International Journal of
Molecular Sciences

MDPI

Review

Big Data Analytics for Genomic Medicine

Karen Y. He [1], Dongliang Ge [2,*] and Max M. He [2,3,*]

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA; kyh9@case.edu
[2] BioSciKin Co., Ltd., Nanjing 210042, China
[3] Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA
[*] Correspondence: dgeduke@outlook.com (D.G.); maxm.he@outlook.com (M.M.H.); Tel.: +86-25-8556-6666 (ext. 3308) (D.G.); Fax: +86-25-8532-2781 (D.G.)

Academic Editor: William Chi-shing Cho
Received: 24 October 2016; Accepted: 9 February 2017; Published: 15 February 2017

Abstract: Genomic medicine attempts to build individualized strategies for diagnostic or therapeutic decision-making by utilizing patients' genomic information. Big Data analytics uncovers hidden patterns, unknown correlations, and other insights through examining large-scale various data sets. While integration and manipulation of diverse genomic data and comprehensive electronic health records (EHRs) on a Big Data infrastructure exhibit challenges, they also provide a feasible opportunity to develop an efficient and effective approach to identify clinically actionable genetic variants for individualized diagnosis and therapy. In this paper, we review the challenges of manipulating large-scale next-generation sequencing (NGS) data and diverse clinical data derived from the EHRs for genomic medicine. We introduce possible solutions for different challenges in manipulating, managing, and analyzing genomic and clinical data to implement genomic medicine. Additionally, we also present a practical Big Data toolset for identifying clinically actionable genetic variants using high-throughput NGS data and EHRs.

Keywords: Big Data analytics; clinically actionable genetic variants; electronic health records; healthcare; next-generation sequencing

1. Introduction

Next-generation sequencing (NGS) technologies, such as whole-genome sequencing (WGS), whole-exome sequencing (WES), and/or targeted sequencing, are progressively more applied to biomedical study and medical practice to identify disease- and/or drug-associated genetic variants to advance precision medicine [1,2]. Precision medicine allows scientists and clinicians to predict more accurately which therapeutic and preventive approaches to a specific illness can work effectively in subgroups of patients based on their genetic make-up, lifestyle, and environmental factors [3]. To date, over 6000 Mendelian disorders have been studied at the genetic level [4,5] and over 1500 clinically-relevant complex traits have been studied with genome-wide association study (GWAS) approaches [6]. Clinical research leveraging electronic health records (EHRs) has become feasible as EHRs have been widely implemented [7]. Additionally, a number of studies have been designed to combine genomic and EHR data to improve clinical research and/or healthcare outcome (Table 1).

Leveraging large-scale genomic data with comprehensive clinical data derived from EHRs can implicate disease- and/or drug-associated variants for individualized diagnosis and therapy. NGS technological advancements in clinical genome sequencing and the adoption of EHRs will very likely create patient-centered precision medicine in clinical practice. Genomic data generated by NGS technologies are a vital component in supporting genomic medicine, but the volume and complexity of the data raise challenges for its use in clinical practice [8]. For instance, sequencing a single whole genome generates more than 100 gigabytes of data. Therefore, the development of novel bioinformatics infrastructures is required to implement NGS in clinical practice.

Big Data is a term used to describe data sets with such large volume or complexity that conventional data processing methods are not good enough to deal with them. Big Data has been described disparately by different people [9]. The most popular definition of Big Data is the 5Vs, which are Volume, Velocity, Variety, Verification/Veracity, and Value [10]. The definition of Big Data might be subjected to technological advances in the future. Big Data infrastructure is a framework, which covers important components including Hadoop (hadoop.apache.org), NoSQL databases, massively parallel processing (MPP), and others, that is used for storing, processing, and analyzing Big Data. Big Data analytics covers collection, manipulation, and analyses of massive, diverse data sets that contain a variety of data types including genomic data and EHRs to reveal hidden patterns, cryptic correlations, and other intuitions on a Big Data infrastructure [11]. Due to its effectiveness, Big Data analytics is widely used in different research fields [12]. In this review, we describe how one type of Big Data, genomic data, is applied to improve clinical research and healthcare. We give an overview of the challenges in processing genomic data and EHRs, provide possible solutions to overcome these challenges using approaches that ensure the safety of genomic data, and present a Big Data solution for identifying clinically actionable variants in sequence data. We also discuss the requirement for the efficient integration of genomic information into EHRs.

Table 1. Studies and efforts of leveraging genomic data and EHRs for genomic research/medicine.

Project	Start Year	Aims	Website	Country
deCODE genetics	1996	To utilize population-based genomic data and EHRs to investigate inherited causes of common diseases	http://www.decode.com/ [13]	USA
PMRP	2002	To enroll >20,000 participants to form a resource enabling researchers to study which genes cause diseases, which genes predict reactions to drugs, and how environment and genes work together to cause diseases	http://www.marshfieldresearch.org/chg/pmrp/ [14]	USA
I2B2	2004	To enable clinical researchers to use existing clinical data and genomic data for discovery research; to facilitate the design of targeted therapies for individual patients with diseases having genetic origins	http://www.i2b2.org/ [15]	USA
CKB	2004	To identify the complex interplay between genes and environmental factors on the risks of common chronic diseases	http://www.ckbiobank.org/ [16]	China
eMERGE	2007	To develop methods and best strategies for utilizing EHRs for genomic research in support of implementing genomic medicine	http://emerge-network.org/ [17]	USA
UK Biobank	2007	To improve the prevention, diagnosis, and treatment of a wide range of serious and life-threatening illnesses through a collection of 500,000 volunteers' biosamples and medical records	http://www.ukbiobank.ac.uk/ [18]	UK
GANI_MED	2009	To develop targeted strategies for the prevention, diagnosis, and therapy of diseases, tailored to the specific characteristics of an individual patient or a well-defined patient group. Specifically, these strategies should improve prediction models for health and disease outcomes and also avoid inefficient therapy strategies and adverse side effects	http://www2.medizin.uni-greifswald.de/gani_med/index.php?L=1&id=603 [19]	Germany
KP RPGEH	2009	To explore the genetic and environmental factors that influence common disease	http://www.rpgeh.kaiser.org/ [20]	USA
SCAN-B Initiative	2010	To improve survival and quality of life for breast cancer patients through the introduction of gene expression and genomic tumor profiling into the clinical routine for breast cancer	http://scan.bmc.lu.se/index.php/Main_Page [21]	Sweden
PGPop	2010	To understand how a person's genetic make-up affects his or her response to medications	http://pgpop.mc.vanderbilt.edu/ [22]	USA
MVP	2011	To enroll one million volunteers and use their clinical and genetic data to improve health care for veterans	http://www.research.va.gov/mvp/ [23]	USA
Cancer 2015 Study	2015	To classify cancers molecularly using MPS to promote more targeted treatment of cancer patients and improve patient survival and outcomes	[24]	Australia
Precision Medicine Initiative	2016	To gain better insights into the biological, environmental, and behavioral influences for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle by using the genomic and clinical data of a million Americans	https://www.nih.gov/precision-medicine-initiative-cohort-program [1]	USA

2. Challenges of Handling Genomic and Clinical Data

2.1. Challenges in Manipulating Genomic Data

Although more than 6000 Mendelian disorders have been studied at the genetic level so far, we still do not have a clear understanding of the majority of their roles in health and diseases [25]. Over the past eight years, the size of the NIH sequence read archive (SRA) database has grown exponentially (Figure 1). While the development of NGS technologies has made it increasingly easier to sequence a whole genome or exome, there continue to be considerable challenges in terms of handling, analyzing, and interpreting the genomic information generated by NGS. Since there are over three billion base pairs (sites) on a human genome, sequencing a whole genome generates more than 100 gigabytes of data in BAM (the binary version of sequence alignment/map) and VCF (Variant Call Format) file formats. The actual size of a BAM file is determined by the coverage (the average number of times each base is read; read depth) and read length in a sequencing experiment. Given a 30× WGS data for a single sample, the size of its FASTQ file can be approximately 250 GB, the BAM file can be approximately 100 GB, the VCF file can be about 1 GB, and the annotated files can be approximately 1 GB as well. The approximate file sizes of different NGS data formats and running times of generating those different format files are described in Figure 2. Big Data infrastructures can greatly facilitate the analysis of these data. For example, Big Data-based Burrows-Wheeler Aligner (BWA) can increase the alignment speed 36-fold compared to the original BWA [26]. Currently, most analytical methods for sequencing data use VCF files that assume all "no-call sites" are the same as reference alleles. In fact, many "no-call sites" may be caused by low quality coverage. Therefore, the data quality information, such as coverage and Phred-scaled base quality scores for every site, needs to be utilized to pinpoint whether "no-call sites" are reference-consistent with high coverage or reference-inconsistent caused by low coverage in the downstream data analysis [27]. A number of toolsets for data compression, cloud computing, variant prioritization, copy number variation (CNV) detection, data sharing, and phenotypes on exome sequencing data have been reviewed by Lelieveld et al. [28]. Because VCFs are much smaller than BAM files, analytical tools on VCFs may not always require a Big Data infrastructure. However, researchers are currently facing substantial challenges in storing, managing, manipulating, analyzing, and interpreting WGS data for moderate numbers of individuals if they need to take into account of data quality information stored in BAM files. These challenges will become exacerbated when millions of individuals are sequenced, which embodies the goals of the precision medicine initiative (PMI) in the U.S. and similar efforts of the same scale elsewhere in the world. Leveraging the distribution and scalability inherent in Big Data's infrastructures, it can be feasible to develop a Big Data system to manage and analyze the extensive genomic data compatible with clinical workflows.

2.2. Challenges in Manipulating Clinical Data

Up until the previous decade, approximately 90% of clinicians in the U.S. routinely recorded patient medical records by hand and stored them in color-coded files. In the past five years, the percentage of clinicians using certified EHR systems has grown dramatically [29]. Clinical data extracted from the EHRs for each patient can include the international classification of diseases (ICD) codes, drugs, treatments, procedure (CPT) codes, laboratory values, clinician notes, as well as self-reported dietary and physical activity data. The ICD code is a clinical cataloging system utilized by clinics and hospitals to classify and code diagnosis, symptom, procedure, and treatment in the U.S. Not only are ICD codes used for disease classification, but also as medical billing codes [30]. The volume of clinical data extracted from EHRs can be considerable. For example, the EHR data of ~20,000 patients enrolled in the Personalized Medicine Research Project (PMRP) at Marshfield Clinic is approximately 3.3 GB. The elements in clinical data can be used to classify and measure associations between environmental exposures and clinical consequences. An important application of mining clinical data is patient classification [31–34]. Without stringent and appropriate phenotyping approaches, the classification cannot be appropriately measured, resulting in false positive or negative associations [31]. Machine learning (ML) involves training an algorithm to systematically classify patients into phenotypic groups [35]. To do this, the ML classifier needs to learn which elements in clinical data are providing useful insights for distinguishing the different phenotypic groups. With the proliferation of EHR adoption, computational phenotyping characterization has shown its advantage in classifying research subjects [36]. In addition, millions of data points regarding tens of thousands of clinical elements within the EHRs are available for EHR-based phenotyping. Like sequence data, it will also become a significant challenge to store, manage, manipulate, and mine the complete clinical data of millions of individuals. Therefore, it is necessary to develop advanced and efficient ML approaches for subject characterization and/or better phenotyping. Meanwhile, some ML tasks may take one or two days or even several days to run specific data mining algorithms. For example, to mine large-scale literature, ML approaches on a Big Data infrastructure could be performed 100 times faster than any of the existing ML tools without using any Big Data infrastructures [37].

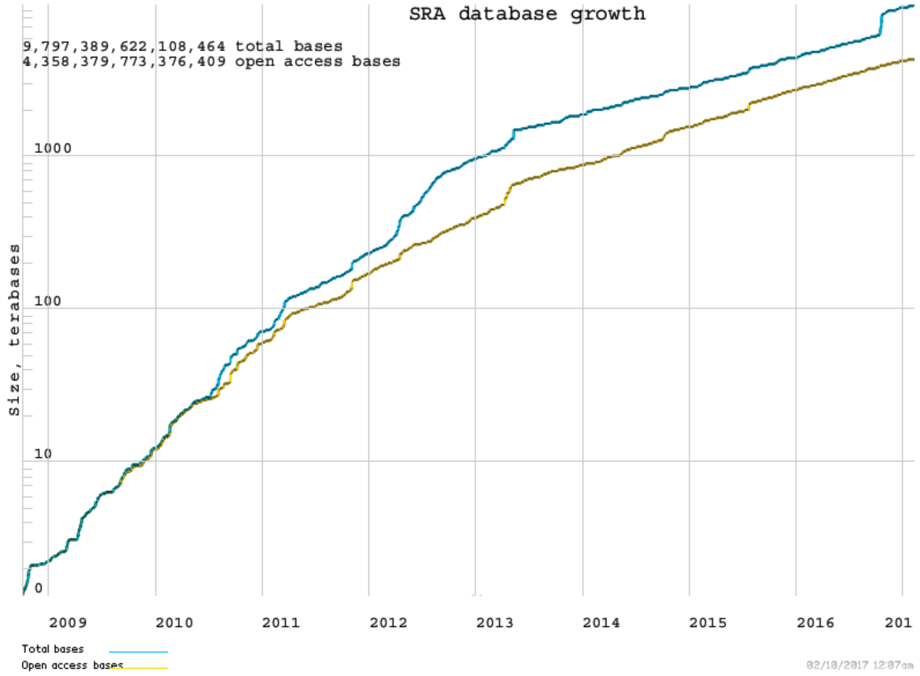

Figure 1. The SRA database growth in the past eight years.

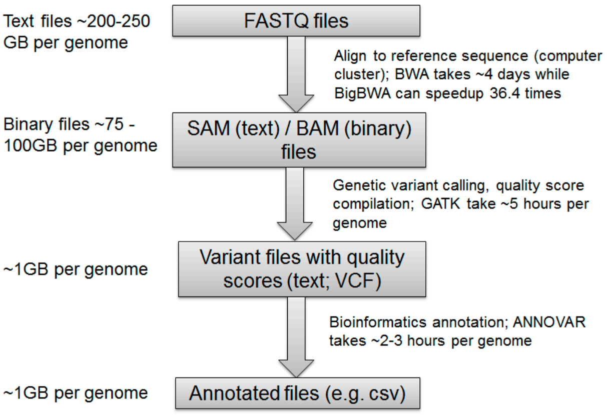
Text files ~200-250 GB per genome
FASTQ files
Align to reference sequence (computer cluster); BWA takes ~4 days while BigBWA can speedup 36.4 times
Binary files ~75 - 100GB per genome
SAM (text) / BAM (binary) files
Genetic variant calling, quality score compilation; GATK take ~5 hours per genome
~1GB per genome
Variant files with quality scores (text; VCF)
Bioinformatics annotation; ANNOVAR takes ~2-3 hours per genome
~1GB per genome
Annotated files (e.g. csv)

Figure 2. The approximately files sizes of different NGS data formats and running times of generating those different format files. BWA: Burrows-Wheeler aligner, GATAK: genome analysis toolkit, BAM: the binary version of sequence alignment/map, FASTQ: a text-based format for representing either nucleotide sequences or peptide sequences, VCF: variant call format.

3. Big Data on the Cloud

3.1. Cloud Computing

Cloud computing providers offer services that provide the infrastructure, software, and programming platforms to clients, and are accountable for the cost for development and maintenance [38]. Compared to creating and maintaining an in-house database, cloud computing is an economical approach to genomic data management because clients pay only for the services that they need. An example of an open-source framework used to develop infrastructure for processing genomic data in a cloud computing environment is Hadoop. It breaks the data into small fragments, distributes them across many data nodes, delivers the computational code to the nodes so that they are processed in parallel, and collectively assembles the results at the end. The parallel processing of many small pieces of data, known as MapReduce, greatly shortens the computing time. Challenges of using cloud computing for genomic data include lengthy data transfers for uploading NGS data to the cloud server, the perceived lack of information safety in cloud computing, and the requirement for developers with advanced programming skills to develop programs on the Hadoop [38].

3.2. Privacy and Security Challenges of Cloud Computing

Cloud computing infrastructures could be deployed with miscellaneous platforms and configurations in which each platform could be configured with diverse security, confidentiality, and authentication settings [39]. These unique aspects could exacerbate security and privacy challenges [40]. Some cloud service providers including Microsoft Azure [41] and Amazon Web Services [42] provide the health insurance portability and accountability act (HIPAA) amenable services for analyzing biomedical data. In addition, data security and privacy in cloud computing is an active research field involving the use of virtual machines [43] and sandboxing techniques [44] for biomedical data management on the cloud.

4. Big Data Analytics in Genomic Studies

4.1. NGS Read Alignment

NGS involves breaking DNA into large amounts of segments. Each segment is called a 'read'. Due to biases in sample processing, library preparation, sequencing-platform chemistry, and bioinformatics methods for genomic alignment and assembly of the reads, the distribution and/or length of reads across the genome can be uneven [45,46]. Therefore, some genomic regions are covered with more reads and others with fewer reads. As mentioned previously, read depth denotes the average number of times each base is read. For instance, a 10× read depth means that each base is present in an average of 10 reads. For RNAseq, read depth is more often designated as number of millions of reads. Read alignment involves lining up the sequence reads to a reference sequence [47,48] to allow comparison of sequence data from a sample sequenced with the reference genome. A number of alignment tools including CloudBurst [49], Crossbow [50], and SEAL [51] have been developed on Big Data infrastructures. More programs designed for short-read sequence alignment are shown in Table S1. Alignment allows a number of quality control (QC) measures, such as the proportion of all reads aligned to a reference sequence, the ratio of unique reads aligned to a reference sequence, and the number of reads aligned at a specific locus. These QC measures affect the accuracy of variant calling.

4.2. Calling Variants

Variant calling is more reliable with higher read depth, which is especially valuable for detecting rare genetic variants with higher confidence. The read depth needed for accurately calling variants relies on various factors, including presence of repetitive genomic regions, error rate of the sequencing platform, and algorithm used for assembling reads into a genomic sequence. Read depth, such as 100× for heterozygous single nucleotide variant (SNV) detection by WES [52], 35× for genotype detection by WGS [53], and 60× for detecting insertions/deletions (INDELs) by WGS [54], may be required. Some widely-used programs for germline variant calling include SAMtools [55], GATK [56], FreeBayes [57], and Atlas2 [58]. SAMtools comprises a number of utilities for manipulating aligned sequence reads and calling SNV and/or INDEL variants. GATK is a NGS analysis suite designed to identify SNVs and INDELs in germline DNA and RNAseq data. It estimates the likelihood of genotype based on the observed sequence reads at a locus by leveraging a Bayesian model. In addition, it employs a MapReduce infrastructure to accelerate the procedure of processing large amounts of sequence aligned reads in parallel [59,60]. Now, it has been expanded to include somatic variant calling tools by incorporating MuTect [61], and to tackle CNVs and structural variations (SVs) as well. The major difference between SAMtools and GATK is the estimation of the genotype likelihood of SNVs and INDELs for calling variants. Regarding the filtering steps, SAMtools uses predefined filters while GATK learns the filters from the data. FreeBayes is a haplotype-based tool that concurrently discovers SNVs, INDELs, multiallelic sites, polyploidy, and CNVs in a sample, pooled multiple samples, or mixed populations [62]. Atlas2 [58] can be used to analyze data generated by the SOLiDTM platform via logistic regression models trained on validated WES data to detect SNVs and INDELs. This tool can also analyze data generated by the Illumina platform using logistic regression models to call INDELs and a mixture of logistic regression and a Bayesian model to call SNVs [63]. To evaluate various programs/tools, Hwang et al. have systematically examined 13 variant calling programs using gold standard personal exome variants [64].

4.3. Variant Annotation

Large amounts of sequence data are being generated by NGS. To pinpoint a small subset of functional variants, many annotation programs have been developed. As one of the most widely used annotation programs, ANNOVAR [65] annotates SNVs, INDELs, and CNVs by exploring their functional consequences on genes, conjecturing cytogenetic bands, and reporting biological functions and various functional scores, including PolyPhen-2 score [66], Sorting Intolerant From Tolerant (SIFT) score [67], the Combined Annotation Dependent Depletion (CADD) score [68], and others. It also discovers variants in conserved regions and identifies variants present in dbSNP [69], the 1000 Genomes Project [70], the NHLBI EPS6500 project [71], and the ExAC [72]. Furthermore, ANNOVAR can employ annotation databases from the UCSC Genome Browser or any other data resources conforming to Generic Feature Format version 3 (GFF3). Other commonly used annotation programs include snpEff [73], and the Ensembl Variant Effect Predictor (VEP) [74]. Xin et al. have developed a web-based service that can be run on the cloud [75]. In order to annotate a WGS data in a short period of time, we are currently developing a cloud-based version of ANNOVAR, which is built on a Hadoop framework and a Cassandra NoSQL database. Additional variant annotation programs are shown in Table S2. In addition, variant annotation depends on biological knowledge in order to provide information on the known or likely impact of variants on gene regulation and protein function [65,73]. 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 [76] variant database, released on 5 July 2016 by the National Center for Biotechnology Information (NCBI), contains 126,315 unique genetic variants with clinical interpretations.

4.4. Statistical Analysis of Genomic Data

Family-based analysis: Family-based NGS data enable the discovery of disease-contributing *de novo* mutations [77–79]. Meanwhile, family-based research strategies can uncover many mutations that may be contributing to recessive, inherited as homozygous or compound heterozygous diseases. SeqHBase [27] is a reliable and scalable computational program that manipulates genome-wide variants, functional annotations and every-site coverage, and analyzes WGS/WES data to identify disease-contributing genes effectively. It is a Big Data-based toolset designed to analyze large-scale family-based sequencing data to quickly discover *de novo*, inherited homozygous, and/or compound heterozygous mutations.

Population-based analysis: A number of large-scale population-based sequencing studies are undergoing. For example, the PMI cohort program attempts to sequence one million or more American participants for improving our ability to preclude and cure diseases based on one's differences in genetic make-up, lifestyle, and environmental factors. By 2025, over 100 million human genomes could be sequenced [80]. Therefore, it is critical to develop statistical toolsets on a Big Data infrastructure for analyzing the genomic data of millions of people.

4.5. Security of Genomic Data

Genomic data need to be protected. Therefore, its privacy and confidentiality should be preserved similarly to other protected health information. Privacy safeguards include the utilization of data encryption, password protection, secure data transmission, auditions of data transferring methods, and the operation of institutional strategies against data breeches and mischievous abuse of the data [81]. The Fair Information Practices Principles (FIPPs) offer a framework for enabling data sharing and usage based on the guidelines adopted by the U.S. Department of Health and Human Services [82]. These principles include: individual access, data correction, data transparency, individual choice, data collection and disclosure limitation, data quality and integrity, safeguards, and accountability. The Workgroup for Electronic Data Interchange (WEDI) has released a report outlining the challenges in regards to the infrastructure, workflows, and coordination of health IT integration [83]. These challenges include data access and integration, data exchange, and data governance. Cloud-computing technology advancements offer easier solutions to store large genomic data files and to consolidate data to make them more easily accessible. The use of cloud computing presents extra security concerns because data storage and/or processing services are provided by an entity external to the healthcare organization. Cloud services qualify as a business associate and they must sign a business associate agreement (BAA) in order to adhere to the modifications to the HIPAA privacy, security, enforcement, and breach notification rules [84]. Cloud service providers can address these concerns by including controlled access to the data and building a user role based access system. Additional security measures should be taken, such as protecting the security of the computer network using warning alarms to monitor when changes are made to stored data, and guaranteeing the complete removal of data from its servers if the cloud storage service is no longer being used [39].

5. Analysis of Genomic and Clinical Data

5.1. Clinically Actionable Genetic Variants

In clinical practice, the identification and return of incidental findings (IFs) for clinically disease-contributing variants in a set of 56 "highly medically actionable" genes associated with 24 inherited conditions have been recommended by the American College of Medical Genetics and Genomics (ACMG) [85,86]. A web-based tool for detecting clinically actionable variants in the 56 ACMG genes is developed by Daneshjou et al. [87], and a variant characterization framework for targeted analysis of relevant reads from high-throughput sequencing data is developed by Zhou et al. [88]. SeqHBase [27] is a bioinformatics toolset for analyzing family-based WGS/WES data on a Big Data infrastructure. To deduce biological perceptions from large amounts of NGS data and inclusive clinical data, we have expanded analysis functions within SeqHBase (Figure 3) to detect disease- and/or drug-associated genetic variants quickly.

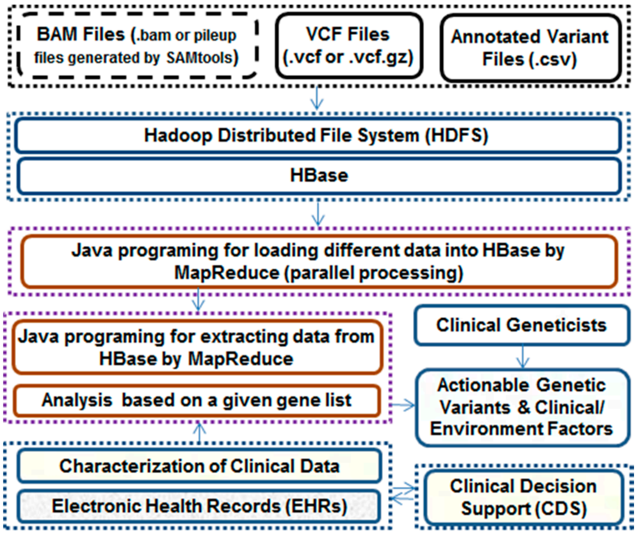
BAM Files (.bam or pileup files generated by SAMtools)
VCF Files (.vcf or .vcf.gz)
Annotated Variant Files (.csv)
Hadoop Distributed File System (HDFS)
HBase
Java programing for loading different data into HBase by MapReduce (parallel processing)
Java programing for extracting data from HBase by MapReduce
Analysis based on a given gene list
Clinical Geneticists
Actionable Genetic Variants & Clinical/ Environment Factors
Characterization of Clinical Data
Electronic Health Records (EHRs)
Clinical Decision Support (CDS)

Figure 3. The basic framework of SeqHBase for identifying clinically actionable genetic variants.

Even though many variant prioritization tools are available, it remains a challenge to detect clinically actionable variants. Additional efforts are required to distinguish truly clinically actionable variants that can be used to guide clinical decisions. As one variant can be classified as different pathogenicity by multiple clinical laboratories [89], more stringent criteria [90] and the latest ACMG guidelines [91] should be complied to report pathogenic variants [92]. To classify the pathogenicity of new variants, which are not recorded in the ClinVar database [76], and to reach some level of concordance on the clinical variant interpretations, assessments from experts, such as medical geneticists, and/or further biological functional studies are needed. To apply actionable results in clinical practice, genetic findings need to be further complemented with highly strong pathological evidence, along with being reviewed by clinical geneticists.

5.2. Clinically Actionable Pharmacogenetic Variants

Substantial efforts have been made to identify clinically actionable pharmacogenetics variants, and it is instructive to review the approach being used. The Coriell Personalized Medicine Collaborative [93], the Clinical Pharmacogenetics Implementation Consortium [94], the Pharmacogenetics Working Group established by the Royal Dutch Association for the Advancement of Pharmacy [95], and the Evaluation of Genomic Applications in Practice and Prevention initiative sponsored by the Centers for Disease Control and Prevention [96] have individually developed similar processes for selecting candidate drugs, reviewing published literature to identify drug-gene associations, scoring evidence supporting associations between genetic variants and drug response, and interpreting the evidence to provide therapeutic guidelines. The approach involving review and interpretation of scientific literature by an expert committee can be considered as the gold standard for determining whether a variant is clinically relevant or actionable, but it also can be costly and labor-intensive. It will not be feasible for experts, either individually or in committees, to review a large number of genetic variants identified in NGS data. Tools such as POLYPHEN-2 [66], VEP [97], Mutation Assessor [98], and SIFT [99] can be used to predict variant effects. However, because these tools are sometimes inaccurate [100] and often differ in their predictions for a same variant [101,102], there will likely be many variants with no clear predicted, clinical interpretation. New methods and toolsets need to be developed to accurately predict the pathogenicity of genetic variants generated by NGS. More importantly, the methods should comply with the U.S. Food and Drug Administration (FDA) guideline [103] and the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline [104].

6. Big Data Analytics in Health Research

Clinical data derived from EHRs have expanded from digital formats of individual patient medical records into high-dimensional data of enormous complexity. Big Data refers to novel technological tools delivering scalable capabilities for managing and processing immense and diverse data sets. On a single level, approaches such as natural language processing (NLP) allow incorporation and exploration of textual data with structured sources. On a population level, Big Data provides the possibility to conduct large-scale exploration of clinical consequences to uncover hidden patterns and correlations [105]. The large amounts of EHRs currently available have enabled us to overcome previously challenging obstacles, such as analyses of rare conditions, more sophisticated analyses, and in-depth analyses of specific data elements [106]. Big Data analytics facilitates the improvement of healthcare from depiction and record to prediction and optimized decision-making.

6.1. Health Informatics

In clinical practice, disease characterization is routinely collected from a number of different streams, such as imaging, pathology, genomics, and electrophysiology. However, much of the deeper insights into disease processes and mechanisms remain to be uncovered and interpreted from routinely acquired clinical data. Clinical data of millions of patients at a clinic/hospital or in a large study (e.g., PMI) exhibit many of the features of Big Data. The volume comes from large amounts of records that can be derived from the EHRs for patients; for example, medical images including magnetic resonance imaging (MRI) or neuroimaging data for each patient can be large, while social media data gathered from a population can be large-scale as well. The velocity occurs when data is accumulated at high speeds, which can be seen when monitoring a patient's real-time conditions through medical sensors for sleep apnea (http://www.sleepapnea.org/), for instance. The variety refers to data sets with a large amount of varying types of independent attributes, such as data sets that are gathered from many different resources. Veracity is a concern when working with possibly noisy, incomplete, or erroneous data where such data need to be properly evaluated using other relevant true evidence. Value portrays the usefulness for improving healthcare outcome. The advance in the fields of health informatics is a vital driving force in the expansion of Big Data, due to either the volume of clinical information produced or the complexity and variety of biomedical data that encompasses discoveries from basic science, translational research, medical system, and population-based study on the determining factors of healthiness. It is essential to develop novel data analytics tools with scalable, accessible and sustainable data infrastructure to effectively manage large, multiscale, and heterogeneous data sets and convert these data into knowledge that can be used for cost-effective decision support, disease management, and healthcare delivery. It is also necessary to develop Big Data infrastructures/systems to store, manage, manipulate, and analyze large-scale clinical data.

6.2. Medical Imaging Analysis

Medical imaging data is a type of Big Data in medical research. Imaging genomics is a rapidly growing field derived from recent advancement in multimodal imaging data and high-throughput omics data. The remarkable complexity of these datasets present critical computational challenges. Kitchen et al. reviewed methods for overcoming the challenges associated with integrating genomic, transcriptomic, and neuroproteomic data [107]. There is an increasing interest in integrating neuroimaging data into frameworks for promoting data mining and meta-analyses [108]. In the past century, a central interest in cognitive neuroscience has been trying to understand the human brain [109,110]. Recently, a group of neuroscience researchers used ML methods to map the human brain in order to understand the incredibly complex human cerebral cortex [111]. Human brain mapping is another monumental step toward precision medicine. For example, an important advance in developing operational management and therapeutics of neurological and psychiatric disease empowers researchers to collect and explore data from approximately 100 billion neurons from the brain in a much greater capacity and at an even more rapid speed. As the human brain controls multiple spatial and chronological scales, the data can be used to understand how the brain works by combining all relevant information [112]. Therefore, developing high-performance computing tools based on a Big Data framework becomes critical to neuroscience for improving healthcare [113,114].

6.3. Data Sharing

In order to share EHRs across multiple healthcare providers, several key components need to be taken into account: (1) *functional interoperability*, which allows data (e.g., medical records) to be exchangeable from one EHR system to other EHR systems without any restrictions; (2) *structural interoperability*, which permits the data structure to be exchangeable across all systems; (3) *semantic interoperability*, which allows multiple systems to exchange data and to easily make use of the data exchanged; and (4) *interpretation*, which allows clinicians to properly interpret the health records (e.g., symptoms) as carrying the same meaning. Currently, several universal EHR providers, including Epic, Cerner, MEDITECH, and Allscripts, have joined to establish an interoperability initiative (http://www.beckershospitalreview.com/healthcare-information-technology/epic-cerner-meditech-dozens-more-make-interoperability-pledge-at-himss16-5-things-to-know.html). All of these vendors have agreed to overcome the challenges of medical record interchangeability, information sharing, and positive patient engagement. The settlement is a critical step towards constructing an allied healthcare system, where information is shared smoothly and in a secure manner across various EHR systems [115].

7. Discussion

In order to maximize preventive measures of serious but preventable diseases, it is critical to understand as much about the patient as early as possible. Generally, precautionary health interventions are simpler or more cost-effective than therapies implemented at a later stage. In addition, knowing patients' individual characteristics is often helpful in providing effective and individualized therapy to a disease because individual patients can respond to the same treatment differently. Genomic medicine could change the path for preventing and treating human diseases. However, the translation of these advances into healthcare will rely critically on our ability to identify disease- and/or drug-associated clinically actionable variants and on our knowledge of the roles of the genetic alterations in the illness procedure.

To conduct pilot studies on incorporating genomic data into clinical care, a number of healthcare systems have developed bioinformatics infrastructures to process NGS data through a group of databases supplementary to the EHRs [116–118]. Most of the infrastructures are 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 requirements. It requires substantial investment in resources and personnel for developing and deploying an efficient bioinformatics infrastructure for incorporating NGS data in clinical care. Thus, healthcare providers might want to consider cooperatively establishing a cloud computing service designed to store and process genomic data securely for the healthcare community. The cost of sequencing instruments may need to be taken into account as part of the infrastructure cost by clinical laboratories. Targeted sequencing instruments are less expensive and generate less data than the ones that perform WGS/WES. Therefore, more laboratories are likely to implement targeted sequencing before attempting to build a framework to support WGS/WES. For instance, a study conducted by Regeneron Genetics Center and the Geisinger Health System has highlighted the value of integrating genomic data and EHRs to uncover a genetic variant that results in reduced levels of triglycerides and a lower risk of coronary artery disease [119]. In addition, a study on the large-scale analysis of more than 50,000 exomes of patients and their EHRs by Regeneron and Geisinger has found clinically actionable genetic variants in 3.5% of individuals and several known and/or potential drug targets as well [120]. However, there are still challenges with integrating genomic data into EHRs in clinical practice, including reliable bioinformatics systems/pipelines that translate raw genomic data into meaningful and actionable variants, the role of human curation in the interpretation of genetic variants, and the requirement for consistent standards to genomic and clinical data [121].

A vital challenge of incorporating genomic data into clinical practice is the lack of standards for generating NGS data, bioinformatics processing, data storage, and clinical decision support. Standards could promote interoperability in data quality. Obedience to standards would enable the routine use of genomic data in clinical care. However, it is challenging to build standards when NGS technology and bioinformatics tools are frequently evolving. Furthermore, approaches to clinical decision support differ among healthcare institutions [116]. Appropriately integrating genomic data with EHRs for the discovery of clinically actionable variants can generate novel insights into disease mechanisms and provide better treatments. To improve our understanding on the nature of the disease from comprehensive EHRs, new methods such as ML, NLP, and other artificial intelligence approaches 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-scale genomic and clinical data and how to interpret discovered variants properly. In the meantime, targeted therapies are not yet available for many important genes, and regulatory issues need to be solved before some useful bioinformatics tools can be applied to clinical setting.

8. Conclusions

In conclusion, as EHRs are exceptionally private, methods of protecting patient data need to make certain that patient information is only shared with those with authorized access. Even with the existing challenges, the prospective advantages that genomic data can bring to healthcare are much more important than the potential disadvantages. The increasing development of integrating genomic data with EHRs may cause concerns, but genomic data will certainly play an important role in advancing genomic medicine only if patient privacy and data security can be strictly protected.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/2/412/s1.

Acknowledgments: Max M. He is greatly appreciative of the support from the leaderships in the Center for Human Genetics and Biomedical Informatics Research Center at Marshfield Clinic Research Foundation. This work was supported by the National Institutes of Health (HL007567 to Karen Y. He, UL1TR000427 to Max M. He); and the MCRF to Max M. He.

Author Contributions: Karen Y. He, Dongliang Ge, and Max M. He contributed to designing, drafting, and performing critical review of the manuscript. Karen Y. He, Dongliang Ge, and Max M. He are the guarantors of the manuscript.

Conflicts of Interest: Dongliang Ge and Max M. He are employed and may hold stock of and/or stock options with BioSciKin Co., Ltd. This does not alter our adherence to the journal's policies. The other authors declare no conflict of interest.

References

1. Collins, F.S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795. [CrossRef] [PubMed]
2. Carter, T.C.; He, M.M. Challenges of identifying clinically actionable genetic variants for precision medicine. *J. Healthc. Eng.* **2016**, *2016*. [CrossRef] [PubMed]
3. Vassy, J.L.; Korf, B.R.; Green, R.C. How to know when physicians are ready for genomic medicine. *Sci. Transl. Med.* **2015**, *7*, 287fs219. [CrossRef] [PubMed]
4. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **2007**, *80*, 588–604. [CrossRef] [PubMed]
5. Brunham, L.R.; Hayden, M.R. Hunting human disease genes: Lessons from the past, challenges for the future. *Hum. Genet.* **2013**, *132*, 603–617. [CrossRef] [PubMed]

6. Welter, D.; MacArthur, J.; Morales, J.; Burdett, T.; Hall, P.; Junkins, H.; Klemm, A.; Flicek, P.; Manolio, T.; Hindorff, L.; et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **2014**, *42*, D1001–D1006. [CrossRef] [PubMed]
7. Gottesman, O.; Kuivaniemi, H.; Tromp, G.; Faucett, W.A.; Li, R.; Manolio, T.A.; Sanderson, S.C.; Kannry, J.; Zinberg, R.; Basford, M.A.; et al. The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genet. Med.* **2013**, *15*, 761–771. [CrossRef] [PubMed]
8. Gullapalli, R.R.; Lyons-Weiler, M.; Petrosko, P.; Dhir, R.; Becich, M.J.; LaFramboise, W.A. Clinical integration of next-generation sequencing technology. *Clin. Lab. Med.* **2012**, *32*, 585–599. [CrossRef] [PubMed]
9. Baro, E.; Degoul, S.; Beuscart, R.; Chazard, E. Toward a literature-driven definition of big data in healthcare. *BioMed Res. Int.* **2015**, *2015*, 639021. [CrossRef] [PubMed]
10. Huang, Q.; Jing, S.; Yi, J.; Zhen, W. *Innovative Testing and Measurement Solutions for Smart Grid*; John Wiley & Sons: Singapore, 2015.
11. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef] [PubMed]
12. Chute, C.G.; Ullman-Cullere, M.; Wood, G.M.; Lin, S.M.; He, M.; Pathak, J. Some experiences and opportunities for big data in translational research. *Genet. Med.* **2013**, *15*, 802–809. [CrossRef] [PubMed]
13. Gulcher, J.R.; Jonsson, P.; Kong, A.; Kristjansson, K.; Frigge, M.L.; Karason, A.; Einarsdottir, I.E.; Stefansson, H.; Einarsdottir, A.S.; Sigurthoardottir, S.; et al. Mapping of a familial essential tremor gene, FET1, to chromosome 3q13. *Nat. Genet.* **1997**, *17*, 84–87. [CrossRef] [PubMed]
14. McCarty, C.A.; Nair, A.; Austin, D.M.; Giampietro, P.F. Informed consent and subject motivation to participate in a large, population-based genomics study: The marshfield clinic personalized medicine research project. *Community Genet.* **2007**, *10*, 2–9. [CrossRef] [PubMed]
15. Butte, A.J.; Kohane, I.S. Creation and implications of a phenome-genome network. *Nat. Biotechnol.* **2006**, *24*, 55–62. [CrossRef] [PubMed]
16. Chen, Z.; Lee, L.; Chen, J.; Collins, R.; Wu, F.; Guo, Y.; Linksted, P.; Peto, R. Cohort profile: The Kadoorie Study of Chronic Disease in China (KSCDC). *Int. J. Epidemiol.* **2005**, *34*, 1243–1249. [CrossRef] [PubMed]
17. Rasmussen-Torvik, L.J.; Stallings, S.C.; Gordon, A.S.; Almoguera, B.; Basford, M.A.; Bielinski, S.J.; Brautbar, A.; Brilliant, M.H.; Carrell, D.S.; Connolly, J.J.; et al. Design and anticipated outcomes of the eMERGE-PGx project: A multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clin. Pharmacol. Ther.* **2014**, *96*, 482–489. [CrossRef] [PubMed]
18. Munoz, M.; Pong-Wong, R.; Canela-Xandri, O.; Rawlik, K.; Haley, C.S.; Tenesa, A. Evaluating the contribution of genetics and familial shared environment to common disease using the UK biobank. *Nat. Genet.* **2016**, *48*, 980–983. [CrossRef] [PubMed]
19. Grabe, H.J.; Assel, H.; Bahls, T.; Dorr, M.; Endlich, K.; Endlich, N.; Erdmann, P.; Ewert, R.; Felix, S.B.; Fiene, B.; et al. Cohort profile: Greifswald approach to individualized medicine (GANI_MED). *J. Transl. Med.* **2014**, *12*, 144. [CrossRef] [PubMed]
20. Hoffmann, T.J.; Kvale, M.N.; Hesselson, S.E.; Zhan, Y.; Aquino, C.; Cao, Y.; Cawley, S.; Chung, E.; Connell, S.; Eshragh, J.; et al. Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. *Genomics* **2011**, *98*, 79–89. [CrossRef] [PubMed]
21. Saal, L.H.; Vallon-Christersson, J.; Hakkinen, J.; Hegardt, C.; Grabau, D.; Winter, C.; Brueffer, C.; Tang, M.H.; Reutersward, C.; Schulz, R.; 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 Med.* **2015**, *7*, 20. [CrossRef] [PubMed]
22. Postmus, I.; Trompet, S.; Deshmukh, H.A.; Barnes, M.R.; Li, X.; Warren, H.R.; Chasman, D.I.; Zhou, K.; Arsenault, B.J.; Donnelly, L.A.; et al. Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat. Commun.* **2014**, *5*, 5068. [CrossRef] [PubMed]
23. Reiber, G.E.; LaCroix, A.Z. Older women veterans in the women's health initiative. *Gerontologist* **2016**, *56* (Suppl. 1), S1–S5. [CrossRef] [PubMed]
24. Wong, S.Q.; Fellowes, A.; Doig, K.; Ellul, J.; Bosma, T.J.; Irwin, D.; Vedururu, R.; Tan, A.Y.; Weiss, J.; Chan, K.S.; et al. Assessing the clinical value of targeted massively parallel sequencing in a longitudinal, prospective population-based study of cancer patients. *Br. J. Cancer* **2015**, *112*, 1411–1420. [CrossRef] [PubMed]

25. Rehm, H.L.; Berg, J.S.; Brooks, L.D.; Bustamante, C.D.; Evans, J.P.; Landrum, M.J.; Ledbetter, D.H.; Maglott, D.R.; Martin, C.L.; Nussbaum, R.L.; et al. Clingen—The clinical genome resource. *N. Engl. J. Med.* **2015**, *372*, 2235–2242. [CrossRef] [PubMed]
26. Abuin, J.M.; Pichel, J.C.; Pena, T.F.; Amigo, J. Bigbwa: Approaching the burrows-wheeler aligner to big data technologies. *Bioinformatics* **2015**, *31*, 4003–4005. [CrossRef] [PubMed]
27. He, M.; Person, T.N.; Hebbring, S.J.; Heinzen, E.; Ye, Z.; Schrodi, S.J.; McPherson, E.W.; Lin, S.M.; Peissig, P.L.; Brilliant, M.H.; et al. Seqhbase: A big data toolset for family based sequencing data analysis. *J. Med. Genet.* **2015**, *52*, 282–288. [CrossRef] [PubMed]
28. Lelieveld, S.H.; Veltman, J.A.; Gilissen, C. Novel bioinformatic developments for exome sequencing. *Hum. Genet.* **2016**, *135*, 603–614. [CrossRef] [PubMed]
29. Jamoom, E.W.; Yang, N.; Hing, E. Adoption of certified electronic health record systems and electronic information sharing in physician offices: United states, 2013 and 2014. *NCHS Data Brief* **2016**, *236*, 1–8.
30. Slee, V.N. The international classification of diseases: Ninth revision (ICD-9). *Ann. Intern. Med.* **1978**, *88*, 424–426. [CrossRef] [PubMed]
31. Wojczynski, M.K.; Tiwari, H.K. Definition of phenotype. *Adv. Genet.* **2008**, *60*, 75–105. [PubMed]
32. Rice, J.P.; Saccone, N.L.; Rasmussen, E. Definition of the phenotype. *Adv. Genet.* **2001**, *42*, 69–76. [PubMed]
33. Gurwitz, D.; Pirmohamed, M. Pharmacogenomics: The importance of accurate phenotypes. *Pharmacogenomics* **2010**, *11*, 469–470. [CrossRef] [PubMed]
34. Samuels, D.C.; Burn, D.J.; Chinnery, P.F. Detecting new neurodegenerative disease genes: Does phenotype accuracy limit the horizon? *Trends Genet.* **2009**, *25*, 486–488. [CrossRef] [PubMed]
35. Richesson, R.L.; Sun, J.; Pathak, J.; Kho, A.N.; Denny, J.C. Clinical phenotyping in selected national networks: Demonstrating the need for high-throughput, portable, and computational methods. *Artif. Intell. Med.* **2016**, *71*, 57–61. [CrossRef] [PubMed]
36. Kho, A.N.; Pacheco, J.A.; Peissig, P.L.; Rasmussen, L.; Newton, K.M.; Weston, N.; Crane, P.K.; Pathak, J.; Chute, C.G.; Bielinski, S.J.; et al. Electronic medical records for genetic research: Results of the emerge consortium. *Sci. Transl. Med.* **2011**, *3*, 79re71. [CrossRef] [PubMed]
37. Ye, Z.; Tafti, A.P.; He, K.Y.; Wang, K.; He, M.M. Sparktext: Biomedical text mining on big data framework. *PLoS ONE* **2016**, *11*, e0162721. [CrossRef] [PubMed]
38. O'Driscoll, A.; Daugelaite, J.; Sleator, R.D. 'Big data', Hadoop and cloud computing in genomics. *J. Biomed. Inf.* **2013**, *46*, 774–781. [CrossRef] [PubMed]
39. Rodrigues, J.J.; de la Torre, I.; Fernandez, G.; Lopez-Coronado, M. Analysis of the security and privacy requirements of cloud-based electronic health records systems. *J. Med. Internet Res.* **2013**, *15*, e186. [CrossRef] [PubMed]
40. Takabi, H.; Joshi, J.B.D.; Ahn, G.-J. Security and privacy challenges in cloud computing environments. *IEEE Secur. Priv.* **2010**, *8*, 24–31. [CrossRef]
41. Calder, B.; Wang, J.; Ogus, A.; Nilakantan, N.; Skjolsvold, A.; McKelvie, S.; Xu, Y.; Srivastav, S.; Wu, J.; Simitci, H.; et al. Windows azure storage: A highly available cloud storage service with strong consistency. In Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, Cascais, Portugal, 23–26 October 2011; pp. 143–157.
42. Fusaro, V.A.; Patil, P.; Gafni, E.; Wall, D.P.; Tonellato, P.J. Biomedical cloud computing with Amazon Web Services. *PLoS Comput. Biol.* **2011**, *7*, e1002147. [CrossRef] [PubMed]
43. Kong, J. A practical approach to improve the data privacy of virtual machines. In Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology, Bradford, UK, 29 June–1 July 2010; pp. 936–941.
44. Aziz, A.; Kawamoto, K.; Eilbeck, K.; Williams, M.S.; Freimuth, R.R.; Hoffman, M.A.; Rasmussen, L.V.; Overby, C.L.; Shirts, B.H.; Hoffman, J.M.; et al. The genomic CDS sandbox: An assessment among domain experts. *J. Biomed. Inf.* **2016**, *60*, 84–94. [CrossRef] [PubMed]
45. Lander, E.S.; Waterman, M.S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **1988**, *2*, 231–239. [CrossRef]
46. Sims, D.; Sudbery, I.; Ilott, N.E.; Heger, A.; Ponting, C.P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **2014**, *15*, 121–132. [CrossRef] [PubMed]

47. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]
48. Church, D.M.; Schneider, V.A.; Steinberg, K.M.; Schatz, M.C.; Quinlan, A.R.; Chin, C.S.; Kitts, P.A.; Aken, B.; Marth, G.T.; Hoffman, M.M.; et al. Extending reference assembly models. *Genome Biol.* **2015**, *16*, 13. [CrossRef] [PubMed]
49. Schatz, M.C. Cloudburst: Highly sensitive read mapping with mapreduce. *Bioinformatics* **2009**, *25*, 1363–1369. [CrossRef] [PubMed]
50. Langmead, B.; Schatz, M.C.; Lin, J.; Pop, M.; Salzberg, S.L. Searching for snps with cloud computing. *Genome Biol.* **2009**, *10*, R134. [CrossRef] [PubMed]
51. Pireddu, L.; Leo, S.; Zanetti, G. Seal: A distributed short read mapping and duplicate removal tool. *Bioinformatics* **2011**, *27*, 2159–2160. [CrossRef] [PubMed]
52. Clark, M.J.; Chen, R.; Lam, H.Y.; Karczewski, K.J.; Euskirchen, G.; Butte, A.J.; Snyder, M. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **2011**, *29*, 908–914. [CrossRef] [PubMed]
53. Ajay, S.S.; Parker, S.C.; Abaan, H.O.; Fajardo, K.V.; Margulies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **2011**, *21*, 1498–1505. [CrossRef] [PubMed]
54. Fang, H.; Wu, Y.; Narzisi, G.; O'Rawe, J.A.; Barron, L.T.; Rosenbaum, J.; Ronemus, M.; Iossifov, I.; Schatz, M.C.; Lyon, G.J. Reducing indel calling errors in whole genome and exome sequencing data. *Genome Med.* **2014**, *6*, 89. [CrossRef] [PubMed]
55. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and samtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
56. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef] [PubMed]
57. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing. Available online: http://arxiv.org/abs/1207.3907 (accessed on 15 October 2016).
58. Evani, U.S.; Challis, D.; Yu, J.; Jackson, A.R.; Paithankar, S.; Bainbridge, M.N.; Jakkamsetti, A.; Pham, P.; Coarfa, C.; Milosavljevic, A.; et al. Atlas2 Cloud: A framework for personal genome analysis in the cloud. *BMC Genom.* **2012**, *13* (Suppl. 6), S19. [CrossRef] [PubMed]
59. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
60. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*. [CrossRef]
61. Cibulskis, K.; Lawrence, M.S.; Carter, S.L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E.S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **2013**, *31*, 213–219. [CrossRef] [PubMed]
62. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *Genomics* **2012**.
63. Bao, R.; Huang, L.; Andrade, J.; Tan, W.; Kibbe, W.A.; Jiang, H.; Feng, G. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inf.* **2014**, *13*, 67–82.
64. Hwang, S.; Kim, E.; Lee, I.; Marcotte, E.M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **2015**, *5*, 17875. [CrossRef] [PubMed]
65. Wang, K.; Li, M.; Hakonarson, H. Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef] [PubMed]
66. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* **2010**, *7*, 248–249. [CrossRef] [PubMed]
67. Kumar, P.; Henikoff, S.; Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **2009**, *4*, 1073–1081. [CrossRef] [PubMed]

68. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef] [PubMed]
69. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]
70. Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; Abecasis, G.R. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]
71. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. Available online: http://evs.gs.washington.edu/EVS/ (accessed on 15 October 2016).
72. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291. [CrossRef] [PubMed]
73. Cingolani, P.; Platts, A.; Wang le, L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. 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* **2012**, *6*, 80–92. [CrossRef] [PubMed]
74. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome Biol.* **2016**, *17*, 122. [CrossRef] [PubMed]
75. Xin, J.; Mark, A.; Afrasiabi, C.; Tsueng, G.; Juchler, M.; Gopal, N.; Stupp, G.S.; Putman, T.E.; Ainscough, B.J.; Griffith, O.L.; et al. High-performance web services for querying gene and variant annotation. *Genome Biol.* **2016**, *17*, 91. [CrossRef] [PubMed]
76. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. Clinvar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **2014**, *42*, D980–D985. [CrossRef] [PubMed]
77. Sanders, S.J.; Murtha, M.T.; Gupta, A.R.; Murdoch, J.D.; Raubeson, M.J.; Willsey, A.J.; Ercan-Sencicek, A.G.; DiLullo, N.M.; Parikshak, N.N.; Stein, J.L.; et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **2012**, *485*, 237–241. [CrossRef] [PubMed]
78. O'Roak, B.J.; Vives, L.; Girirajan, S.; Karakoc, E.; Krumm, N.; Coe, B.P.; Levy, R.; Ko, A.; Lee, C.; Smith, J.D.; et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **2012**, *485*, 246–250.
79. Allen, A.S.; Berkovic, S.F.; Cossette, P.; Delanty, N.; Dlugos, D.; Eichler, E.E.; Epstein, M.P.; Glauser, T.; Goldstein, D.B.; Han, Y.; et al. De novo mutations in epileptic encephalopathies. *Nature* **2013**, *501*, 217–221. [CrossRef] [PubMed]
80. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big data: Astronomical or genomical? *PLoS Biol.* **2015**, *13*, e1002195. [CrossRef] [PubMed]
81. Hazin, R.; Brothers, K.B.; Malin, B.A.; Koenig, B.A.; Sanderson, S.C.; Rothstein, M.A.; Williams, M.S.; Clayton, E.W.; Kullo, I.J. Ethical, legal, and social implications of incorporating genomic information into electronic health records. *Genet. Med.* **2013**, *15*, 810–816. [CrossRef] [PubMed]
82. Baker, D.B.; Kaye, J.; Terry, S.F. Governance through privacy, fairness, and respect for individuals. *EGEMS* **2016**, *4*, 1207. [PubMed]
83. The Workgroup for Electronic Data Interchange. *Issues and Trends in Electronic Genomic Data Exchange*; The Workgroup for Electronic Data Interchange: Washington, DC, USA, 2015.
84. Department of Health and Human Services Office of the Secretary. *Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act*; Department of Health and Human Services, Ed.; Federal Register: Washington, DC, USA, 2013.
85. Green, R.C.; Berg, J.S.; Grody, W.W.; Kalia, S.S.; Korf, B.R.; Martin, C.L.; McGuire, A.L.; Nussbaum, R.L.; O'Daniel, J.M.; Ormond, K.E.; et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **2013**, *15*, 565–574. [CrossRef] [PubMed]
86. Hampel, H.; Bennett, R.L.; Buchanan, A.; Pearlman, R.; Wiesner, G.L. 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. *Genet. Med.* **2015**, *17*, 70–87. [CrossRef] [PubMed]

87. Daneshjou, R.; Zappala, Z.; Kukurba, K.; Boyle, S.M.; Ormond, K.E.; Klein, T.E.; Snyder, M.; Bustamante, C.D.; Altman, R.B.; Montgomery, S.B. Path-scan: A reporting tool for identifying clinically actionable variants. *Pac. Symp. Biocomput.* **2014**, 229–240.
88. Zhou, W.; Zhao, H.; Chong, Z.; Mark, R.J.; Eterovic, A.K.; Meric-Bernstam, F.; Chen, K. Clinsek: A targeted variant characterization framework for clinical sequencing. *Genome Med.* **2015**, *7*, 34. [CrossRef] [PubMed]
89. Van Driest, S.L.; Wells, Q.S.; Stallings, S.; Bush, W.S.; Gordon, A.; Nickerson, D.A.; Kim, J.H.; Crosslin, D.R.; Jarvik, G.P.; Carrell, D.S.; et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *J. Am. Med. Assoc.* **2016**, *315*, 47–57. [CrossRef] [PubMed]
90. Biesecker, L.G. Long QT syndrome and potentially pathogenic genetic variants. *J. Am. Med. Assoc.* **2016**, *315*, 2467. [CrossRef] [PubMed]
91. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.W.; Hegde, M.; Lyon, E.; Spector, E.; et al. 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.* **2015**, *17*, 405–424. [CrossRef] [PubMed]
92. He, K.Y.; Zhao, Y.; McPherson, E.W.; Li, Q.; Xia, F.; Weng, C.; Wang, K.; He, M.M. Pathogenic mutations in cancer-predisposing genes: A survey of 300 patients with whole-genome sequencing and lifetime electronic health records. *PLoS ONE* **2016**, *11*, e0167847. [CrossRef] [PubMed]
93. Gharani, N.; Keller, M.A.; Stack, C.B.; Hodges, L.M.; Schmidlen, T.J.; Lynch, D.E.; Gordon, E.S.; Christman, M.F. The coriell personalized medicine collaborative pharmacogenomics appraisal, evidence scoring and interpretation system. *Genome Med.* **2013**, *5*, 93. [CrossRef] [PubMed]
94. Relling, M.V.; Klein, T.E. CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clin. Pharmacol. Ther.* **2011**, *89*, 464–467. [CrossRef] [PubMed]
95. Swen, J.J.; Nijenhuis, M.; de Boer, A.; Grandia, L.; Maitland-van der Zee, A.H.; Mulder, H.; Rongen, G.A.; van Schaik, R.H.; Schalekamp, T.; Touw, D.J.; et al. Pharmacogenetics: From bench to byte—An update of guidelines. *Clin. Pharmacol. Ther.* **2011**, *89*, 662–673. [CrossRef] [PubMed]
96. Teutsch, S.M.; Bradley, L.A.; Palomaki, G.E.; Haddow, J.E.; Piper, M.; Calonge, N.; Dotson, W.D.; Douglas, M.P.; Berg, A.O. The evaluation of genomic applications in practice and prevention (EGAPP) initiative: Methods of the EGAPP working group. *Genet. Med.* **2009**, *11*, 3–14. [CrossRef] [PubMed]
97. McLaren, W.; Pritchard, B.; Rios, D.; Chen, Y.; Flicek, P.; Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **2010**, *26*, 2069–2070. [CrossRef] [PubMed]
98. Reva, B.; Antipin, Y.; Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **2011**, *39*, e118. [CrossRef] [PubMed]
99. Sim, N.L.; Kumar, P.; Hu, J.; Henikoff, S.; Schneider, G.; Ng, P.C. Sift web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **2012**, *40*, W452–W457. [CrossRef] [PubMed]
100. Gnad, F.; Baucom, A.; Mukhyala, K.; Manning, G.; Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genom.* **2013**, *14* (Suppl. 3), S7.
101. Flanagan, S.E.; Patch, A.M.; Ellard, S. Using sift and polyphen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomark.* **2010**, *14*, 533–537. [CrossRef] [PubMed]
102. Castellana, S.; Mazza, T. Congruency in the prediction of pathogenic missense mutations: State-of-the-art web-based tools. *Brief. Bioinform.* **2013**, *14*, 448–459. [CrossRef] [PubMed]
103. Table of Pharmacogenomic Biomarkers in Drug Labeling. Available online: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm (accessed on 30 July 2016).
104. The Clinical Pharmacogenetics Implementation Consortium (CPIC). Available online: https://www.pharmgkb.org/ (accessed on 30 July 2016).
105. Peters, S.G.; Buntrock, J.D. Big data and the electronic health record. *J. Ambul. Care Manag.* **2014**, *37*, 206–210. [CrossRef] [PubMed]
106. DeFrances, C. Electronic Health Records and "Big Data" for Health Care. Available online: http://www.cdc.gov/nchs/data/bsc/bscpres_defrances_may_2016.pdf (accessed on 15 October 2016).
107. Kitchen, R.R.; Rozowsky, J.S.; Gerstein, M.B.; Nairn, A.C. Decoding neuroproteomics: Integrating the genome, translatome and functional anatomy. *Nat. Neurosci.* **2014**, *17*, 1491–1499. [CrossRef] [PubMed]

108. Laird, A.R.; Eickhoff, S.B.; Fox, P.M.; Uecker, A.M.; Ray, K.L.; Saenz, J.J., Jr.; McKay, D.R.; Bzdok, D.; Laird, R.W.; Robinson, J.L.; et al. The brainmap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res. Notes* **2011**, *4*, 349. [CrossRef] [PubMed]
109. Raichle, M.E. Functional brain imaging and human brain function. *J. Neurosci.* **2003**, *23*, 3959–3962. [PubMed]
110. Albrecht, J.; Kopietz, R.; Frasnelli, J.; Wiesmann, M.; Hummel, T.; Lundstrom, J.N. The neuronal correlates of intranasal trigeminal function—An ALE meta-analysis of human functional brain imaging data. *Brain Res. Rev.* **2010**, *62*, 183–196. [CrossRef] [PubMed]
111. Glasser, M.F.; Coalson, T.S.; Robinson, E.C.; Hacker, C.D.; Harwell, J.; Yacoub, E.; Ugurbil, K.; Andersson, J.; Beckmann, C.F.; Jenkinson, M.; et al. A multi-modal parcellation of human cerebral cortex. *Nature* **2016**, *536*, 171–178. [CrossRef] [PubMed]
112. Dinov, I.; Lozev, K.; Petrosyan, P.; Liu, Z.; Eggert, P.; Pierce, J.; Zamanyan, A.; Chakrapani, S.; van Horn, J.; Parker, D.S.; et al. Neuroimaging study designs, computational analyses and data provenance using the LONI zpipeline. *PLoS ONE* **2010**, *5*. [CrossRef] [PubMed]
113. Alyass, A.; Turcotte, M.; Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genom.* **2015**, *8*, 33. [CrossRef] [PubMed]
114. Dinov, I.D. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *GigaScience* **2016**, *5*, 12. [CrossRef] [PubMed]
115. Lin, C.P.; Stephens, K.A.; Baldwin, L.M.; Keppel, G.A.; Whitener, R.J.; Echo-Hawk, A.; Korngiebel, D. Developing governance for federated community-based EHR data sharing. *AMIA Jt. Summits Transl. Sci. Proc.* **2014**, *2014*, 71–76. [PubMed]
116. Tarczy-Hornoch, P.; Amendola, L.; Aronson, S.J.; Garraway, L.; Gray, S.; Grundmeier, R.W.; Hindorff, L.A.; Jarvik, G.; Karavite, D.; Lebo, M.; et al. A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. *Genet. Med.* **2013**, *15*, 824–832. [CrossRef] [PubMed]
117. Peissig, P.L.; Nikolai, A.; Brilliant, M. Personalized medicine. In *Drug Discovery and Evaluation: Pharmacological Assays*; Hock, F.J., Ed.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 1–16.
118. Peterson, J.F.; Bowton, E.; Field, J.R.; Beller, M.; Mitchell, J.; Schildcrout, J.; Gregg, W.; Johnson, K.; Jirjis, J.N.; Roden, D.M.; et al. Electronic health record design and implementation for pharmacogenomics: A local perspective. *Genet. Med.* **2013**, *15*, 833–841. [CrossRef] [PubMed]
119. Dewey, F.E.; Gusarova, V.; O'Dushlaine, C.; Gottesman, O.; Trejos, J.; Hunt, C.; van Hout, C.V.; Habegger, L.; Buckler, D.; Lai, K.M.; et al. Inactivating variants in ANGPTL4 and risk of coronary artery disease. *N. Engl. J. Med.* **2016**, *374*, 1123–1133. [CrossRef] [PubMed]
120. Dewey, F.E.; Murray, M.F.; Overton, J.D.; Habegger, L.; Leader, J.B.; Fetterolf, S.N.; O'Dushlaine, C.; van Hout, C.V.; Staples, J.; Gonzaga-Jauregui, C.; et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the discovehr study. *Science* **2016**, *354*. [CrossRef] [PubMed]
121. Warner, J.L.; Jain, S.K.; Levy, M.A. Integrating cancer genomic data into electronic health records. *Genome Med.* **2016**, *8*, 113. [CrossRef] [PubMed]