frontiers in Genetics

REVIEW
published: 12 February 2019
doi: 10.3389/fgene.2019.00049

Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics

Pawel Suwinski[1], ChuangKee Ong[2,3], Maurice H. T. Ling[2], Yang Ming Poh[2], Asif M. Khan[2,4] and Hui San Ong[2]**

[1] Malaysian Genomics Resource Centre Berhad, Kuala Lumpur, Malaysia, [2] Centre for Bioinformatics, School of Data Sciences, Perdana University, Serdang, Malaysia, [3] Centre of Genomics Research, Precision Medicine and Genomics, AstraZeneca UK Limited, London, United Kingdom, [4] Graduate School of Medicine, Perdana University, Serdang, Malaysia

OPEN ACCESS

Edited by:
Prashanth N. Suravajhala, Birla Institute of Scientific Research, India

Reviewed by:
Vincenza Colonna, Italian National Research Council (CNR), Italy
Sandeep Kumar Dhanda, La Jolla Institute for Allergy and Immunology (LJI), United States
Rahul Kumar, Columbia University Irving Medical Center, United States

***Correspondence:**
Asif M. Khan
asif@perdanauniversity.edu.my
Hui San Ong
huisan.ong@perdanauniversity.edu.my

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *03 September 2018*
Accepted: *21 January 2019*
Published: *12 February 2019*

Citation:
Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM and Ong HS (2019) Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Front. Genet. 10:49. doi: 10.3389/fgene.2019.00049

There is a growing attention toward personalized medicine. This is led by a fundamental shift from the 'one size fits all' paradigm for treatment of patients with conditions or predisposition to diseases, to one that embraces novel approaches, such as tailored target therapies, to achieve the best possible outcomes. Driven by these, several national and international genome projects have been initiated to reap the benefits of personalized medicine. Exome and targeted sequencing provide a balance between cost and benefit, in contrast to whole genome sequencing (WGS). Whole exome sequencing (WES) targets approximately 3% of the whole genome, which is the basis for protein-coding genes. Nonetheless, it has the characteristics of big data in large deployment. Herein, the application of WES and its relevance in advancing personalized medicine is reviewed. WES is mapped to Big Data "10 Vs" and the resulting challenges discussed. Application of existing biological databases and bioinformatics tools to address the bottleneck in data processing and analysis are presented, including the need for new generation big data analytics for the multi-omics challenges of personalized medicine. This includes the incorporation of artificial intelligence (AI) in the clinical utility landscape of genomic information, and future consideration to create a new frontier toward advancing the field of personalized medicine.

Keywords: big data, exome, personalized medicine, sequencing, precision, analytics

INTRODUCTION

Advances in next generation sequencing (NGS) technologies have resulted in an unprecedented proliferation and deluge of genomic sequence data. Harnessing the information encoded in a person's genome is far-reaching and has been instrumental in assessing the substantial portion of person-to-person variability in response to diagnosis, treatment, and prevention strategies (Seripa et al., 2010). This is done by comparing an individual's genomic information to the DNA sequence of another "reference," leading to a variability map of the population when done at a broader scale. The notion of individual variability dates back to Garrod, who in 1902 coined the term "chemical individuality" (Garrod, 1996). The definition has since become more precise; however,

the "reference" still remains vague because of the heterogeneity that exists in the population genome (Huang et al., 2015; Lim et al., 2015; Spratt et al., 2016; Caswell-Jin et al., 2018). These genetic variations stand to impact significantly on the risk and survival outcome of a patient's health (Dawood et al., 2008; Jemal et al., 2011; Eheman et al., 2012). This factor also points toward the potential challenges for advancing personalized medicine – in the hope of incorporating patient genetics within the management and treatment modalities toward better clinical outcomes.

The conventional approach of using candidate genes alone is not sufficient to explain the differences in disease risks that occur between ethnic groups, let alone individuals. The revolution of genotyping technologies has allowed focus on a specific region of the genome, thus enabling deeper coverage of the variants. This approach was successful in identifying prostate cancer risk loci (8q24 and 17q21) in men of African descent (Haiman et al., 2007, 2011; Yeager et al., 2007), which helped explicate the 50% increased risks of getting prostate cancer in these men (Amundadottir et al., 2006). In contrast to genotyping, the advent of the targeted sequencing approach has enabled the focus on specific regions of interest within the genome. This includes targeted amplicon sequencing and whole exome sequencing (WES). Going broader, the whole genome sequencing (WGS) approach provides the most comprehensive analyses of the entire genome; that is ∼3 billion bases for a single "representative" haploid copy, in the case of a human. Notably, the complete set of protein-coding regions, the exome only constitutes ∼3.09% (over 90 million nucleotides) of the latest release of the human reference genome, GRCh38 (Guo et al., 2017). Compared to WGS, targeted sequencing is a more cost-effective method and delivers a higher coverage, allowing for detection of rare variants. Coverage (breadth) for WES is referred to as capture of coding sequence targets (genes and their flanking regions) and in most cases include 22,000 genes. Coverage (depth) refers to the number of sequences for a locus based on independent reads. For clinical purposes, a target depth of 100× from Illumina machines is considered sufficient.

The lowest cost estimate for running a single WES test has fallen to £382 ($555) per exome, which is ∼3.5 factor lower compared to the lowest cost estimate for WGS using HiSeq X (in Germany), £1,312 ($1,906) (Schwarze et al., 2018). This is in stark contrast to the cost per genome of ∼$100 million, back in 2001 after the completion of the first Human Genome Project (National Human Genome Research Institute, 2016). The significant price reduction has taken the democratization of the sequencing to an entire new plateau.

Whole exome sequencing is attractive for clinical application mainly because it covers actionable areas of the genome to determine the variations in the exon regions and identify causal variants of a disease or disease-causing mutations (Gorski et al., 2016; LaHaye et al., 2016; Gambin et al., 2017; Gupta et al., 2017; Hixson et al., 2017; Mueller et al., 2018; Weigelt et al., 2018). There has been a tremendous boost in the generation of WES data at the population scale. The WES has proven its successful application in discovering of the gene associated with the Miller Syndrome, Mendelian phenotypes (Chong et al., 2015) and complex disorder (O'Roak et al., 2012; Jeste and Geschwind, 2014). Since 2011, WES has been routinely offered as a diagnostic tool in clinical genetics laboratories (Pierson et al., 2011; Yang et al., 2013). WES has since been incorporated into the 1000 Genome Project (Genomes Project et al., 2012), the NHLBI "Grand Opportunity" Exome Sequencing Project (GO-ESP) (Tennessen et al., 2012) and the efforts by the Exome Aggregation Consortium (ExAC) (Lek et al., 2016) to catalog population variants and to identify diseases associated with rare variants. These efforts bring us closer to the development of personalized medicine, by matching specific treatments to the genetic makeup of specific patients for maximum benefit. Recent breakthroughs heralding the new era for personalized medicine include approvals by the United States Food and Drug Administration (FDA) for monoclonal antibody pembrolizumab, targeting tumors expressing PD-L1 (Khoja et al., 2015) and olaparib, a poly(ADP-ribose) polymerase (PARP) inhibitor for ovarian cancer patients who carry mutations in BRCA1/2 genes (Rezende, 2014). More recently, the FDA approved larotrectinib (Vitrakvi), the first targeted therapeutic based on the tumor biomarker, instead of tumor origin in the body (Honey, 2018). The market size of personalized medicine is expected to reach USD 87.7 billion by the year 2023 (Newswire, 2016), while the digital genome market is expected to be worth over 45 billion by 2024 (Global Market Insights, 2017).

Herein, we review the application of WES genomic information in clinical practice. The review covers the big data characteristics of WES, discussing existing biological databases and bioinformatics tools to deal with the big data, including new generation artificial intelligence (AI) platforms. Concluding with the clinical utility landscape of genomic information, and future consideration to creating a new frontier toward advancing the field of personalized medicine.

FROM GENETIC MEDICINE TO GENOMIC MEDICINE, PAVING THE WAY FOR PERSONALIZED MEDICINE

The extent of genomic information utilization in medical practice is strongly linked to the advances in genomic technologies and sciences. The relatively small scope of clinical utility and the slow early uptake can be attributed to the lack of clinical evidence supporting the use of medical genomics in multifactorial diseases. Thus, the early focus on variants with high or near certain genotype–phenotype correlation probability (high penetrance) (Lobo, 2006). It soon became obvious that sequencing data alone is not sufficient to explain the genotype–phenotype correlation for multifactorial diseases, because they are characterized by a complex etiology, with variable genetic and environmental contributions. The genetic risk of developing multifactorial conditions is brought about by small and discrete alterations at the genomic/genic levels at multiple loci. Furthermore, these DNA changes exhibit low-to-medium penetrance power that is highly influenced by external factors related to the environment and lifestyle (Centre for Genetics Education, 2015; Abdullah Said et al., 2018).

Knowing the sequence data was simply not enough to understand the etiological and pathogenic processes in complex diseases – the genomic data had a low predictive power and penetrance.

The sequencing of the human genome was more of a technological achievement rather than scientific. Knowing the exact position of all nucleic acids within the DNA molecule (99.9%) did not automatically mean we understand the functional implications of the sequence (Galas, 2001). To this end, several new initiatives were created to uncover the biological message behind the linear combination of the four nucleotides. One year before the full human genome was published, the HapMap project was launched to document the variations in the genome (Eichler et al., 2007). In 2005, the first GWAS was conducted to annotate medically documented genetic variants (Ikegawa, 2012). Soon, hundreds of studies were underway rapidly generating a clinical context for genomic data. The scope of GWAS was wide-ranging, but for most cases, focused on the risk factors and metabolic pathways related to multifactorial diseases. The GWAS contribution was crucial in uncovering strong polygenic-phenotype associations. Based on the GWAS discoveries, it was possible to identify essential metabolic pathways in many traits and medical conditions, paving the way for the first predictive and prognostic genetic test related to multifactorial diseases, and drug metabolism and response (pharmacogenetics).

In parallel, fast progress was made in researching for somatic variants from cancer cells. Neoplasms can be defined as acquired genetic diseases (∼70% of all cancers), where the etiological genetic component is brought about by environmental factors (Malhotra et al., 2014). The cancerous transformation of cells is closely linked to genetic alterations in specific genes, e.g., proto-oncogenes, tumor suppressor genes, and DNA repair genes. It is possible to characterize the histological type of cancer cells based on the patterns of somatic mutations. This knowledge has been successfully explored to develop a range of cancer genetic tests:

- Predictive (e.g., testing BRCA1/2 genes for the genetic risk of developing Hereditary Breast and Ovarian Cancer) (McCartan and Chatterjee, 2018)
- Prognostic (metastatic potential and conventional treatment response) (Maman and Witz, 2018)
- Targeted treatments (small molecule therapeutics targeting specific gene mutations, e.g., imatinib for c-KIT gene mutation in Chronic Myeloid Leukemia and Gastrointestinal Stromal Tumors) (Druker et al., 2006; Grandori and Kemp, 2018)

Cancer genomics is a well-established field of medical practice and research. It is supported by strong clinical evidence and knowledge through many high-profile projects and initiatives such as:

- COSMIC (Catalogue of Somatic Mutations in Cancer) (Forbes et al., 2017)
- TCGA (The Cancer Genome Atlas) (Weinstein et al., 2013)
- ICGC (International Cancer Genome Consortium) (Zhang et al., 2011)

However, it took another 10 years, after the first GWAS was published, for research activities to elucidate the mechanisms underlying the genotype–phenotype association in multifactorial diseases. Since the environment is an essential modifier of the genetic effect, the inclusion of environmental and genetic factors as well as their combined effect on the downstream biological process in the assessment process is necessary to increase the predictive power of genetic alterations. In this approach, genes rather than single variants are assessed for their functional effect. It is a departure from the main GWAS assessment methodology, where the statistical association between the genetic variant and phenotype is measured without often accounting for the underlying biological process. It was a new concept that needed to be tested. In order to validate the clinical utility of functional genomic analysis, two sets of tools were required: (i) data from every layer of the molecular network involved in the translation of genetic effect to observed phenotype, and (ii) powerful computational tools capable of processing large volumes of data and making associations. It was to this end that numerous projects were launched, either to generate the data or to provide integrated bioinformatics tools for the clinical, functional analysis of the data.

BIG DATA CHARACTERISTICS OF WHOLE EXOME SEQUENCING

In 2025, genomics is expected to surpass the three biggest players in big data domains: Twitter, Astronomy, and YouTube (Stephens et al., 2015). Stephen's team had mapped the key technologies that are needed to support big data genomics, in terms of data acquisition, storage, distribution and analysis. Data in genomics had also been mapped to the five Vs, characteristic of big data (He et al., 2017): volume, velocity, variety, veracity, and value. Below, we present the mapping of WES to not just the five, but the expanded 10 Vs of big data (Firican, 2017):

(i) **Volume – WES data size, which can vary by coverage and number of samples.** For the same sample at about 100× coverage, WES will generate ∼5–6 GB of data. Although this is substantially lesser than ∼90 GB for WGS (AllSeq, 2018) at the same coverage, the data size can grow substantially for a large number of subjects. Variant calling on exome sequence data in ExAC v0.3.1 from 60,706 individuals spanned 540 GB (Karczewski et al., 2017). Nowadays, many research studies involving tens of thousands of samples use WES for cost effectiveness, but it is clear that data generation is not the main issue, instead the bottleneck lies in data processing and analysis.

(ii) **Velocity – speed at which WES data per sample is generated and accumulated.** For example, a sequencing facility in 2013, equipped with 50 or so Illumina HiSeq 2000s and 2500s sequenced four exomes for every whole genome and had a capacity of some 2,000 exomes per week (Perkel, 2013). By 2018, the latest Illumina NovaSeq 6000 System is able to sequence a human genome (30×, >120 GB) at a pace of every 55 min, and an exome

(100×, ∼8 GB) every ∼5 min (Illumina, 2018). This empowers users to high-throughput sequence up to 48 human genomes or close to 500 exomes per run in less than 45 h.

(iii) **Variety – the different attributes of WES data.** One aspect of this can be in terms of the five Ws and one H: what (WES), who (gender/age/ethnicity), why (diseased versus healthy), where (organ/tissue/cell), when (day/month/year), and how (accuracy/coverage)? For example, a 100× (how?) WES dataset (what?) can be generated from a centenarian (who?) with tumor (why?) in the neck and bladder (where?) that is in the late stage (when?).

(iv) **Veracity – confidence or trustability in WES data.** Various sources of errors and confounding factors can affect the confidence or trustability of the sequencing data. For example, because of mismapped shortreads, mosaicism, and sequencing errors, variant callers can end up predicting close to sevenfold more than the ∼3 million variants in an individual human genotype (Robasky et al., 2014). It is challenging to differentiate small mutations from random errors generated during sequencing (Hofmann et al., 2017). Additionally, a major shortcoming of WES is the uneven coverage of sequence reads over the exome targets, contributing to many low coverage regions, which affect the downstream analysis, and thus, hinder accurate variant calling (Wang Q. et al., 2017). For example, some regions are still poorly captured (coverage as low as 10×) in a sample with a high average read depth (>75×), which can cause potentially out-turn in missed variant calls (Hoischen et al., 2014).

(v) **Variability – inconsistencies and multitude of dimensions in WES data.** Inconsistencies can include anomalies and outliers, which can be picked up using analytical methods; it can also include inconsistent speed at which data is loaded into the repository. A patient could have a totally or partially rearranged genome as seen in those with autism in one extreme of anomaly (Tabet et al., 2015). Multiple data dimensions can result from disparate data types and sources.

(vi) **Validity – WES data accuracy and readiness for analysis.** In 2017, the accuracy of various variants calling pipelines was investigated for exome sequencing by the PrecisionFDA Hidden Treasures – Warm Up challenge, a contest run by the FDA to promote more accurate genetic screening (PrecisionFDA, 2017). Edico DRAGEN received the highest overall score and Saphetor was the second. Besides a choice in computational pipeline, sequencing artifact can also affect the search for reliable results in the exome sequencing data, particularly in identifying the properties that distinguish false positive variants from true variants. To overcome this, a trio design strategy (father, mother and child) had been used to filter out (removing sequencing artifacts) and retain true mutations (Patel et al., 2014). As for readiness from raw data to analysis, for example, DeepVariant, using Google Cloud, can take ∼70 min (time estimate does not include mapping) for a whole genome at 30× coverage, and ∼25 min for an exome (DeepVariant, 2016).

(vii) **Vulnerability – WES data security and data breach.** Human genomic data has the potential to reveal sensitive information and is potentially re-identifiable, as such privacy and security are often at risk. Several studies have reported vulnerability of the human genomic data, which enables re-identification of patients from an 'anonymous database' (Homer et al., 2008; Gymrek et al., 2013; Harmanci and Gerstein, 2016). Shringarpure and Bustamante (2015) demonstrated that an individual can be re-identified by repeatedly querying the genome data sets via an open-access Beacon for alleles associated with an individual's genome. Moreover, there are also concerns surrounding the policy and practice of returning genome sequences back to research participants (Wright et al., 2017), whereby substantial resources are required to ensure the safety return of that whole data to individual participants (Kaye et al., 2014).

(viii) **Volatility – how long before the WES data is considered obsolete or irrelevant.** Currently, the driving factor behind WES is the favorable cost, when compared to WGS. WGS is more powerful than WES for the detection of potential disease-causing mutation within WES regions, especially in those regions due to single nucleotide variants (SNVs) (Belkadi et al., 2015). Additionally, WGS is also more comprehensive than WES, and thus more useful when the disease causing variant is not in the exome, as in the case of limb malformation due to mutation in the limb enhancer of sonic hedgehog gene (SHH) (Visel et al., 2009). Thus, in the future, when the cost for WGS reduces to the point of being equivalent or lower than the current cost of WES, then the relevance of WES data becomes questionable. Thus, one may consider that the attractiveness of WES for clinical use is of a limited shelf-life, subject to WGS becoming affordable to the masses. It is estimated that by 2020 or later, the cost for WGS may be as low as USD 100 (Herper, 2017).

(ix) **Visualization – how challenging it is to visualize WES data**. Visualization of sequence data is an important tool for researchers and clinicians, especially those without extensive IT skills. Exome data is currently visualized through various popular browsers (**Table 1**) that provide a gene- and transcript-centric display of variation (Karczewski et al., 2017), with extensive functionality for comparative analysis, as well aggregation of available knowledge. However, plotting graphical representation of NGS data in real-time comes with a cost: higher computational requirements (computing power, memory, and storage) and faster Internet. Additionally, many of the genome Internet viewer use older annotation databases than those installed locally, which might be a significant restriction. For example, viewers only accepting sequences aligned to GRCh37/hg19 assembly (current version GRCh38), support dbSNP version 141 (current version is 151), and Ensembl VEP 85 (current version 94). Increasing complexity of viewing data adds

TABLE 1 | List of biological databases and bioinformatics tools relevant for data-warehousing, alignment, processing or analysis of sequence reads.

Category	Bioinformatics tools	Reference
Read alignment	BWA	Li and Durbin, 2009
	Bowtie	Langmead, 2010
Annotation	Annovar (Qiagen)	Qiagen, 2018a
	Variant Effect Predictor (Ensembl)	McLaren et al., 2016
	SNPsift and SNPeffect	Cingolani et al., 2012
	Variant Annotation Integrator (UCSC)	Hinrichs et al., 2016
	NCBI Variant Annotation	Church et al., 2013
	Sift4G	Vaser et al., 2016
	WGS annotator (runnable on the Amazon Compute Cloud)	Liu et al., 2016a
Visualization	NCBI Variant Viewer	National Center for Biotechnology Information, 2018
	UCSC Genome Browser	Kent et al., 2002
	ENSEMBL Genome Browser	Stalker et al., 2004
	ExAC browser	Karczewski et al., 2017
	Integrative Genomics Viewer (IGV)	Thorvaldsdottir et al., 2013
	Personal Genome Browser (PGB)	Juan et al., 2014
	3D Genome Browser	Wang et al., 2018b
Data-warehousing	ClinVar (clinical significance)	Landrum et al., 2014
	dbSNP (NCBI main variant annotation database)	Sherry et al., 2001
	dbNSFP (variants damage prediction using many *in silico* algorithms)	Liu et al., 2011
	COSMIC (Catalogue of Somatic Mutations in Cancer)	Forbes et al., 2017
	GWAS Catalog	Welter et al., 2014
	GWAS Central	Beck et al., 2014
	Cancer Atlas	Liu et al., 2018
	RefSeq	Pruitt et al., 2005
	PANTHER	Thomas et al., 2003
	TCGA (The Cancer Genome Atlas)	Weinstein et al., 2013
	ICGC (International Cancer Genome Consortium	Zhang et al., 2011
Analytics	Genome Analysis Toolkit (GATK)	DePristo et al., 2011
	MuTect	Cibulskis et al., 2013
	OTG-snpcaller	Zhu et al., 2014
	ASEQ	Romanel et al., 2015
	Halvade-RNA	Decap et al., 2017
	GT-WGS	Wang et al., 2018a
	EXCAVATOR2	D'Aurizio et al., 2016
	KaryoScan	Maxwell et al., 2017
AI-based analytics	Exomiser	Smedley et al., 2015
	DeepVariant	Knight, 2017
	Deep Genomics	Knight, 2017
	Qiagen (Ingenuity Variant Analysis and Ingenuity Pathway Analysis)	QIAGEN, 2018b
	Golden Helix (VarSeq, VSCkinical)	Golden Helix, 2017
	Advaita (iVariant/iPatway/iBio Guides)	ADVAITA, 2018
	Lifemap Sciences	TGex™, 2018

additional needs for storage, as in the example of the 3D Genome Browser requiring at least 10 GB for compressed data (1T for uncompressed). Newer genome viewers utilizing cloud computing technology are gaining popularity as they provide good resource optimization, satisfactory performance and affordability for those requiring commercial license (e.g., DNAnexus).

(x) **Value – usefulness of WES data.** Genomic data has clearly established its fundamental value, while exome data as a focus on the coding sequences does have its contribution in improving health outcomes. For example, WES provides value to the medical system through better ability to give patient-directed care, to anticipate future medical needs and avoid unnecessary interventions. As a diagnosis to a family, it diminishes the need for other testing; and allows new gene discovery and re-analysis of old data with new information (Mayo Clinic, 2017).

The 10 Vs, characteristic of big data are applicable to WES (**Figure 1**), and thus, they naturally extend to WGS. The value each sequencing approach brings would be useful at different levels. The limitation of WES, however, relative to WGS is the focus on the coding sequences. With the expected cost reduction of WGS, it remains to be seen if WES remains useful for discovery and statistical analysis. Nonetheless, targeted sequencing, both WES and amplicon, are expected to remain relevant, similar to genotyping, as a way to concentrate the research resources, akin to "less is more."

NEW GENERATION OF BIG DATA ANALYTICS

NGS Technological Platforms and Approaches

The completion of the human genome project marked the start of an era of significant growth in genome sequencing technologies, termed as "Next Generation Sequencing." This resulted in various NGS techniques, besides WGS and WES, such as RNA-seq, Chip-seq, and Bisulfite-seq and the accompanying development of tools for data analysis (**Table 2**).

There are currently two major approaches in NGS technology, whether performing WES or WGS. Short read sequencing approach, such as by use of Illumina HiSeq X, provides a reduced cost and higher accuracy data, which are geared toward population level studies and clinical variant discovery, whilst, long read approaches, such as by use of PacBio's single molecule real-time (SMRT) sequencing machines, are designed more for *de novo* genome assembly applications or isoforms discovery (Goodwin et al., 2016). Short read massive parallel sequencing has emerged as a standard tool for clinical use (Ardui et al., 2018). However, there are inherent limitations, such as GC bias, difficulties mapping to repetitive elements, trouble discriminating paralogous sequences, and difficulties in phasing alleles. These obstacles can be addressed by long read single molecule sequencers. Additionally, they offer higher consensus accuracies and detection of epigenetic modifications.

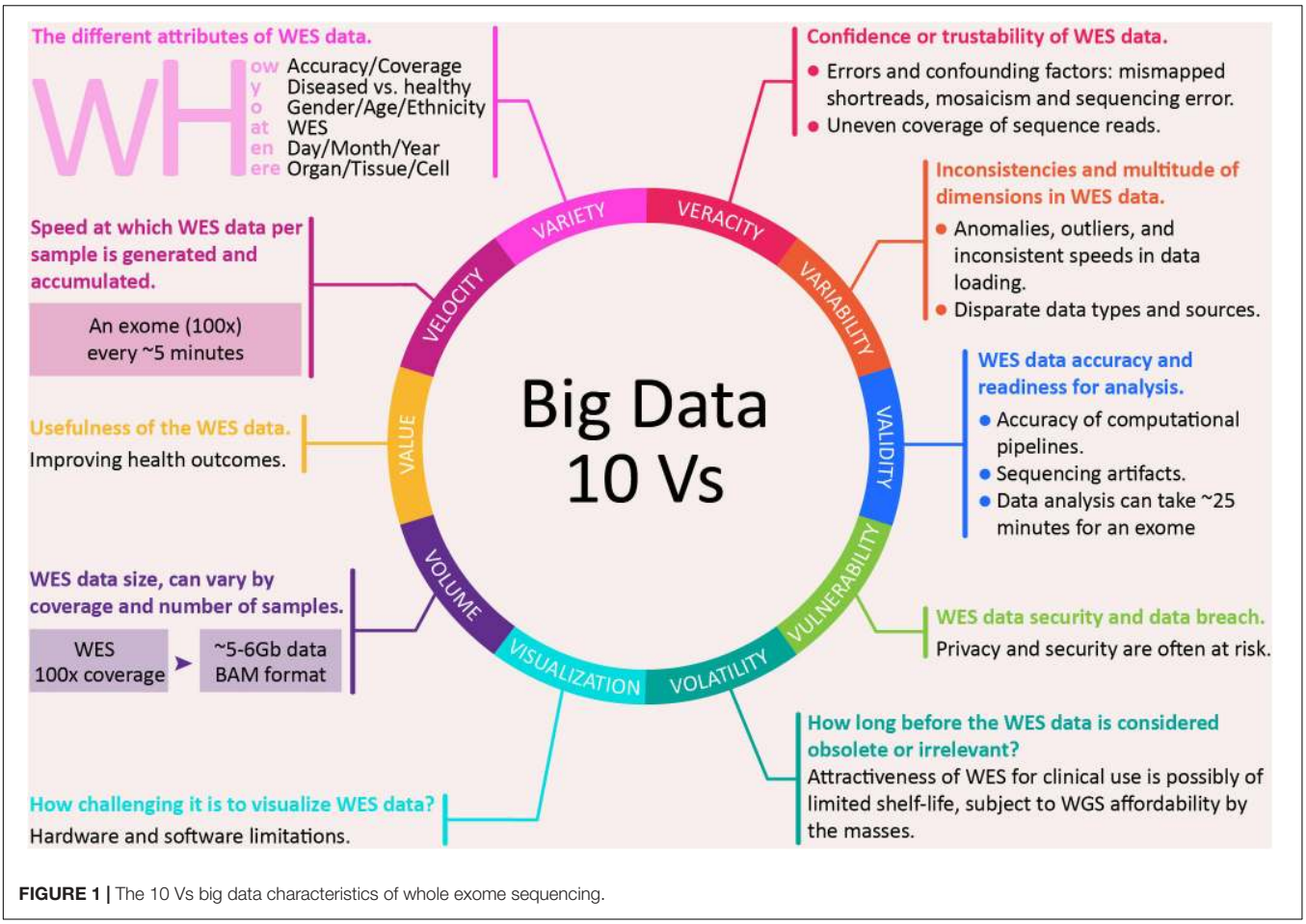

FIGURE 1 | The 10 Vs big data characteristics of whole exome sequencing.

TABLE 2 | Comparison of various NGS technique and primary analysis tools.

NGS techniques	Study aim(s)	Data size per sample	Tool(s) used	Reference
WGS	*De novo* assembly	~90 GB	Velvet, SOAPdenovo	Zerbino and Birney, 2008; Luo et al., 2012
WES	Protein-coding variant identification	~5–6 GB	Edico DRAGEN, GATK, Samtools	Li et al., 2009; McKenna et al., 2010; Edico Genome, 2018
RNA-seq	Gene expression, novel isoform discovery	~3–4 GB	DESeq, Cufflinks	Anders and Huber, 2010; Trapnell et al., 2012
ChIP-seq	Protein–DNA interaction study, i.e., identification of histone marks and transcription factor binding sites	~1–2 GB	QuEST, MACS	Valouev et al., 2008; Liu, 2014
Bisulfite-seq	DNA methylation sites identification	~1–2 GB	BS Seeker	Chen et al., 2010

Nonetheless, their utility in the clinical setting has been limited because of low throughput and high cost.

The WES data can be obtained using different technological platforms. First generation sequencing, e.g., Sanger sequencing, is based on chain termination and electrophoretic separation for the detection of newly incorporated nucleotide. It is a slow and costly process, but a highly accurate method. It is routinely used for confirmation of genomic alteration discovered by other methods (Sanger et al., 1977). To speed up the sequencing process, new technology was developed that uses chemical reaction and optical detection in a massive parallel process. These technologies are often called the NGS or second generation sequencing and include proprietary methods, such as sequencing by synthesis (SOLEXA/Illumina), sequencing by ligation (SOLiD/Life Technologies), pyrosequencing (454/Roche), and semiconductor sequencing (Ion Torrent) (Buermans and den Dunnen, 2014; Kchouk et al., 2017). Each of them has specific application based on their advantages and weak points. Very often the use of these technologies is determined by the length of the reads length, the accuracy of base calling, and the cost per base-pair. Third generation of sequencing technology is characterized by departure from amplification via sequencing of just one DNA

molecule (or one cell DNA) using physical properties of DNA. Oxford Nanopore is one of the industry leading companies that commercialized the technology, which uses electrical impedance to detect the nucleotide passing through a membrane. Some sources distinguish 4th generation of sequencing technology for real-time single molecule sequencing (SMaRT). Although they accuracy is still below the second generation sequencing machines, they are the perfect tools for point-of-care (Laver et al., 2015; Lee et al., 2016).

When performing WES, a key consideration factor is the selection of the exome capture kit, more than the choice of platform. Various commercials kits are available, such as Agilent SureSelect XT, Agilent SureSelect QXT, NimbleGen SeqCap EZ and Illumina Nextera Rapid Capture Exome. They use biotinylated DNA or RNA baits, which are hybridized to genomic fragment libraries. Yet they differ in target region selection, bait length, bait density, molecule used for capture and genomic fragmentation method. If the aim is to detect SNVs and indels in untranslated regions (UTRs), then NimbleGen platform stands-out, while both Agilent XT and Illumina perform similarly for SNV and indel detection in coding regions (Shigemizu et al., 2015).

NGS Data Analysis

Medical conditions that are genetically determined or have a strong genetic component arise from a variety of DNA alterations. These molecular events include SNV [referred to as single nucleotide polymorphism (SNP)] if they occur to some appreciable degree (>1%) in a population) and structural DNA changes, such as copy number variation (CNV), short insertions and deletions (indels), repetitions, large insertions and deletions, translocations (can result in fusion genes), inversions, aneuploidy, and ploidy (Ye et al., 2016). WES is primarily used for the detection of SNV/SNPs and indels within the coding regions of a genome.

Massive parallel sequencing of short reads through NGS generate big data, which has to be aligned (mapped to a reference genome or generate *de novo* genome sequence) for analysis. When a reference genome is available, the first step in data analysis is mapping the reads onto the reference genome (Shang et al., 2014). The intention is to "stack" each reads on the reference genome "floorplan." If the template molecules are mRNA (thus, known as RNA-seq), the "height" of each stack corresponds to the abundance of mRNA for the specific genomic locus (Conesa et al., 2016; Zhao et al., 2018), at the resolution of each nucleotide. In the case where the template molecules are DNA (thus, known as DNA-seq), the "height" of the stack corresponds to the multiple of copy number and number of haploids. In this case, SNPs is rendered as mismatches on the stack (Kumar et al., 2012). In an event where a reference genome is not available, *de novo* assembly or genome annotation can be used. *De novo* assembly is based on the premise that each read may be overlapping and can be used to generate a contig assembly (Cho et al., 2015; Deng et al., 2015; de Sá et al., 2018), much like an assembly from shotgun sequencing (Staden, 1979; Hung et al., 2013). Once a contig is rendered, it can be used as a proxy to a reference genome. Genome annotation (Nagasaki et al., 2013; Menon et al., 2016), on the other hand, is direct analysis of the reads by two steps. In the first step, each read is annotated using tools such as BLAST (Altschul et al., 1990), functional annotations using tools such as InterProScan (Jones et al., 2014), or pathways by sequence similarities to known enzymes. This is sometimes known as read annotation. This is followed by the second step during which reads are mapped onto a scaffold; such as, a genome or a pathway map. When mapped by BLAST to another genome; for example, BLAST of *Bacillus subtilis* NGS data to *Escherichia coli* genome; then *E. coli* genome can be used as a reference genome. As such, there are commonalities between all these methods (mapping to reference genome, *de novo* sequencing, and genome annotation) of data analysis as the end result requires the mapping of reads onto some form of scaffolding substrate. When NGS data is functionally annotated to known proteins or pathways, the set of proteins or pathways will be used as a reference and transcript abundance or SNP calls can be made.

The Broad Institute had developed a set of tools, the Genome Analysis Toolkit or GATK (DePristo et al., 2011), for analysis of reads with the ability to combine various tools within GATK into a workflow for better documentation and reproducibility. GATK can be accessed at https://software.broadinstitute.org/gatk and various example workflows are also publicly available at https://software.broadinstitute.org/gatk. As more tools are added to do GATK, the possibility of workflows is virtually endless. For example, do Valle et al. (2016) had combined GATK and MuTect (Cibulskis et al., 2013), another tool by the Broad Institute and had been included into GATK, for more accurate SNP calls. Hence, it is foreseeable that combinations of existing tools may yield better results than individual tools, which also demonstrates the advantage of workflows. A volume of recent studies (Ahn et al., 2016; Engelhardt et al., 2017; Kim et al., 2017; Coudray et al., 2018; Han et al., 2018) had used GATK for mutation/SNP analysis using WES data. For example, Artomov et al. (2017) performed WES on more than 10,000 patients and analyzed the data using GATK to identify rare variants in hereditary melanoma. From this study, a mutational landscape of cutaneous and ocular melanoma, and implicated Early B Cell Factor 3 (EBF3) as a potential cutaneous melanoma pre-deposition gene. **Table 1** provides a list of biological databases and bioinformatics tools relevant for data-warehousing, alignment, processing or analysis of sequence reads.

Since GATK and MuTect, several other tools had been published, including a number that utilize GATK. For example, OTG-snpcaller (Zhu et al., 2014) combined Ion Torrent's Mapping Alignment Program (TMAP) and GATK for SNP calls. This had been used in WES analyses, leading to the identification of a missense mutation in sodium voltage-gated channel alpha subunit 8 (SCN8A) in a clinical presentation of early infantile epileptic encephalopathy type 13 (Malcolmson et al., 2016). ASEQ (Romanel et al., 2015) is designed to perform gene-level allele-specific expression analysis from genomic and transcriptomic NGS data to identify allele specific features, and had been used to analyze chemotherapy-resistant urothelial carcinoma for insight that can be used to develop new treatment modalities (Faltas et al., 2016). Halvade-RNA (Decap et al., 2017) re-implements GATK workflow to take advantage of parallel

processing to reduce processing time and achieve 93.8% overlaps in variant identification.

Besides SNP calls, tools for detecting structural variations are also developed. For example, CNNdel (Wang J. et al., 2017) uses convolutional neural network on the output from various feature analysis tools to identify structural variations. GT-WGS (Wang et al., 2018a) takes advantage of Amazon Web Services to process NGS data and achieves 99.9% consistency with GATK best practice in SNP and indel calls. CNVs and larger structural changes still can be identified as long as they are limited to exonic regions. This is possible through the application of bioinformatic algorithms capable of accurately measuring read's depth and allelic imbalances in the aligned sequence (BAM file). EXCAVATOR2 and KaryoScan are examples of such methods with the former being able to detect CNVs and the latter large chromosomal aberrations and changes to chromosome numbers (D'Aurizio et al., 2016; Maxwell et al., 2017). WES is not recommended to be used for translocations and repetitions (e.g., tandem repeats), because of their tendency of having break-points or extending beyond genic space (Belkadi et al., 2015).

New Generation Analytics for Multi-Omics Big Data

Although data generation is not an issue with the advent of NGS and there are bioinformatics tools and databases to handle the resulting big data, the upcoming long read, single DNA molecule sequencing, such as the Oxford Nanopore, can offset the volume of data generation from the second generation NGS. However, while the sequencing data can be decreased, the omics data needed for personalized medicine presents higher complexity and is more voluminous than second-generation sequencing data, and would require continuous evolution or new generation of bioinformatics tools and data-warehousing approaches. For example, in April 2016, AstraZeneca announced an integrative genomics initiative to transform drug discovery and development by delivering novel insights into the biology of diseases, identifying new drug targets, supporting patients' selection for clinical trials and matching patients to the therapies most likely to benefit them, a.k.a personalized medicine (Gameiro, 2016). The initiative included collaborations with Human Longevity, The Wellcome Trust Sanger Institute, United Kingdom, and The Institute for Molecular Medicine, Finland. In order to deliver the bold initiative, AstraZeneca established an in-house Centre for Genomics Research, which will sequence and analyze up to two million genome sequences (WGS and WES), including 500,000 samples from their clinical trials by 2026. Working in collaboration with DNAnexus (Business Wire, 2017), the use of a secure cloud-based translational informatics platform was adopted (Business Wire, 2017) to allow for warehousing and analyses of unprecedented massive volume of raw sequencing data rapidly and economically. This was aimed at enabling the processing of samples from thousands of patients per week and the sharing of data easily and safely with collaborators around the world. The platform also provides a secure environment where genetic data can be combined with de-identified clinical data, paving the way for novel scientific insights.

THE CLINICAL UTILITY LANDSCAPE OF GENOMIC INFORMATION

Pharmacogenetics

Personalized medicine, as the tailoring of clinical interventions, is mostly pharmacological, based on a person's ability to respond favorably; for pharmacological agents this entails metabolic capability to process them. The CYP450 family of enzymes are responsible for phase one of xenobiotics metabolism, and their activity can be altered by genetic variants located in their respective genes. Identifying such genetic variants can help in predicting drugs' pharmacokinetics and pharmacodynamics, which can then assist clinicians in selection of interventions that will achieve desirable therapeutic effect without toxicity (Evans and Relling, 2004; Feero et al., 2008; Whirl-Carrillo et al., 2012; Carr et al., 2014). For drugs with a narrow therapeutic range, such as blood-thinning agents, a small functional activity change can result in either a too low or a too high physiological effect that can lead to health complications. Adverse drug reactions (ADRs) are reported to be one of the major causes of morbidity and mortality that can easily be avoided. In the United States, 3% of registered drugs carry FDA recommendation for genetic tests (FDA, 2018).

Cancer Therapeutics

Besides predicting the response to common drugs, genetic information is also used in matching targeted cancer therapeutics (Johannessen and Boehm, 2017). While pharmacogenetics for common drugs detects germline variants, cancer pharmacogenetics is for selecting small molecule inhibitors and analyzing somatic variants from tumor cells. As cancer is predominantly a genetic disease, tumor DNA analysis is routinely deployed for molecular characterization of the cancer cells, as well as treatment prognostics and monitoring. Obtaining tumor samples for genetic analysis can be a challenge if the growth is small or inaccessible. In recent years, liquid biopsy has been successfully applied to obtain a tumor circulating free DNA. It is now possible to use liquid biopsy for early cancer detection, prognostics, and treatment selection and monitoring. Unfortunately, the cost of cancer genetic tests and targeted treatments are still very high, making them inaccessible in less developed countries.

Reproductive Health

Reproductive health is another area that has benefited from WGS and WES. Shallow WGS (3X) is performed for preimplantation assessment of embryos. It is also used for gender selection. Non-Invasive Prenatal Test (NIPT) is a combination of liquid biopsy and WGS for the detection of trisomies or other large chromosomal rearrangements in the fetus cells. It is possible to replace WGS with a higher coverage WES for both tests, which could make the tests more affordable (Pray, 2008).

Multifactorial Diseases

The clinical utility of genomic information for multifactorial diseases still lacks enough predictive power and strong scientific evidence. However, the advances in bioinformatics technologies,

allowing multi-omics analysis, is showing promising results. There are already reports about polygenic risk score for complex medical conditions attaining similar predictive power as genetic risk assessment for monogenic diseases (Khera et al., 2018).

THE RISE OF ARTIFICIAL INTELLIGENCE

AI-Driven Genomics

High costs and limitations in terms of technologies have remained the main barriers for the greater omics-based implementation of personalized medicine. AI-driven machines, are being deployed to cut costs, especially in overcoming the enormous volume of collected patient data. For instance, Congenica's Sapientia uses the Exomiser tool to accelerate the annotation and prioritization of variants from whole-exome sequencing in the diagnosis of rare diseases (Smedley et al., 2015). Sapientia empowers clinical decision-making by organizing the data into an easily comprehensible fashion, which helps to cut diagnosis times down from 5 years to 5 days (Congenica, 2018). AI-driven machines are even predicted to perform better than humans, from driving a truck (as autonomous vehicles) by 2027, writing a bestselling book by 2049, to performing a surgery by 2053 (Grace et al., 2018).

Meanwhile, tech giants, such as Google and its competitors are furiously adding machine-learning features to their cloud platforms in an effort to attract people to tap into the latest AI techniques (Knight, 2017). For instance, Deep Genomics uses deep learning to tease out genetic causes of diseases and potential drug therapies, and Wuxi's Nextcode, which invested heavily in machine learning methods, are among the companies behind such efforts.

The Google Brain team, a group that focuses on developing an AI application and Verily, another Alphabet subsidiary that focuses on life sciences, released a tool known as DeepVariant that uses the latest AI techniques to construct a more accurate picture of a person's genome from their sequencing data (Knight, 2017). It automatically identifies insertion, deletion and single-base-pair mutation in sequencing data. Millions of high-throughput reads and genomes from the Genome in a Bottle (BIAB) project, No Author (2015), were collected to feed the data to the deep-learning system and the parameters of the model was painstakingly tweaked until it learned to interpret the sequences data with a high level of accuracy (Knight, 2017). In 2016, DeepVariant won the first place in the PrecisionFDA Truth Challenge, in the best SNP performance category, and thus highly accurate. DeepVariant is also extensively fast, robust, cost efficient, flexible, easy to use, and where you need it by using Google Cloud Platform (DeepVariant, 2016).

Omics Analytics Powered by AI Technologies

AI can improve statistical computation, but it needs more data to do the guess-work (Lopes et al., 2012; Topol, 2014; Carter and He, 2016; Camacho et al., 2018). Although the size of NGS data is significantly dropping, thanks to the introduction of single-molecule sequencing (Oxford Nanopore) (Rabbani et al., 2014; Halvaei et al., 2018), the downstream AI analysis requires exponential volumes of longitudinal data for making the genotype–phenotype connection as accurate as possible. While the quest for more robust causal algorithms is underway, a number of bioinformatics tools have been developed aiming to link sequence variants with biological metadata and phenotype. These new generation tools provide *in silico* assessment of omics data, derived from WGS or WES, and analytical capacity (often deploying AI) for variants prioritization/phenotype scoring (Shihab et al., 2013; Liu et al., 2016b). This approach has already proven to generate sufficient predictive power that can be compared to the prediction of Mendelian diseases (Khera et al., 2018). The next step entails the translation of scientific findings into easily understood medical standards, similarly to how pathology test results are reported, and there are already available templates developed for reporting WES findings.

Still, it might take a decade before the new technologies will enter mainstream medicine. The main reason for the slow adoption of genomic information, besides regulatory barriers, is the clinicians' readiness and acceptance of incorporating the NGS findings into their routine case management (Metcalfe et al., 2009; Vassy et al., 2015a). Having clinician-friendly reporting will definitely speed-up the uptake process (Vassy et al., 2015b; Manolio, 2017).

In recent years, some companies have made inroads into NGS clinical reporting using omics analytics powered by AI technologies. In the industry sector of integrated WGS/WES clinical reporting, there are at least four commercial entities that offer clinician-friendly analytics and reporting:

- Qiagen (Ingenuity Variant Analysis and Ingenuity Pathway Analysis) (QIAGEN, 2018b)
- Golden Helix (VarSeq, VSCkinical) (Golden Helix, 2017)
- Advaita (iVariant/iPatway/iBio Guides) (ADVAITA, 2018)
- Lifemap Sciences (TGex^{TM}, 2018)

All four solutions are available through a Web-based interface and offer clinical prioritization (using as input Variant Calling File – VCF) that deploys some aspect of AI. Qiagen applications (Annovar is part of the suite) are the clear leaders, as traditionally most genomic laboratory companies use their offerings (Krämer et al., 2014). In terms of innovation, the sheer depth of knowledge and the ease of generating clinical reports makes Lifemap Sciences and its clinical exome analysis suite (TGex) the top scorer (Ben-Ari Fuchs et al., 2016; Stelzer et al., 2016a,b). It is by far the most clinician-friendly WES analysis pipeline and reporting. It is also one of the most affordable on the market. It compiles over 110 different biological databases, ranging from gene ontology (GO) and biological pathways, through network interdependencies, transcriptional expression, and ending at phenotype essentialities. Because of its coverage of omics data (width and depth), many bioinformatics analytical tools utilize their resources including the companies mentioned above.

AI is most commonly deployed at two levels within clinical bioinformatics: *in silico* gene damage scoring (mostly Markov Hidden Model) (Liu et al., 2016b; McLaren et al., 2016; Feng, 2017) and prioritization and phenotype scoring, where various

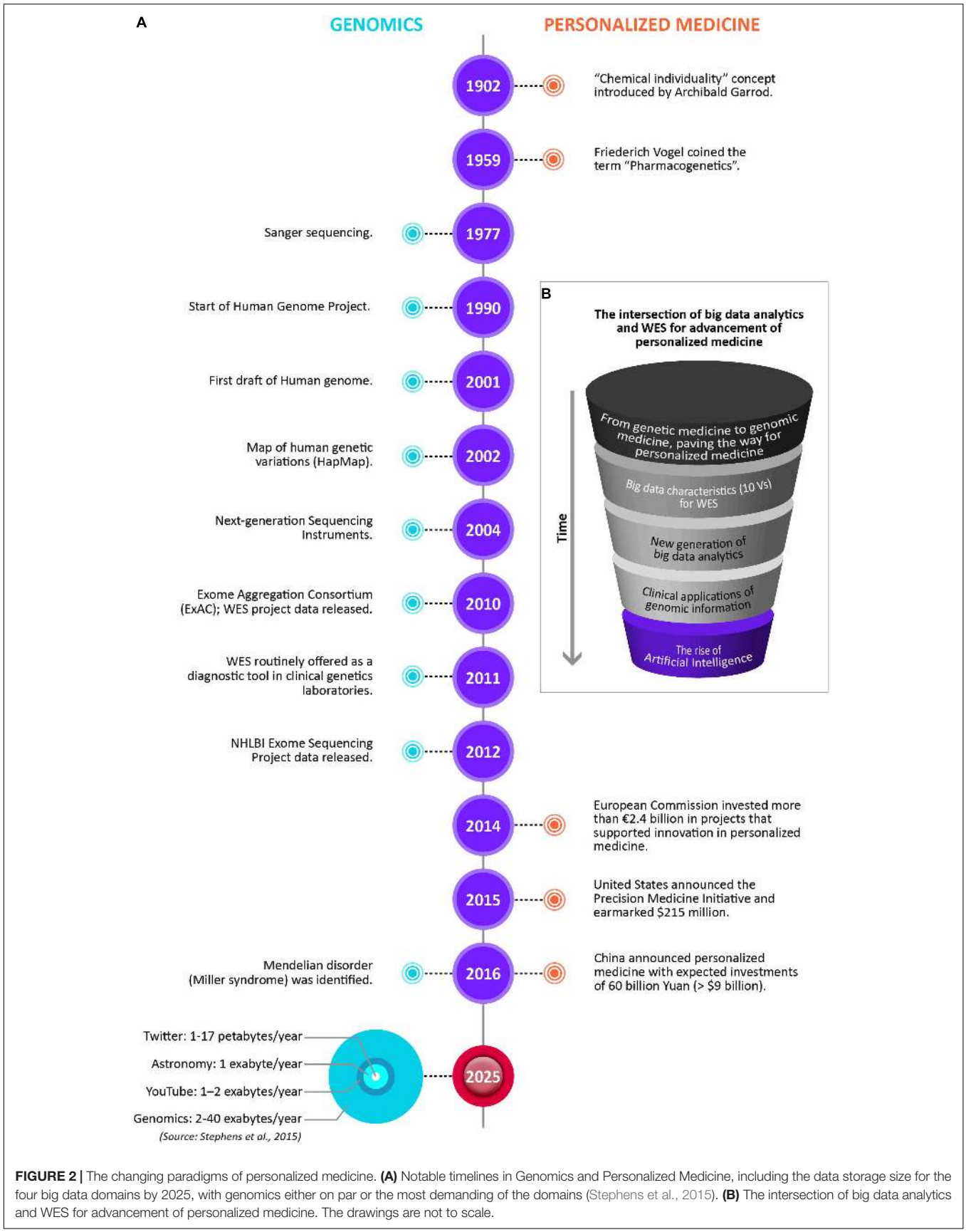

FIGURE 2 | The changing paradigms of personalized medicine. **(A)** Notable timelines in Genomics and Personalized Medicine, including the data storage size for the four big data domains by 2025, with genomics either on par or the most demanding of the domains (Stephens et al., 2015). **(B)** The intersection of big data analytics and WES for advancement of personalized medicine. The drawings are not to scale.

text mining algorithms are adopted. Other than that, AI is still a research tool until large longitudinal data, and more robust informatics frameworks are available. It is worth mentioning that one of the main strengths of AI in clinical practice is the area of image recognition (Alyass et al., 2015; He et al., 2017; Lytras and Papadopoulou, 2018). Many research studies are incorporating AI image processing with pathology and clinical imaging to improve diagnostic decision-making.

Artificial intelligence tools incorporating omics data are still a nascent development; they are a valuable addition to the existing bioinformatics application arsenal and a valuable connection between medical molecular geneticists and frontline clinicians.

FUTURE CONSIDERATIONS

The advancement of personalized medicine in many ways is being driven by the intersection of big data analytics and WES. **Figure 2** illustrates the changing paradigms of personalized medicine. Notable timelines in Genomics and Personalized Medicine are showcased, including the data storage size of the four big data domains by 2025, with genomics either on par or the most demanding of the domains (Stephens et al., 2015). However, there are many barriers still for WES to have a wider use in mainstream medical practice. The major challenges include results reproducibility, reporting standards, and affordability.

Results Reproducibility

A recent study conducted by the American College of Medical Genetics and Genomics (ACMG) showed significant variability in results reproducibility between different genetic laboratories. As a result, the Association together with industry players have developed the standards for genetic tests assessment. Although it is still voluntary, laboratories are encouraged to validate their products against industry standards (Amendola et al., 2016).

Reporting Standards

Standards for reporting results of genetic tests have also been developed by ACMG; however, they only address pathogenic variants detected in ACMG recommended 59 genes. The Harvard School of Medicine, in collaboration with Healthcare Partners, designed a more comprehensive template for reporting results related to genetic diseases, polygenic/multifactorial diseases, and pharmacogenetics (Vassy et al., 2015b). It has to be noted that the polygenic risk score is based on odds ratios reported in the GWAS database (**Table 1**). Conditions with variants without odds ratios or *P*-value score cannot be assessed. The main objective of the reporting template was to present genomic tests results in a clinician-friendly manner so that it can be even used at all levels of health care services, including primary care physicians (Metcalfe et al., 2009; Harvard Medical School, 2015; Manolio, 2017).

For omics based genomic analysis, no standard template exists and each laboratory reports use their own standards. As the goal of multi-omics prioritization is to detect variants, functional effect on genes and possible genes' phenotypic essentiality, the practical way of reporting would be to focus on the Loss of Function/Partial Loss of Function (LOF/PLOF) and phenotype essentiality scoring.

Affordability

Generally, genetic tests are expensive (Topol, 2014; Kong et al., 2015; Bomba et al., 2017; He et al., 2017). The tests can be divided into two technological groups: genotyping and sequencing. Genotyping tests are less costly (USD 100–400), but analyze a limited number of variants, genes (regions). Since scientific progress produces new information on a daily basis, genotyping tests need to be repeated when current findings are included. Sequencing, on the other hand, is much more expensive (>USD 400) but detects variants at any location within the queried region. Also, the clinical utility of sequencing is higher, and there is no need for repeated tests. The cost of WGS is still prohibitive (>USD 1000) for routine application in medical practice. It also produces a large amount of unusable data. A more practical approach is the sequencing of all coding and flanking regions (WES), which covers between 3 and 6% of the genome, and the cost for commercial use can be as low as USD 400 (Macrogen Korea, 2017). The affordability, actionable data, no repeated tests required, and lower junk data makes WES a genetic test of choice (Vissers et al., 2017). Unfortunately, the total cost of WES clinical interpretation is still high (>USD 1000), which makes it more of a premium service rather than first line modality. The affordability of WGS and WES sequencing tests can be dramatically increased provided health insurance companies agree to reimburse the cost.

AUTHOR CONTRIBUTIONS

PS, CKO, MHTL, YMP, AMK, and HSO contributed in the writing of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Ms. Nur Atiqah Azhar for her assistance in the artwork of **Figures 1**, **2** herein.

REFERENCES

Abdullah Said, M., Verweij, N., and Van Der Harst, P. (2018). Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK biobank study. *JAMA Cardiol.* 3, 693–702. doi: 10.1001/jamacardio.2018.1717

ADVAITA (2018). *ADVAITA [Online]*. Available at: https://apps.advaitabio.com/oauth-provider (accessed December 22, 2018).

Ahn, D. H., Ozer, H. G., Hancioglu, B., Lesinski, G. B., Timmers, C., and Bekaii-Saab, T. (2016). Whole-exome tumor sequencing study in biliary cancer patients with a response to MEK inhibitors. *Oncotarget* 7, 5306–5312. doi: 10.18632/oncotarget.6632

AllSeq (2018). *WGS vs. WES*. Available at: http://allseq.com/kb/wgsvswes/ [accessed November 16, 2018].

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genomics* 8:33. doi: 10.1186/s12920-015-0108-y

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., et al. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am. J. Hum. Genet.* 98, 1067–1076. doi: 10.1016/j.ajhg.2016.03.024

Amundadottir, L. T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A., Agnarsson, B. A., et al. (2006). A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* 38, 652–658. doi: 10.1038/ng1808

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 46, 2159–2168. doi: 10.1093/nar/gky066

Artomov, M., Stratigos, A. J., Kim, I., Kumar, R., Lauss, M., Reddy, B. Y., et al. (2017). Rare variant, gene-based association study of hereditary melanoma using whole-exome sequencing. *J. Natl. Cancer Inst.* 109:djx083. doi: 10.1093/jnci/djx083

Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., and Brookes, A. J. (2014). GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* 22, 949–952. doi: 10.1038/ejhg.2013.274

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., et al. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5473–5478. doi: 10.1073/pnas.1418631112

Ben-Ari Fuchs, S., Lieder, I., Stelzer, G., Mazor, Y., Buzhor, E., Kaplan, S., et al. (2016). GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. *OMICS* 20, 139–151. doi: 10.1089/omi.2015.0168

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* 18, 77–77. doi: 10.1186/s13059-017-1212-4

Buermans, H. P., and den Dunnen, J. T. (2014). Next generation sequencing technology: advances and applications. *Biochim. Biophys. Acta* 1842, 1932–1941. doi: 10.1016/j.bbadis.2014.06.015

Business Wire (2017). *DNAnexus to Partner With AstraZeneca's Centre for Genomics Research*. Available at: https://www.businesswire.com/news/home/20170523005582/en/DNAnexus-Partner-AstraZeneca%E2%80%99s-Centre-Genomics-Research [accessed August 6, 2018].

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173, 1581–1592. doi: 10.1016/j.cell.2018.05.015

Carr, D., Alfirevic, A., and Pirmohamed, M. (2014). Pharmacogenomics: current State-of-the-Art. *Genes* 5, 430-443. doi: 10.3390/genes5020430

Carter, T. C., and He, M. M. (2016). Challenges of identifying clinically actionable genetic variants for precision medicine. *J. Healthc. Eng.* 2016:3617572 doi: 10.1155/2016/3617572

Caswell-Jin, J. L., Gupta, T., Hall, E., Petrovchich, I. M., Mills, M. A., Kingham, K. E., et al. (2018). Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genet. Med.* 20, 234–239. doi: 10.1038/gim.2017.96

Centre for Genetics Education (2015). Fact sheet 11 – Environmental and genetic interactions. *Centre Genet. Educ.* 1–3.

Chen, P. Y., Cokus, S. J., and Pellegrini, M. (2010). BS seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11:203. doi: 10.1186/1471-2105-11-203

Cho, N., Hwang, B., Yoon, J. K., Park, S., Lee, J., Seo, H. N., et al. (2015). De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries. *Nat. Commun.* 6:8351. doi: 10.1038/ncomms9351

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., et al. (2015). The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* 97, 199–215. doi: 10.1016/j.ajhg.2015.06.009

Church, D., Sherry, S., Phan, L., Ward, M., Landrum, M., and Maglott, D. (2013). *Variation Overview*. Available at: http://www.ncbi.nlm.nih.gov/variation [accessed November 28, 2018].

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. doi: 10.1038/nbt.2514

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., et al. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a New Program, SnpSift. *Front. Genet.* 3:35. doi: 10.3389/fgene.2012.00035

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

Congenica (2018). *Artificial Intelligence & Machine Learning in Genomics*. Available at: https://www.congenica.com/2018/01/09/artificial-intelligence-machine-learning-genomics/ [accessed November 18, 2018].

Coudray, A., Battenhouse, A. M., Bucher, P., and Iyer, V. R. (2018). Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* 6:e5362. doi: 10.7717/peerj.5362

D'Aurizio, R., Pippucci, T., Tattini, L., Giusti, B., Pellegrini, M., and Magi, A. (2016). Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* 44:e154. doi: 10.1093/nar/gkw695

Dawood, S., Broglio, K., Gonzalez-Angulo, A. M., Buzdar, A. U., Hortobagyi, G. N., and Giordano, S. H. (2008). Trends in survival over the past two decades among white and black patients with newly diagnosed stage IV breast cancer. *J Clin. Oncol.* 26, 4891–4898. doi: 10.1200/JCO.2007.14.1168

de Sá, P. H. C. G., Guimarães, L. C., Graças, D. A. D., de Oliveira Veras, A. A., Barh, D., Azevedo, V., et al. (2018). "Chapter 11 next-generation sequencing and data analysis strategies, tools, pipelines and protocols," in *Omics Technologies and Bio-Engineering*, eds D. Barh and V. Azevedo, 191–207. Belém: Federal University of Para.

Decap, D., Reumers, J., Herzeel, C., Costanza, P., and Fostier, J. (2017). Halvade-RNA: parallel variant calling from transcriptomic data using MapReduce. *PLoS One* 12:e0174575. doi: 10.1371/journal.pone.0174575

DeepVariant (2016). *DeepVariant is an Analysis Pipeline that Uses a Deep Neural Network to Call Genetic Variants From Next-Generation DNA Sequencing Data.* Available at: https://github.com/google/deepvariant myfootnote1 [accessed November 17, 2018].

Deng, X., Naccache, S. N., Ng, T., Federman, S., Li, L., Chiu, C. Y., et al. (2015). An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic Acids Res.* 43:e46. doi: 10.1093/nar/gkv002

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

do Valle, I. F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., et al. (2016). Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics* 17(Suppl. 12):341. doi: 10.1186/s12859-016-1190-7

Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., et al. (2006). Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *New Engl. J. Med.* 355, 2408–2417. doi: 10.1056/NEJMoa062867

Edico Genome (2018). *DRAGEN Onsite Solutions*. Available at: http://edicogenome.com/dragen-bioit-platform/ [accessed November 28, 2018].

Eheman, C., Henley, S. J., Ballard-Barbash, R., Jacobs, E. J., Schymura, M. J., Noone, A. M., et al. (2012). Annual report to the nation on the status of cancer, 1975-2008, featuring cancers associated with excess weight and lack of sufficient physical activity. *Cancer* 118, 2338–2366. doi: 10.1002/cncr.27514

Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., et al. (2007). Completing the map of human genetic variation. *Nature* 447, 161–165. doi: 10.1038/447161a

Engelhardt, K. R., Xu, Y., Grainger, A., Germani Batacchi, M. G., Swan, D. J., Willet, J. D., et al. (2017). Identification of Heterozygous Single- and Multi-exon Deletions in IL7R by Whole Exome Sequencing. *J. Clin. Immunol.* 37, 42–50. doi: 10.1007/s10875-016-0343-9

Evans, W. E., and Relling, M. V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, 464–468. doi: 10.1038/nature02626

Faltas, B. M., Prandi, D., Tagawa, S. T., Molina, A. M., Nanus, D. M., Sternberg, C., et al. (2016). Clonal evolution of chemotherapy-resistant urothelial carcinoma. *Nat Genet* 48, 1490–1499. doi: 10.1038/ng.3692

FDA (2018). *Science & Research (Drugs) – Table of Pharmacogenomic Biomarkers in Drug Labeling*. Silver Spring, MD: FDA.

Feero, W. G., Guttmacher, A. E., and Collins, F. S. (2008). The genome gets personal – Almost. *JAMA* 299, 1351–1352. doi: 10.1001/jama.299.11.1351

Feng, B. J. (2017). PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* 38, 243–251. doi: 10.1002/humu.23158

Firican, G. (2017). *The 10 Vs of Big Data*. Available at: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx?m=1 [accessed August 17, 2018].

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 45, D777–D783. doi: 10.1093/nar/gkw1121

Galas, D. J. (2001). Making sense of the sequence. *Science* 291, 1257–1260. doi: 10.1126/science.291.5507.1257

Gambin, T., Akdemir, Z. C., Yuan, B., Gu, S., Chiang, T., Carvalho, C. M. B., et al. (2017). Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res.* 45, 1633–1648. doi: 10.1093/nar/gkw1237

Gameiro, D. N. (2016). *AstraZeneca Partners up With Genomics Elite for new Biobank*. Available at: https://labiotech.eu/medical/astrazeneca-partners-up-with-genomics-elite-for-new-biobank/ [accessed August 6, 2018].

Garrod, A. E. (1996). The incidence of alkaptonuria: a study in chemical individuality. 1902. *Mol. Med.* 2, 274–282. doi: 10.1007/BF03401625

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Global Market Insights (2017). *Digital Genome Market worth over $45 billion by 2024*. Available at: https://www.gminsights.com/pressrelease/digital-genome-market [accessed December, 2 2018].

Golden Helix (2017). *Clinical Interpretation of Variants Based on ACMG Guidelines*. London: Golden Helix Inc.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49

Gorski, M. M., Blighe, K., Lotta, L. A., Pappalardo, E., Garagiola, I., Mancini, I., et al. (2016). Whole-exome sequencing to identify genetic risk variants underlying inhibitor development in severe hemophilia A patients. *Blood* 127, 2924–2933. doi: 10.1182/blood-2015-12-685735

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell.* 62, 729–754. doi: 10.1613/jair.1.11222

Grandori, C., and Kemp, C. J. (2018). Personalized Cancer Models for Target Discovery and Precision Medicine. *Trends Cancer* 4, 634–642. doi: 10.1016/j.trecan.2018.07.005

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 109, 83–90. doi: 10.1016/j.ygeno.2017.01.005

Gupta, S., Chatterjee, S., Mukherjee, A., and Mutsuddi, M. (2017). Whole exome sequencing: uncovering causal genetic variants for ocular diseases. *Exp. Eye Res.* 164, 139–150. doi: 10.1016/j.exer.2017.08.013

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science* 339, 321–324. doi: 10.1126/science.1229566

Haiman, C. A., Chen, G. K., Blot, W. J., Strom, S. S., Berndt, S. I., Kittles, R. A., et al. (2011). Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat. Genet.* 43, 570–573. doi: 10.1038/ng.839

Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A., et al. (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* 39, 638–644. doi: 10.1038/ng2015

Halvaei, S., Daryani, S., Eslami-S, Z., Samadi, T., Jafarbeik-Iravani, N., Bakhshayesh, T. O., et al. (2018). Exosomes in cancer liquid biopsy: a focus on breast cancer. *Mol. Ther. – Nucleic Acids* 10, 131–141. doi: 10.1016/j.omtn.2017.11.014

Han, Z., Xiao, S., Li, W., Ye, K., and Wang, Z. Y. (2018). The identification of growth, immune related genes and marker discovery through transcriptome in the yellow drum (Nibea albiflora). *Genes Genomics* 40, 881–891. doi: 10.1007/s13258-018-0697-x

Harmanci, A., and Gerstein, M. (2016). Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat. Methods* 13, 251–256. doi: 10.1038/nmeth.3746

Harvard Medical School (2015). *Cardiac Risk Report*. Boston, MA: Harvard Medical School.

He, K., Ge, D., and He, M. (2017). Big data analytics for genomic medicine. *Int. J. Mol. Sci.* 18:412. doi: 10.3390/ijms18020412

Herper, M. (2017). *Illumina Promises To Sequence Human Genome For $100 – But Not Quite Yet*. Available at: https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#6924957a386d [accessed November, 28 2018].

Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., et al. (2016). UCSC data integrator and variant annotation integrator. *Bioinformatics* 32, 1430–1432. doi: 10.1093/bioinformatics/btv766

Hixson, J. E., Jun, G., Shimmin, L. C., Wang, Y., Yu, G., Mao, C., et al. (2017). Whole exome sequencing to identify genetic variants associated with raised atherosclerotic lesions in young persons. *Sci. Rep.* 7:4091. doi: 10.1038/s41598-017-04433-x

Hofmann, A. L., Behr, J., Singer, J., Kuipers, J., Beisel, C., Schraml, P., et al. (2017). Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics* 18:8. doi: 10.1186/s12859-016-1417-7

Hoischen, A., Krumm, N., and Eichler, E. E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat. Neurosci.* 17, 764–772. doi: 10.1038/nn.3703

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4:e1000167. doi: 10.1371/journal.pgen.1000167

Honey, K. (2018). *FDA Approves First Targeted Therapeutic Based on Tumor Biomarker, Not Tumor Origin*. Available at: https://blog.aacr.org/fda-approves-first-targeted-therapeutic-based-on-tumor-biomarker-not-tumor-origin/ [ACCESSED November, 30 2018].

Huang, T., Shu, Y., and Cai, Y. D. (2015). Genetic differences among ethnic groups. *BMC Genomics* 16:1093. doi: 10.1186/s12864-015-2328-0.

Hung, C. M., Lin, R. C., Chu, J. H., Yeh, C. F., Yao, C. J., and Li, S. H. (2013). The de novo assembly of mitochondrial genomes of the extinct passenger pigeon (Ectopistes migratorius) with next generation sequencing. *PLoS One* 8:e56301. doi: 10.1371/journal.pone.0056301

Ikegawa, S. (2012). A short history of the genome-wide association study: where we were and where we are going. *Genomics Informatics* 10:220. doi: 10.5808/GI.2012.10.4.220

Illumina (2018). *Scalability for Sequencing Like Never Before*. Available at: https://sapac.illumina.com/systems/sequencing-platforms/novaseq/specifications.html [accessed November, 30 2018].

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Jeste, S. S., and Geschwind, D. H. (2014). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat. Rev. Neurol.* 10, 74–81. doi: 10.1038/nrneurol.2013.278

Johannessen, C. M., and Boehm, J. S. (2017). Progress towards precision functional genomics in cancer. *Curr. Opin. Syst. Biol.* 2, 74–83. doi: 10.1016/j.coisb.2017.02.002

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Juan, L., Teng, M., Zang, T., Hao, Y., Wang, Z., Yan, C., et al. (2014). The personal genome browser: visualizing functions of genetic variants. *Nucleic Acids Res.* 42, W192–W197. doi: 10.1093/nar/gku361

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 45, D840–D845. doi: 10.1093/nar/gkw971

Kaye, J., Kanellopoulou, N., Hawkins, N., Gowans, H., Curren, L., and Melham, K. (2014). Can I access my personal genome? The current legal position in the UK. *Med. Law Rev.* 22, 64–86. doi: 10.1093/medlaw/fwt027

Kchouk, M., Gibrat, J. F., and Elloumi, M. (2017). Generations of sequencing technologies: from first to next generation. *Biol. Med.* 9:395. doi: 10.4172/0974-8369.1000395

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Khoja, L., Butler, M. O., Kang, S. P., Ebbinghaus, S., and Joshua, A. M. (2015). Pembrolizumab. *J. Immunother. Cancer* 3:36. doi: 10.1186/s40425-015-0078-9

Kim, B. Y., Park, J. H., Jo, H. Y., Koo, S. K., and Park, M. H. (2017). Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. *PLoS One* 12:e0182272. doi: 10.1371/journal.pone.0182272

Knight W. (2017). *Google Has Released an AI Tool That Makes Sense of Your Genome*. Available at: https://www.technologyreview.com/s/609647/google-has-released-an-ai-tool-that-makes-sense-of-your-genome/ [accessed November, 17 2018].

Kong, S. W., Lee, I. H., Leshchiner, I., Krier, J., Kraft, P., Rehm, H. L., et al. (2015). Summarizing polygenic risks for complex diseases in a clinical whole-genome report. *Genet. Med.* 17, 536–544. doi: 10.1038/gim.2014.143

Krämer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530. doi: 10.1093/bioinformatics/btt703

Kumar, S., Banks, T. W., and Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *Int. J. Plant Genomics* 2012:831460. doi: 10.1155/2012/831460

LaHaye, S., Corsmeier, D., Basu, M., Bowman, J. L., Fitzgerald-Butt, S., Zender, G., et al. (2016). Utilization of whole exome sequencing to identify causative mutations in familial congenital heart disease. *Circ. Cardiovasc. Genet.* 9, 320–329. doi: 10.1161/CIRCGENETICS.115.001324

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi: 10.1093/nar/gkt1113

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics Chapter* 11:17. doi: 10.1002/0471250953.bi1107s32

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3, 1–8. doi: 10.1016/j.bdq.2015.02.001

Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., et al. (2016). Third-generation sequencing and the future of genomics. *bioRxiv* [Preprint]. doi: 10.1101/048603

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616), 285-291. doi: 10.1038/nature19057

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352

Lim, E., Miyamura, J., and Chen, J. J. (2015). Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among asians, blacks, hispanics, and white. *Hawaii J. Med. Public Health* 74, 302–310.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400.e11–416 e11. doi: 10.1016/j.cell.2018.02.052

Liu, T. (2014). Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol. Biol.* 1150, 81–95. doi: 10.1007/978-1-4939-0512-6-4

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32: 894–899. doi: 10.1002/humu.21517

Liu, X., White, S., Peng, B., Johnson, A. D., Brody, J. A., Li, A. H., et al. (2016a). WGSA: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.* 53, 111–112. doi: 10.1136/jmedgenet-2015-103423

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016b). dbNSFP v3.0: a One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Lobo, I. (2006). Multifactorial inheritance and genetic disease. *Nat. Educ.* 1:5.

Lopes, M. C., Joyce, C., Ritchie, G. R. S., John, S. L., Cunningham, F., Asimit, J., et al. (2012). A combined functional annotation score for non-synonymous variants. *Hum. Hered.* 73, 47–51. doi: 10.1159/000334984

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. doi: 10.1186/2047-217X-1-18

Lytras, M. D., and Papadopoulou, P. (2018). *Applying Big Data Analytics in Bioinformatics and Medicine*. IGI Global. Philadelphia, PA. doi: 10.4018/978-1-5225-2607-0

Macrogen Korea (2017). *Humanizing Genomics*. Available at: https://dna.macrogen.com/eng/index.jsp.

Malcolmson, J., Kleyner, R., Tegay, D., Adams, W., Ward, K., Coppinger, J., et al. (2016). SCN8A mutation in a child presenting with seizures and developmental delays. *Cold Spring Harb. Mol. Case Stud.* 2:a001073. doi: 10.1101/mcs.a001073

Malhotra, A., Levine, S., and Allingham-Hawkins, D. (2014). Whole exome sequencing for cancer – is there evidence of clinical utility? *Adv. Genom. Genet.* 4, 115–115. doi: 10.2147/AGG.S58809

Maman, S., and Witz, I. P. (2018). A history of exploring cancer in context. *Nat. Rev. Cancer* 18, 359–376. doi: 10.1038/s41568-018-0006-7

Manolio, T. A. (2017). Incorporating whole-genome sequencing into primary care: falling barriers and next steps. *Ann. Internal Med.* 167, 204–204. doi: 10.7326/M17-1518

Maxwell, E. K., Gonzaga-Jauregui, C., McCarthy, S. E., O'Dushlaine, C., Staples, J., Lopez, A. E., et al. (2017). KaryoScan: abnormal karyotype detection from whole-exome sequence. *bioRxiv* [Preprint]. doi: 10.1101/204719

Mayo Clinic (2017). *Measuring the Value of Whole Exome Sequencing – Beyond the Numbers*. Available at: http://www.necla.org/resources/Value+of+WES.pdf [accessed December, 3 2018].

McCartan, D. P., and Chatterjee, S. (2018). Hereditary and familial cancer. *Surgery* 36, 134–138. doi: 10.1016/j.mpsur.2017.12.003

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17(1);122. doi: 10.1186/s13059-016-0974-4

Menon, R., Patel, N. V., Mohapatra, A., and Joshi, C. G. (2016). VDAP-GUI: a user-friendly pipeline for variant discovery and annotation of raw next-generation sequencing data. *3 Biotech* 6:68. doi: 10.1007/s13205-016-0382-1

Metcalfe, A., Wilson, S., McCahon, D., Sleightholme, H. V., Gill, P., and Cole, T. (2009). Integrating genetic risk assessment for multi-factorial conditions into primary care. *Prim. Health Care Res. Dev.* 10, 200–209. doi: 10.1017/S1463423609001200

Mueller, J. J., Schlappe, B. A., Kumar, R., Olvera, N., Dao, F., Abu-Rustum, N., et al. (2018). Massively parallel sequencing analysis of mucinous ovarian carcinomas:

genomic profiling and differential diagnoses. *Gynecol. Oncol.* 150, 127–135. doi: 10.1016/j.ygyno.2018.05.008

Nagasaki, H., Mochizuki, T., Kodama, Y., Saruhashi, S., Morizaki, S., Sugawara, H., et al. (2013). DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.* 20, 383–390. doi: 10.1093/dnares/dst017.

National Center for Biotechnology Information (2018). *Variation Viewer*. Available at: https://www.ncbi.nlm.nih.gov/variation/view/ [accessed November, 30 2018].

National Human Genome Research Institute (2016). *The Cost of Sequencing a Human Genome*. Available at: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ [accessed August, 6 2018].

Newswire, P. (2016). *Precision Medicine Market Size to Exceed $87 Billion by 2023: Global Market Insights Inc*. Available at: https://www.prnewswire.com/news-releases/precision-medicine-market-size-to-exceed-87-billion-by-2023-global-market-insights-inc-599454691.html [accessed November, 17 2018].

No Author (2015). Genome in a bottle—a human DNA standard. *Nat. Biotechnol.* 33:675. doi: 10.1038/nbt0715-675a

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Patel, Z. H., Kottyan, L. C., Lazaro, S., Williams, M. S., Ledbetter, D. H., Tromp, H., et al. (2014). The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. *Front. Genet.* 5:16. doi: 10.3389/fgene.2014.00016

Perkel, J. M. (2013). LIFE SCIENCE TECHNOLOGIES: exome sequencing: toward an interpretable genome. *Science* 342, 262–264. doi: 10.1126/science.342.6155.262

Pierson, T. M., Adams, D., Bonn, F., Martinelli, P., Cherukuri, P. F., Teer, J. K., et al. (2011). Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. *PLoS Genet.* 7:e1002325. doi: 10.1371/journal.pgen.1002325

Pray, L. A. (2008). Embryo Screening and the Ethics of Human Genetic Engineerng. *Nat. Educ.* 1:207.

PrecisionFDA (2017). *Hidden Treasures – Warm Up*. Available at: https://precision.fda.gov/challenges/1/view/results [accessed November, 26 2018].

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501–D504. doi: 10.1093/nar/gki025

Qiagen (2018a). *ANNOVAR*. Available at: https://www.qiagenbioinformatics.com/products/annovar/.

QIAGEN (2018b). *Ingenuity Variant Analysis*. Available at: https://www.qiagenbioinformatics.com/products/ingenuity-variant-analysis [accessed December 22, 2018].

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* 59, 5–15. doi: 10.1038/jhg.2013.114

Rezende, L. (2014). *FDA Approves the First PARP Inhibitor for Treatment of Ovarian Cancer in BRCA Mutation Carriers*. Available at: http://www.facingourrisk.org/research-clinical-trials/research-findings/lynparza.php [accessed October 31, 2018].

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62. doi: 10.1038/nrg3655

Romanel, A., Lago, S., Prandi, D., Sboner, A., and Demichelis, F. (2015). ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med. Genomics* 8:9. doi: 10.1186/s12920-015-0084-2

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* 20, 1122–1130. doi: 10.1038/gim.2017.247

Seripa, D., Pilotto, A., Panza, F., Matera, M. G., and Pilotto, A. (2010). Pharmacogenetics of cytochrome P450 (CYP) in the elderly. *Ageing Res. Rev.* 9, 457–474. doi: 10.1016/j.arr.2010.06.001

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed. Res. Int.* 2014:309650. doi: 10.1155/2014/309650.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shigemizu, D., Momozawa, Y., Abe, T., Morizono, T., Boroevich, K. A., Takata, S., et al. (2015). Performance comparison of four commercial human whole-exome capture platforms. *Sci. Rep.* 5:12742. doi: 10.1038/srep12742

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum. Mutat.* 34, 57–65. doi: 10.1002/humu.22225

Shringarpure, S. S., and Bustamante, C. D. (2015). Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* 97, 631–646. doi: 10.1016/j.ajhg.2015.09.010

Smedley, D., Jacobsen, J. O., Jager, M., Kohler, S., Holtgrewe, M., Schubach, M., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* 10, 2004–2015. doi: 10.1038/nprot.2015.124

Spratt, D. E., Chan, T., Waldron, L., Speers, C., Feng, F. Y., Ogunwobi, O. O., et al. (2016). Racial/ethnic disparities in genomic sequencing. *JAMA Oncol.* 2, 1070–1074. doi: 10.1001/jamaoncol.2016.1854

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 6, 2601–2610. doi: 10.1093/nar/6.7.2601

Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H. R., et al. (2004). The Ensembl Web site: mechanics of a genome browser. *Genome Res.* 14, 951–955. doi: 10.1101/gr.1863004

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., et al. (2016a). VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* 17(Suppl. 2):444 doi: 10.1186/s12864-016-2722-2

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016b). The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* 2016, 1.30.31–31.30.33. doi: 10.1002/cpbi.5

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big Data: astronomical or Genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195

Tabet, A., Verloes, A., Pilorge, M., Delaby, E., Delorme, R., Nygren, G., et al. (2015). Complex nature of apparently balanced chromosomal rearrangements in patients with autism spectrum disorder. *Mol. Autism* 6:19. doi: 10.1186/s13229-015-0015-2

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

TGex^{TM} (2018). *Knowledge-Driven NGS Analysis [Online]*. Available at: http://tgex.genecards.org/ (accessed December 22, 2018).

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi: 10.1101/gr.772403

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Topol, E. J. (2014). Individualized medicine from prewomb to tomb. *Cell* 157, 241–253. doi: 10.1016/j.cell.2014.02.012

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., et al. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* 5, 829–834. doi: 10.1038/nmeth.1246

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123

Vassy, J.L., Korf, B.R., and Green, R.C. (2015a). How to know when physicians are ready for genomic medicine. *Sci. Transl. Med.* 7, fs219–fs287. doi: 10.1126/scitranslmed.aaa2401

Vassy, J. L., McLaughlin, H. L., MacRae, C. A., Seidman, C. E., Lautenbach, D., Krier, J. B., et al. (2015b). A one-page summary report of genome sequencing for the healthy adult. *Public Health Genomics* 18, 123–129. doi: 10.1159/000370102

Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature* 461, 199–205. doi: 10.1038/nature08451

Vissers, L. E. L. M., Van Nimwegen, K. J. M., Schieving, J. H., Kamsteeg, E. J., Kleefstra, T., Yntema, H. G., et al. (2017). A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet. Med.* 19, 1055–1063. doi: 10.1038/gim.2017.1

Wang, J., Ling, C., and Gao, J. (2017). CNNdel: calling structural variations on low coverage data based on convolutional neural networks. *Biomed. Res. Int.* 2017:6375059. doi: 10.1155/2017/6375059

Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S., and Girirajan, S. (2017). Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci. Rep.* 7:885. doi: 10.1038/s41598-017-01005-x

Wang, Y., Li, G., Ma, M., He, F., Song, Z., Zhang, W., et al. (2018a). GT-WGS: an efficient and economic tool for large-scale WGS analyses based on the AWS cloud service. *BMC Genomics* 19(Suppl. 1):959. doi: 10.1186/s12864-017-4334-x

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., et al. (2018b). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* 19:151. doi: 10.1186/s13059-018-1519-9

Weigelt, B., Bi, R., Kumar, R., Blecua, P., Mandelker, D. L., Geyer, F. C., et al. (2018). The landscape of somatic genetic alterations in breast cancers from ATM germline mutation carriers. *J. Natl. Cancer Inst.* 110, 1030–1034. doi: 10.1093/jnci/djy028

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., et al. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* 92, 414–417. doi: 10.1038/clpt.2012.96

Wright, C. F., Middleton, A., Barrett, J. C., Firth, H. V., FitzPatrick, D. R., Hurles, M. E., et al. (2017). Returning genome sequences to research participants: policy and practice. *Wellcome Open Res.* 2:15. doi: 10.12688/wellcomeopenres.10942.1

Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* 369, 1502–1511. doi: 10.1056/NEJMoa1306555

Ye, K., Hall, G., and Ning, Z. (2016). Structural variation detection from next generation sequencing. *J. Next Gen. Seq. Appl.* S1:007. doi: 10.4172/2469-9853.S1-007

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* 39, 645–649. doi: 10.1038/ng2022

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International cancer genome consortium data portal–a one-stop shop for cancer genomics data. *Database* 2011:bar026. doi: 10.1093/database/bar026

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and von Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* 8:4781. doi: 10.1038/s41598-018-23226-4

Zhu, P., He, L., Li, Y., Huang, W., Xi, F., Lin, L., et al. (2014). OTG-snpcaller: an optimized pipeline based on TMAP and GATK for SNP calling from ion torrent data. *PLoS One* 9:e97507. doi: 10.1371/journal.pone.0097507

Conflict of Interest Statement: CKO is an employee of AstraZeneca UK Limited with an interest in the deployment of WES for personalized medicine.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.