu, G. Narzisi et al., "Reducing INDEL calling errors in whole genome and exome sequencing data," *Genome Medicine*, vol. 6, no. 10, article 89, 2014.

[22] Y. Liu, J. Zhou, and K. P. White, "RNA-seq differential expression studies: more sequence or more replication?" *Bioinformatics*, vol. 30, no. 3, pp. 301–304, 2014.

[23] Y. Liu, J. F. Ferguson, C. Xue et al., "Evaluating the impact of sequencing depth on transcriptome profiling in human adipose," *PLoS ONE*, vol. 8, no. 6, Article ID e66883, 2013.

[24] C. W. Fuller, L. R. Middendorf, S. A. Benner et al., "The challenges of sequencing by synthesis," *Nature Biotechnology*, vol. 27, no. 11, pp. 1013–1023, 2009.

[25] J. Eid, A. Fehr, J. Gray et al., "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[26] J. M. Rothberg, W. Hinz, T. M. Rearick et al., "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, no. 7356, pp. 348–352, 2011.

[27] S. Balzer, K. Malde, A. Lanzén, A. Sharma, and I. Jonassen, "Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim," *Bioinformatics*, vol. 26, no. 18, pp. i420–i425, 2010.

[28] A. Valouev, J. Ichikawa, T. Tonthat et al., "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Research*, vol. 18, no. 7, pp. 1051–1063, 2008.

[29] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[30] P. Tarczy-Hornoch, L. Amendola, S. J. Aronson et al., "A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record," *Genetics in Medicine*, vol. 15, no. 10, pp. 824–832, 2013.

[31] K. S. Poon, K. M. Tan, and E. S. Koay, "Targeted next-generation sequencing of the ATP7B gene for molecular diagnosis of Wilson disease," *Clinical Biochemistry*, vol. 49, no. 1-2, pp. 166–171, 2016.

[32] V. Rehvathy, M. H. Tan, S. P. Gunaletchumy et al., "Multiple genome sequences of *Helicobacter pylori* strains of diverse disease and antibiotic resistance backgrounds from malaysia," *Genome Announcements*, vol. 1, no. 5, Article ID e00687, 2013.

[33] I. Vanni, S. Coco, A. Truini et al., "Next-generation sequencing workflow for NSCLC critical samples using a targeted sequencing approach by ion torrent PGM platform," *International Journal of Molecular Sciences*, vol. 16, no. 12, pp. 28765–28782, 2015.

[34] M. A. Choudhury, W. B. Lott, S. Banu et al., "Nature and extent of genetic diversity of dengue viruses determined by 454 pyrosequencing," *PLOS ONE*, vol. 10, no. 11, Article ID e0142473, 2015.

[35] B. Maranhao, P. Biswas, A. D. H. Gottsch et al., "Investigating the molecular basis of retinal degeneration in a familial cohort of Pakistani decent by exome sequencing," *PLoS ONE*, vol. 10, no. 9, Article ID e0136561, 2015.

[36] A. Webb, A. C. Papp, A. Curtis et al., "RNA sequencing of transcriptomes in human brain regions: protein-coding and noncoding RNAs, isoforms and alleles," *BMC Genomics*, vol. 16, no. 1, article 990, 2015.

[37] A. Shiroma, Y. Terabayashi, K. Nakano et al., "First complete genome sequences of *Staphylococcus aureus* subsp. *aureus* Rosenbach 1884 (DSM 20231T), determined by PacBio single-molecule real-time technology," *Genome Announcements*, vol. 3, no. 4, Article ID e00800, 2015.

[38] O. Harismendy, P. C. Ng, R. L. Strausberg et al., "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome Biology*, vol. 10, no. 3, article R32, 2009.

[39] L. Liu, Y. Li, S. Li et al., "Comparison of next-generation sequencing systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 251364, 11 pages, 2012.

[40] N. Rieber, M. Zapatka, B. Lasitschka et al., "Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies," *PLoS ONE*, vol. 8, no. 6, Article ID e66621, 2013.

[41] M. A. Quail, M. Smith, P. Coupland et al., "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomics*, vol. 13, article 341, 2012.

[42] A. Ratan, W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster, "Comparison of sequencing platforms for single nucleotide variant calls in a human sample," *PLoS ONE*, vol. 8, no. 2, Article ID e55089, 2013.

[43] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, no. 3, pp. 186–194, 1998.

[44] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[45] M. A. Depristo, E. Banks, R. Poplin et al., "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics*, vol. 43, no. 5, pp. 491–501, 2011.

[46] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, article e164, 2010.

[47] P. Cingolani, A. Platts, L. L. Wang et al., "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3," *Fly*, vol. 6, no. 2, pp. 80–92, 2012.

[48] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[49] D. M. Church, V. A. Schneider, K. M. Steinberg et al., "Extending reference assembly models," *Genome Biology*, vol. 16, article 13, 2015.

[50] H. Li, B. Handsaker, A. Wysoker et al., "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[51] M. J. Landrum, J. M. Lee, G. R. Riley et al., "ClinVar: public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research*, vol. 42, no. 1, pp. D980–D985, 2014.

[52] H. L. Rehm, J. S. Berg, L. D. Brooks et al., "ClinGen—the clinical genome resource," *The New England Journal of Medicine*, vol. 372, no. 23, pp. 2235–2242, 2015.

[53] R. C. Green, J. S. Berg, W. W. Grody et al., "ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing," *Genetics in Medicine*, vol. 15, no. 7, pp. 565–574, 2013.

[54] H. Hampel, R. L. Bennett, A. Buchanan, R. Pearlman, and G. L. Wiesner, "A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment," *Genetics in Medicine*, vol. 17, no. 1, pp. 70–87, 2015.

[55] L. M. Amendola, M. O. Dorschner, P. D. Robertson et al., "Actionable exomic incidental findings in 6503 participants: challenges of variant classification," *Genome Research*, vol. 25, no. 3, pp. 305–315, 2015.

[56] M. He, T. N. Person, S. J. Hebbring et al., "SeqHBase: a big data toolset for family based sequencing data analysis," *Journal of Medical Genetics*, vol. 52, no. 4, pp. 282–288, 2015.

[57] R. Daneshjou, Z. Zappala, K. Kukurba et al., "PATH-SCAN: a reporting tool for identifying clinically actionable variants," *Pacific Symposium on Biocomputing*, pp. 229–240, 2014.

[58] W. Zhou, H. Zhao, Z. Chong et al., "ClinSeK: a targeted variant characterization framework for clinical sequencing," *Genome Medicine*, vol. 7, no. 1, article 34, 2015.

[59] P. N. Robinson, S. Köhler, A. Oellrich et al., "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Research*, vol. 24, no. 2, pp. 340–348, 2014.

[60] A. Antanaviciute, C. M. Watson, S. M. Harrison et al., "OVA: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization," *Bioinformatics*, vol. 31, no. 23, pp. 3822–3829, 2015.

[61] E. M. Ramos, C. Din-Lovinescu, J. S. Berg et al., "Characterizing genetic variants for clinical action," *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*, vol. 166, no. 1, pp. 93–104, 2014.

[62] J. A. Tennessen, A. W. Bigham, T. D. O'Connor et al., "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 336, no. 6090, pp. 64–69, 2012.

[63] P. Danecek, A. Auton, G. Abecasis et al., "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, Article ID btr330, pp. 2156–2158, 2011.

[64] M. J. Bamshad, S. B. Ng, A. W. Bigham et al., "Exome sequencing as a tool for Mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[65] Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels, "Three-stage quality control strategies for DNA re-sequencing data," *Briefings in Bioinformatics*, vol. 15, no. 6, pp. 879–889, 2014.

[66] N. Gharani, M. A. Keller, C. B. Stack et al., "The Coriell personalized medicine collaborative pharmacogenomics appraisal, evidence scoring and interpretation system," *Genome Medicine*, vol. 5, no. 10, article 93, 2013.

[67] M. V. Relling and T. E. Klein, "CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network," *Clinical Pharmacology and Therapeutics*, vol. 89, no. 3, pp. 464–467, 2011.

[68] J. J. Swen, M. Nijenhuis, A. de Boer et al., "Pharmacogenetics: from bench to byte—an update of guidelines," *Clinical Pharmacology and Therapeutics*, vol. 89, no. 5, pp. 662–673, 2011.

[69] S. M. Teutsch, L. A. Bradley, G. E. Palomaki et al., "The evaluation of genomic applications in practice and prevention (EGAPP) initiative: methods of the EGAPP working group," *Genetics in Medicine*, vol. 11, no. 1, pp. 3–14, 2009.

[70] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[71] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," *Bioinformatics*, vol. 26, no. 16, Article ID btq330, pp. 2069–2070, 2010.

[72] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Research*, vol. 39, no. 17, article e118, 2011.

[73] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, "SIFT web server: predicting effects of amino acid substitutions on proteins," *Nucleic Acids Research*, vol. 40, no. 1, pp. W452–W457, 2012.

[74] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, and Z. Zhang, "Assessment of computational methods for predicting the effects of missense mutations in human cancers," *BMC Genomics*, vol. 14, supplement 3, article S7, 2013.

[75] S. E. Flanagan, A.-M. Patch, and S. Ellard, "Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations," *Genetic Testing and Molecular Biomarkers*, vol. 14, no. 4, pp. 533–537, 2010.

[76] S. Castellana and T. Mazza, "Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 448–459, 2013.

[77] D. K. Crockett, E. Lyon, M. S. Williams, S. P. Narus, J. C. Facelli, and J. A. Mitchell, "Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants," *The Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 207–211, 2012.

[78] G. P. Jarvik, L. A. Amendola, H. McLaughlin et al., "Performance of ACMG variant classification guidelines within and across 9 CLIA labs in the Clinical Sequeuncing Exploratory Research (CSER) Consortium; (Abstract/Program #1986)," in *Proceedings of the 65th Annual Meeting of the American Society of Human Genetics*, Baltimore, Md, USA, October 2015.

[79] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir et al., "ACMG clinical laboratory standards for next-generation sequencing," *Genetics in Medicine*, vol. 15, no. 9, pp. 733–747, 2013.

[80] N. Aziz, Q. Zhao, L. Bry et al., "College of American Pathologists' laboratory standards for next-generation sequencing clinical tests," *Archives of Pathology & Laboratory Medicine*, vol. 139, no. 4, pp. 481–493, 2015.

[81] M. M. Weiss, B. Van der Zwaag, J. D. H. Jongbloed et al., "Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of dutch genome diagnostic laboratories," *Human Mutation*, vol. 34, no. 10, pp. 1313–1321, 2013.

[82] A. S. Gargis, L. Kalman, D. P. Bick et al., "Good laboratory practice for clinical next-generation sequencing informatics pipelines," *Nature Biotechnology*, vol. 33, no. 7, pp. 689–693, 2015.

[83] B. M. Welch, S. R. Loya, K. Eilbeck, and K. Kawamoto, "A proposed clinical decision support architecture capable of supporting whole genome sequence information," *Journal of Personalized Medicine*, vol. 4, no. 2, pp. 176–199, 2014.

[84] A. N. Kho, L. V. Rasmussen, J. J. Connolly et al., "Practical challenges in integrating genomic data into the electronic health record," *Genetics in Medicine*, vol. 15, no. 10, pp. 772–778, 2013.

[85] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[86] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biology*, vol. 10, no. 11, article R134, 2009.

[87] A. McKenna, M. Hanna, E. Banks et al., "The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[88] M. Niemenmaa, A. Kallio, A. Schumacher, P. Klemelä, E. Korpelainen, and K. Heljanko, "Hadoop-BAM: directly manipulating next generation sequencing data in the cloud," *Bioinformatics*, vol. 28, no. 6, pp. 876–877, 2012.

[89] M. Lawler, L. L. Siu, H. L. Rehm et al., "All the world's a stage: facilitating discovery science and improved cancer care through the global alliance for genomics and health," *Cancer Discovery*, vol. 5, no. 11, pp. 1133–1136, 2015.

[90] B. Paten, M. Diekhans, B. J. Druker et al., "The NIH BD2K center for big data in translational genomics," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1143–1147, 2015.

[91] R. Hazin, K. B. Brothers, B. A. Malin et al., "Ethical, legal, and social implications of incorporating genomic information into electronic health records," *Genetics in Medicine*, vol. 15, no. 10, pp. 810–816, 2013.

[92] J. J. P. C. Rodrigues, I. de La Torre, G. Fernández, and M. López-Coronado, "Analysis of the security and privacy requirements of cloud-based electronic health records systems," *Journal of Medical Internet Research*, vol. 15, no. 8, article e186, 2013.

[93] P. L. Peissig, A. Nikolai, and M. Brilliant, "Personalized medicine," in *Drug Discovery and Evaluation: Pharmacological Assays*, F. J. Hock, Ed., pp. 1–16, Springer, 2015.

[94] J. F. Peterson, E. Bowton, J. R. Field et al., "Electronic health record design and implementation for pharmacogenomics: a local perspective," *Genetics in Medicine*, vol. 15, no. 10, pp. 833–841, 2013.

[95] B. H. Shirts, J. S. Salama, S. J. Aronson et al., "CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1231–1242, 2015.