Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping

Zeeshan Ahmed[1,2], Eduard Gibert Renart[1] and Saman Zeeshan[3]

[1] Institute for Health, Health Care Policy and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[2] Department of Medicine, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[3] Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

ABSTRACT

Over the last few decades, genomics is leading toward audacious future, and has been changing our views about conducting biomedical research, studying diseases, and understanding diversity in our society across the human species. The whole genome and exome sequencing (WGS/WES) are two of the most popular next-generation sequencing (NGS) methodologies that are currently being used to detect genetic variations of clinical significance. Investigating WGS/WES data for the variant discovery and genotyping is based on the nexus of different data analytic applications. Although several bioinformatics applications have been developed, and many of those are freely available and published. Timely finding and interpreting genetic variants are still challenging tasks among diagnostic laboratories and clinicians. In this study, we are interested in understanding, evaluating, and reporting the current state of solutions available to process the NGS data of variable lengths and types for the identification of variants, alleles, and haplotypes. Residing within the scope, we consulted high quality peer reviewed literature published in last 10 years. We were focused on the standalone and networked bioinformatics applications proposed to efficiently process WGS and WES data, and support downstream analysis for gene-variant discovery, annotation, prediction, and interpretation. We have discussed our findings in this manuscript, which include but not are limited to the set of operations, workflow, data handling, involved tools, technologies and algorithms and limitations of the assessed applications.

Subjects Bioinformatics, Computational Biology, Genomics, Medical Genetics, Data Science
Keywords Annotation, Bioinformatics, Genomics, Genotyping, Pipelines, Prediction, Variants, WGS, WES, Tools

INTRODUCTION

Genetics is a field of biology, and it is about studying inherited variation, transmission, machinery, and processes of cellular reproduction. Genetic differences among Deoxyribonucleic acid (DNA) of the healthy and diseased individuals and populations are

Submitted 29 March 2021
Accepted 14 June 2021
Published 26 July 2021

Corresponding author
Zeeshan Ahmed,
zahmed@ifh.rutgers.edu

Academic editor
Vladimir Uversky

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.11724**

known as the variants. DNA structure was first solved in 1953 by *Watson & Crick (1953)*, using crystallographic data generated by Rosalind Franklin and Maurice Wilkins (*Zallen, 2003*). It is a complex molecule, encapsulated within their cells including all information necessary to build and maintain an organism. Human DNA sequence is composed of coding and non-coding regions. Coding regions are the sequences that translate into protein, when non-coding are the regulatory sequences that do not encode into amino acids but can impact the degree, timing, and tissue specificity of gene expression. Different high-throughput technologies have been developed in the last few decades to sequence the DNA of human and several other species (*Heather & Chain, 2016*). So far, there have been three generations considered reporting progress in DNA sequencing.

First generation started with designing conceptual framework for DNA replication and encoding proteins in nucleic acids (*Watson & Crick, 1953*; *Zallen, 2003*), and then focused on sequencing RNA species (*e.g.*, microbial ribosomal or transfer RNA, genomes of single-stranded RNA bacteriophages) (*Heather & Chain, 2016*). Different techniques were proposed, including but not limited to the measurement of nucleotide composition (*Holley et al., 1961*); detection of radiolabelled partial-digestion fragments (*Sanger, Brownlee & Barrell, 1965*); production of the first complete protein-coding gene sequence with 2-D fractionation (*Min Jou et al., 1972*; *Fiers et al., 1976*); DNA polymerase to synthesize from a primer using 'plus and minus' (*Sanger & Coulson, 1975*; *Maxam & Gilbert, 1977*); 'chain-termination' for DNA sequencing (*Sanger, Nicklen & Coulson, 1977*); and automated DNA sequencing with fluorometric based detection (*Smith et al., 1985*; *Ansorge et al., 1986*; *Luckey et al., 1990*; *Hunkapiller et al., 1991*). During second generation, luminescent method was used to measure pyrophosphate synthesis (*Nyrén & Lundin, 1985*); adapter sequences were introduced to bind libraries of DNA molecules with beads (*Tawfik & Griffiths, 1998*); Genome Analyzer (GA) machines were introduced for very short reads sequencing. Most importantly paradigm started shifting with the increased production of better-quality data using sequencing machines including, 454 (later bought by Roche and called as the GS 20), Solexa by Illumina, which then followed by the MiSeq, NextSeq, HiSeq (*Voelkerding, Dames & Durtschi, 2009*). Additionally, many other novel methodologies were also brought into the field *e.g.*, sequencing by oligonucleotide ligation and detection (SOLiD) system from Applied Biosystems (*McKernan et al., 2009*), sequence-by-ligation by Complete Genomic (*Drmanac et al., 2010*), and post-light sequencing by Ion Torrent (*Rothberg et al., 2011*). Third generation has proven to be the most advanced (*Schadt, Turner & Kasarskis, 2010*; *Niedringhaus et al., 2011*; *Gut, 2013*), and equipped with high volume and much improved quality producing sequencing technologies, including single molecule sequencing (SMS) by Stephen Quake's lab (*Mardis, 2008*), single molecule real time (SMRT) by Pacific Biosciences (*Bragg et al., 2013*), and solid-state technology to generate suitable nanopores by Oxford Nanopore Technologies (ONT) (*Huse et al., 2007*). Major differences among these generations and technologies are based on the depth of sequencing, number of produced DNA nucleotides, structure and overall size of the outcome (*Mardis, 2008*).

Clinical molecular laboratories are progressively discovering novel sequence variants for significantly regulated (up or down), expressed (high or low) and phenotype associated single and multiple genes linked to genetic disorders (*Richards et al., 2015*). The most popular DNA sequencing methodologies today are the whole-genome sequencing (WGS) and whole-exome sequencing (WES). The WGS is widely applied to sequence the entirety of the genome, while the WES is mainly used to only sequence the protein coding structures. Along with the production of in-depth and high-quality DNA sequence data (*Zeeshan et al., 2020*; *Ahmed et al., 2019*), one challenge that survived throughout three generations is the efficient processing of raw sequence data to support downstream analysis and clinical interpretation (*Koboldt et al., 2010*). It is still difficult to identify novel variants with high penetrance (*Cavalleri & Delanty, 2012*). Together with academic, commercial laboratories have also developed expertise to support genetic testing and implementing models of sequence variant interpretation (*Pepin et al., 2016*). Various bioinformatics tools have been developed worldwide to perform standalone and networking operations, which include cleansing of raw sequence data, converting raw signals into base calling, identifying regions of interest in genome, alignment, and assembly of contigs and scaffolds, and variant detection. Furthermore, well curated gene, variant (functional and non-functional), and disease annotation databases are available to support secondary and downstream sequence data analysis and interpretation (*Ahmed et al., 2020b*; *Mailman et al., 2007*; *Tang et al., 2015*). The scope of this study is to review and support the process of genetic testing with the classification of susceptibility genes to detect changes of clinical significance. Finding and interpreting such variations is still a challenge among diagnostic laboratories and clinicians (*Ahmed & Ucar, 2017*). We are interested in understanding and analyzing current state of genomics solutions for the identification of variants, alleles, and haplotypes by processing raw sequence data of variable length (*Pabinger et al., 2014*).

SURVEY METHODOLOGY

In this study, we consulted published literature and tested available material of the freely available bioinformatics pipelines. In this study we have assessed the set of operations, workflow, data handling, involved tools, technologies and algorithms, and limitations of different bioinformatic pipelines published in last 10 years and available through PubMed Central (PMC)—NCBI. We were focused on the standalone and networking bioinformatics applications proposed to efficiently process WGS and WES data, and support downstream analysis for gene-variant discovery, annotation, prediction, and interpretation. Based on our evaluation and conclusions drawn from comparative analysis, we have felt serious need of a solution addressing the current limitations to the field. As most of the Next Generation Sequencing (NGS) data processing and analytic application are based on the nexus of different command-line applications known as pipelines. Their execution not only requires good programming skills but deeper understanding of the computer science fundamentals *e.g.*, UNIX commands, scripting, file management etc.

Pipelines for WGS and WES data processing and variant calling

WGS and WES pipelines can be divided into three types: (I) cloud computing pipelines, (II) centralized pipelines, and (III) standalone pipelines. Cloud computing pipelines are mainly deployed in environments with on-demand compute resources managed and provided by externa vendors *e.g.*, Amazon AWS, Google cloud, or Microsoft Azure. Centralized pipelines are developed to be deployed and executed in local computer. However, standalone pipelines are mostly applied in the high-performance computing environments (HPC). In this study, we have selected and evaluated eleven different WGS and WES pipelines of all three different types, which includes DNAp (*Causey et al., 2018*), STORMseq (*Karczewski et al., 2014*), ExScaliburn (*Bao et al., 2015*), Atlas2 (*Evani et al., 2012*), MC-GenomeKey (*Elshazly et al., 2016*), Simplex (*Fischer et al., 2012*), Whole Exome sequencing Pipeline web tool (WEP) (*D'Antonio et al., 2013*), SeqBench (*Dander et al., 2014*), VDAP-GUI (*Menon et al., 2016*), and fastq2vcf (*Gao, Xu & Starmer, 2015*).

STORMseq is a cloud-based pipeline for processing WES and WGS data (*Karczewski et al., 2014*). It is built as an Amazon Machine Image (AMI), which has all the information required to automatically launch an instance in the Amazon Web Services (AWS). It uses a graphical point-and-click interface for setting up the pipeline parameters. Once complete, all that is left is to upload the data to Amazon Compute cloud and deploy the pipeline. Authors have claimed it to be a pipeline that is customizable, user-friendly, and based on click-to-deploy architecture. Authors have developed STORMseq for mainly processing and analyzing short reads/sequences produced by the Illumina technology. Its operations include read cleansing and mapping to reference genome, removing duplicates, variant calling, and annotation. In read mapping step the FASTQ data is mapped to the reference genome, by using the Burrows-Wheeler Aligner (BWA) (*Li & Durbin, 2009*; *Li & Durbin, 2010*). and SNAP tools (*Johnson et al., 2008*). It is then followed by the read cleansing step, in which data is sorted and duplicates are removed to minimize the number of false positives. Genome Analysis Toolkit (GATK) (*McKenna et al., 2010*; *Franke & Crowgey, 2020*; *Heldenbrand et al., 2019*; *Brouard et al., 2019*; *Poplin et al., 2017*; *Auwera et al., 2013*; *DePristo et al., 2011*) is used for the variant calling step, in which Single Nucleotide Polymorphisms (SNPs) and insertions and deletions (INDELs) are called. In the final step variants are annotated using different reference databases and VEP tool (*McLaren et al., 2016*). Authors have validated STORMSeq (*Karczewski et al., 2014*) at two paired-end 100 bp read datasets (genome and exome) with both SNAP and BWA. STORMSeq took 10 h to complete the analysis of the exome using the BWA pipeline and 5 h using SNAP. When it took 176 h with BWA and 82 h with SNAP for genome analysis.

Atlas2 is a downloadable package containing a suite of all the necessary tools for analyzing WES and WGS data produced from Roche 454, Illumina, and SOLiD platforms (*Puri, Tiwary & Shukla, 2019*). It is built using the Software-as-a-Service (SaaS) technology (*Rumale & Chaudhari, 2017*), which allows it to be deployed using two different cloud scenarios: local cloud and commercial cloud. Atlas2 relies on Genboree

Workbench (*Riehle et al., 2012*) to be deployed in the local cloud. Genboree Workbench is a platform for deploying genomic tools as a service, that is currently being hosted at the Baylor College of Medicine. It offers a web-based interface, which allows scientists to interact with the pipeline, making it perfect for those with no bioinformatics background. Atlas2 relies on AWS to be deployed using a commercial cloud. It is built as an AMI image, and all it is required is to download it, upload the data to amazon S3 service and deploy it. Its operations comprise of a suite of variant detection software packages, which consists of Atlas-SNP2 for calling SNPs and Atlas-Indel2 for calling short indels.
To demonstrate the effectiveness, authors tested it by performing two separate analysis: one with the Genboree Workbench platform and other with the AWS. In the Genboree test, authors sequenced the complete genomes of a two 14-year-old fraternal twins, both diagnosed with dihydroxyphenylalanine responsive dystonia (DRD), and proved that Atlas2 was able to successfully call all the relevant variants in both patients. In the AWS test, authors used one Illumina and one SOLiD file obtained from the 1,000 Genomes phase 1 project and demonstrated that it was able to successfully call all the variants. More importantly it only took 8–11 h to process the data, making it possible to perform WGS analysis on the cloud.

Simplex is a cloud-based pipeline for processing WES and WGS data (*Fischer et al., 2012*). It is built as a ready to use VirtualBox and a Cloud image, which allows users to deploy it quickly and easily in any compute infrastructure. Simplex has been developed for processing and analyzing short reads/sequences produced by the Illumina technology. Its operations include five steps: quality report, sequence alignment and refinement, alignment statistics, variant detection, and annotation and summary. In the quality report step, data is converted from Solexa and Illumina format into Sanger FASTQ. Then the data is checked for quality, and the results are reported. Afterwards, the data is then trimmed to a given read length and quality. Filters are later applied to eliminate errors, to reduce the number of variant false positives. To complete the quality report, step a second round of data quality statistic is generated, offering an overview of the quality improvements. In the sequence alignment and refinement step, two rounds of data alignment are performed to minimize the number of mismatching bases across all reads. BWA is used in the first round, and Genome Analysis Toolkit (GATK) is followed. The last operation in the sequence alignment step is to remove unmapped and improperly paired reads. The alignment statistics is the subsequent step, which evaluates the data quality before performing variant calling. Next is the variant calling step, in which INDELs and SNPs are called simultaneously. To carry out the variant calling processes the GATK tool is used. In the variant annotation and summary step, the variants are annotated using GATK and ANNOVAR (*Wang, Li & Hakonarson, 2010*). GATK is used to add annotation information from existing databases, such as RefSeq (*Pruitt, Tatusova & Maglott, 2007*), KEGG (*Kanehisa et al., 2002*), and dbSNP (*Sherry et al., 2001*) information. ANNOVAR is used to infer the functions of unknown variants. The results are then merged, and a summary report is outputted. Simplex has been evaluated by using data from the Kabuki study, (*Ng et al., 2010*) and demonstrates that they can analyze 42 samples simultaneously in a good timeframe. Furthermore, since the pipeline is free and

opensource the pipeline is continuously tested and evaluated by multiple researchers around the world.

Whole Exome sequencing Pipeline web tool (WEP) is a centralized web-based application that can analyze WES data produced by Illumina platforms (*D'Antonio et al., 2013*). It offers a user-friendly web interface, designed for scientists without much knowledge of bioinformatics. It consists of three layers: submission, monitoring, and results. The submission layer validates the input files and deploys the pipeline. The monitoring layer provides information on the status of currently running pipelines and allows users to download the output data. The results layer graphs the outcome of the pipelines. The operations of WEP include six modules: Quality controls, Alignment, Conversion, Variant preprocessing, Variant calling and Postprocessing. In the quality control module, an overall quality check of the input data is performed, by using the FastQC tool (*Leggett et al., 2013*; *Brown, Pirrung & McCue, 2017*). The data is then passed to the NGS QC Toolkit (*Patel & Jain, 2012*) which filters low quality reads and removes primer/adaptor sequences. The alignment module is next, in which the data is mapped to a reference genome using BWA. The conversion module is followed, where data is converted and sorted by chromosomal coordinates by suing SAMtools (*Li et al., 2009*; *Li, 2011*). In the variant preprocessing module, the Picard tool (*Liu et al., 2013*) is used to mark and remove duplicate reads. The GATK tool is executed subsequently by performing local realignment, to improve the accuracy of variant calling and reduce false positives. The last step in this module is to call NGSrich which provides important information about the quality of data before variant calling is performed. In the variant calling module, WEP uses GATK tool, which allows to simultaneously call both SNPs and INDELs. The last module is the postprocessing, in which ANNOVAR, is used for annotating the variants. The output is then parsed and uploaded to a central database, and reports are generated. Since WEP is a free-to-use pipeline it has been validated by multiple researchers that have used WEP over the years.

MC-GenomeKey is a multi-cloud pipeline, which uses resources from different vendors (Google, AWS, Microsoft) to deploy a single pipeline, and minimize the overall cost of performing WES and WGS in the cloud (*Ardagna, 2015*). MC-GenomeKey consists of three features: Variant analysis, Workflow parallelization and cloud support. The variant analysis feature is comprised of four phases: quality check, read alignment, variant calling and variant annotation. In the first phase, data with low quality scores is trimmed out, this is achieved by using the FASTX-Toolkit (hannonlab.cshl.edu/fastx_toolkit/). In the second phase the data is mapped to the reference genome using the BWA tool. The variant calling phase is followed, in which the data is analyzed to determine where variants exist. It is accomplished by using the GATK tool. The variant annotation phase uses ANNOVAR to annotate all the variants found using knowledge from different structural and functional databases. Next feature is the workflow parallelization, which consists of a Python-based scheduling algorithm called Cosmos that constantly monitors the cost of the computing resources of multiple vendors. Based on that, it moves portions of the pipeline to cloud providers that offer cheapener resources, with the goal to minimize the overall cost processing NGS data in the cloud. The last

feature of MC-GenomeKey is the cloud support, which allows scientists to deploy the MC-GenomeKey in two different cloud scenarios, which include individual cloud, and Multi-Cloud. Scientists have the option to deploy the entire pipeline in an individual cloud platform or deploy it across multiple cloud platforms. The authors tested MC-GenomeKey by using two datasets, a WGS and a WES dataset. The first experiment is to test the performance of MC-GenomeKey across different cloud provides, which included Amazon, Google and Microsoft. The results for both WGS and WES samples indicate that MC-GenomeKey can complete the pipeline faster when is deployed using Amazon resources and can be completed for cheaper when using Google resources. The second experiment involves the migration of portions of the pipeline to other cloud providers. Authors proved that using the migration model can lead to cost reduction in many cases with minor running time increases.

VDAP-GUI is a fully automated standalone pipeline, designed for non-IT scientists (*Menon et al., 2016*). It can analyze WES and WGS data produced by Illumina, Roche 454, and Ion torrent platforms. It is comprised of four steps: quality control and trimming, reference mapping and duplicate marking, variant calling, and annotation. In the first step, the quality and the read length of the data is analyzed using the FastQC tool, then it is filtered using the PRINSEQ (*Schmieder & Edwards, 2011*). In the reference mapping and duplicate marking step, the data is aligned to the reference genome using the BWA tool. Later the data is subject to duplicate marking using the Picard tools to remove false positives. The variant calling step is followed, which performs a technique called MultiCom where three variant calling tools are simultaneously executed and only variants that appear in at least two of them are selected. VDAP-GUI relies on SAMtools, VarScan (*Koboldt et al., 2009*) and Freebayes (*Garrison & Marth, 2012*) to perform the variant calling. In the final step variants are filtered and annotated using the VEP tool. Authors validated VDAP-GUI using a publicly available human WES dataset PDA_033-Tumor. It was able to detect a total of 55,919 SNPs, and in addition, it was able to detect 46,963 SNPs that were not reported in the previous study, solidifying the validity and the effectiveness of VDAP-GUI.

Fastq2vcf is a standalone pipeline for processing WES and WGS data (*Gao, Xu & Starmer, 2015*). It integrates multiple sequencing analysis tools to achieve better data processing efficiency. Fastq2vcf is built around three shell script files: "QC_mapping.sh", "PreCalling.sh" and "Variant.sh". The QC_mapping script performs quality control of the raw data, by using the FastQC tool, which provides a summary of the quality of the data processed. The next is to align the data with the reference genome, by using the BWA tool. The PreCalling script subsequently removes duplicates from the data, to improve the quality of the variant calling, by using the MarkDuplicate command-line tool from Picard. Then GATK recalibration tools are used to perform local realignments and base quality recalibration to help correct the misalignments. Finally, the last script performs variant calling, which uses four tools simultaneously: GATK UnifiedGenotyper, GATK HaplotypeCaller, SAMtools and SNVer (*Wei et al., 2011*), and the outputs of all the tools are consolidated into a single call set. Only the variants that appear in all of them are used. Lastly ANNOVAR, and VEP are used to annotate the variants. Authors evaluated

Fastq2vcf by using a five WES sample, and demonstrated the effectiveness of Fastq2vcf by deploying it on commodity hardware and discovered a total of 55,919 SNPs while only taking 27 h to compute the results.

ExScalibur is a set of high-performance cloud-based pipelines for processing WES data (*Bao et al., 2015*). Authors have designed ExScalibur for processing and analyzing short reads/sequences produced by the Illumina technology. Its operations include seven modules: quality control, preprocessing, alignment, alignment refinement, variant calling and filtering, annotation, and project report generation. In the first module the raw sequencing reads are assessed for base quality, duplication level, and nucleotide composition. The preprocessing module is followed, where adapters and removed. Next, is the alignment module, which maps the raw sequence reads into the reference genome using three different aligners: BWA-aln, BWA-mem, and Novoalign. Later, the alignment refinement module improves the alignment by using the base quality score recalibration tool. Subsequently the variant calling module, performs parallel execution of multiple variant callers to increase the variant calling confidence. The variant callers are split into two groups germline and somatic. The germline callers include GATK UnifiedGenotyper, GATK HaplotypeCaller, FreeBayes, SAMtools mpileup/bcftools, Isaac Variant Caller (*Raczy et al., 2013*) and Platypus (*Rimmer et al., 2014*). The Somatic variant callers include MuTect (*Cibulskis et al., 2013*), Shimmer (*Hansen et al., 2013*), SomaticSniper (*Larson et al., 2012*), Strelka (*Saunders et al., 2012*), VarScan2 (*Koboldt et al., 2012*) and Virmid (*Kim et al., 2013*). The variant filtering module is followed where some custom build filters are applied, to remove false positives. Lastly variants are annotated for gene symbol, functional changes, population frequency, and a comprehensive data analysis report is generated, offering statistics on the variant calling processes. Authors validated ExScalibur by using two different types of WES datasets: germline mutations (GMD) and somatic mutations (SMD) datasets. Each of the datasets where generated using two techniques, simulation, and real data. The simulated dataset was created using the Genome in a Bottle Consortium from the genome of NA12878. In the real dataset they used 30 Acute Myeloid Leukemia (AML) tumor/normal pairs (*Cibulskis et al., 2013*). Each dataset was tested using a combination of aligners and callers, including a one where two aligners and two callers were performed at the same time. For the GMD experiments on the simulated data authors reported achieving a sensitivity of 99.03% when using two aligners and two callers. Similarly, for the AML dataset, authors achieved a 98.03% sensitivity when using the two aligners and callers. In the SMD experiments on simulated data authors achieved 90% sensitivity for all different combinations of aligner and callers. Similarly, results were obtained for the AML dataset, suggesting that using of several aligner and caller variants can be detected with a higher confidence.

DNAp is docker container that users download onto their systems and contains all required tools for analyzing WES and WGS data (*Causey et al., 2018*). It requires necessary computer hardware to execute it, including a minimum of 1.5TB for locally saved inputs and outputs in the working directory. It consists of five steps: quality control, alignment, re-alignment, bam quality control, variant calling, and annotation. The first

step validates the quality of the FASTQ files by using the FastQC tool, next is alignment, in which the data is aligned to the reference genome using the BWA tool. The data is then sorted, index, and duplicates are marked and removed using the Picard tools. The output is then merged into a single file by using the MergeSamFiles tool from Picard. The realignment step is followed, in which the data is re-aligned to reduce the number of variant false positives, to do that the GATK toolset is used. The subsequent steps of the pipeline vary depending on the type of the data that is being processed. When working with WES data the bam quality control step uses GATK diagnoseTargets and the Qualimap tool to examine the sequencing alignment data. When processing WGS data in the bam quality control step uses GATK DepthofCoverage tool. Variant calling is followed, and just like the previous step, when consuming WES data, it uses MuTect2 and Strelka. When processing WGS data it uses BreakDancer and Lumpy. Laster, a consensus algorithm is executed to find the variances that are common in both callers, this is done to reduce the number of variant false positives. In the final step the variants are annotated using the Snpeff (*Cingolani et al., 2012*) and Oncotator (*Ramos et al., 2015*) tools. DNAp was evaluated using two different datasets: a human and a mouse. In the human dataset, it was able to identically call 90% of the variances for the human dataset, when compared to the reference results. In the mouse dataset, it was able to identically call 89.8% of the variances present in the reference dataset. Furthermore, an additional test was performed to show that the pipeline could be deployed in different heterogeneous platforms and the same results were obtained.

SeqBench is a centralized web-based application that merges the management and analysis of WES and WGS data into a single application and combines it with an easy-to-use interface to facilitate the data handling (*Dander et al., 2014*). It is organized into three different modules: data acquisition, the dashboard, and the visualization. The data acquisition module validates the input data provided and deploys the pipeline. The dashboard lays out the status and the statistics of deployed pipelines. Lastly, the visualization module, displays the results obtained. Authors of SeqBench highlight the fact that SeqBench is built on top of an already evaluated and validated cloud-based analysis pipeline named Simplex. For those reasons, no further results are reported.

Pros and cons of reported WGS/WES pipelines

Meeting our earlier discussed survey and review methodology, in this manuscript, we have investigated and reported important elements of eleven different state-of-art WGS and WES pipelines, published with in last ten years and made available through PMC/NCBI. We have compared and highlighted common and variable features of these pipelines in Table 1, and those include, interactive and user friendly graphical interface; open-source code and freely available to the community; can be deployed as a standalone application; able to process multi samples at a time (parallel processing); support debugging; able to process both or individual WES and WGS data; check data quality; generate data quality reports; perform alignment, remove duplicates, call and annotate sequence variant; support long reads-based data processing; provide data simulation and visualization; automatic variant data extraction, transfer and loading into database

Table 1 Feature comparison of STORMseq, Atlas2, Simplex, WEP, MC-GenomeKey, VDAP-GUI, fastq2vcf, ExScaliburn, DNAp, and SeqBench pipelines. The table compares following pipeline features: Interactive and user-friendly graphical interface; Open-source code and freely available to the community; Deployment as a standalone application; Able to process multi samples at a time (parallel processing); Support debugging; Able to process WES data; Able to process WGS data; Generate data quality reports; Check Data Quality; Align to reference genome; Remove Duplicates; Call Variants; Annotate Variants; support long reads-based data processing; provide data simulation and visualization; automatic variant data extraction, transfer and loading into database management system; support SQL based data manipulation; integration with annotation databases.

Features/ Pipelines	STORMseq	Atlas2	Simplex	WEP	MC-GenomeKey	VDAP-GUI	fastq2vcf	ExScaliburn	DNAp	SeqBench
Interactive and user-friendly graphical interface	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No
Open-source code and freely available to the community	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Deployment as a standalone application	Cloud	Cloud	Cloud	Centralized	Cloud	Standalone	Standalone	Cloud	Cloud	Centralized
Able to process multi samples at a time (parallel processing)	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Support debugging	No	No	No	No	Yes	No	No	Yes	No	No
Able to process WES data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Able to process WGS data	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No
Generate data quality reports	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Check Data Quality	Not reported	Not reported	Fastx	FastQC, Ngs-qc	Fastx	FastQC	FastQC	Not reported	FastQC	Fastx
Align to reference genome	BWA	Not reported	BWA	BWA	BWA	BWA	BWA	BWA	BWA	BWA
Remove Duplicates	Not reported	Not reported	Not reported	Picard	Picard	Picard	Picard	Picard	Picard	Not reported
Call Variants	GATK	Atlas-SNP2, and Atlas-Indel2	GATK	GATK	GATK	SAMTools, VarScan, FreeBayes	GATK, SAMtools, and SNVer	GATK, FreeBayes, SAMtools,Isaac Variant Caller, and Platypus	MuTect2, Strelka, BreakDancer, Lumpy	GATK
Annotate Variants	VEP	Not reported	GATK, ANNOVAR	ANNOVAR	ANNOVAR	VEP	ANNOVAR and, VEP	ANNOVAR	SnpEff, and Oncotator	GATK, ANNOVAR
Support long reads-based data processing	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Provide data simulation and visualization	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Automatic variant data extraction, transfer and loading into database management system.	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Support SQL based data manipulation	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Integration with annotation databases	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported

management system; support SQL based data manipulation; and integration with annotation databases.

In our conclusions, not all developed and reported pipelines are interactive, based on user friendly graphical interface and can be used by the users without strong computational and programming background. According to our review, STORMseq, Atlas2, Simplex, MC-GenomeKey, ExScaliburn and DNAp are claimed by their authors to be user friendly pipelines but still these cannot be easily deployed and applied for the processing of WES/WGS data. However, all discussed WGS/WES pipelines in this manuscript are open-source code and freely available to the community, and able to process single (sequential) and multi samples (parallel processing) at a time. VDAP-GUI and fastq2vcf are standalone, WEP and SeqBench are recommended for centralized computational environment, when remaining all pipelines (STORMseq, Atlas2, Simplex, MC-GenomeKey, ExScaliburn and DNAp) are cloud based (*e.g.*, AWS supported and tested). With these, current and major limitations of cloud-based pipelines include affordability (high cost) and unsupervised deployment for data processing. Standalone and centralized environment-based pipelines require creating reference indexes and integration of compatible tools used in the published pipeline. Any change in those do not support successful execution of pipelines. One another important aspect is debugging, and except MC-GenomeKey and ExScaliburn, none of the discussed pipeline supports explicit debugging. All mentioned pipelines can process WES data, however, authors of STORMseq, Atlas2, WEP, ExScaliburn, SeqBench have not mentioned using these for WGS data.

Except STORMseq, all reviewed pipelines explicitly support quality checking of WGS/WES data, and most wide used software application proved to be the FastQC. Other than Atlas2, all pipelines are using BWA for mapping sequences to the reference genome. Except STORMseq, Atlas2, Simplex and SeqBench, all have applied Picard tools for removing duplicates. GATK is used by the for calling variants, the STORMseq, Simplex, WEP, MC-GenomeKey, fastq2vcf, ExScaliburn, and SeqBench. When Atlas2 applied Atlas-SNP2 and Atlas-Indel2; VDAP-GUI used SAMTools, VarScan, FreeBayes; and DNAp implmented MuTect2, Strelka, BreakDancer and Lumpy. To annotate variants, STORMseq, fastq2vcf, and VDAP-GUI used VEP; Simplex and SeqBench applied GATK; and WEP, MC-GenomeKey, fastq2vcf, ExScaliburn, and SeqBench have implemented ANNOVAR. When, Atlas2 have not declared and only DNAp has used SnpEff, and Oncotator.

During our review and comparative analysis, we found none of these pipelines have been recommended to process long reads based WGS/WES data (*e.g.*, generated using PacBio and Oxford nanopore) but only for short reads, mainly generated using Illumina sequencing technology. Data simulation and visualization is a key to perform and report efficient downstream analysis, when all these pipelines do provide any kind data visualization supporting variant data interpretation and annotation. Today, when big data is well supported with the used database management systems and implementation of SQL for data manipulation, all discussed pipelines do not provide any kind of automated variant data extraction, transfer and loading process. Furthermore, outcome of these

pipelines cannot be straight forwardly integrated with any of the available annotation databases for timely interpretation and critical analysis. These all limitations do not support application of these pipelines in clinical settings and users from variable backgrounds cannot easily apply these for timely WGS and WES data processing, analysis, simulation, interpretation, and reporting of reproducible results in unorthodox computation and clinical settings.

Precision medicine requires development of progressive healthcare environments that incorporate heterogeneous genomic data into clinical settings. None of these genomics pipelines can efficiently incorporate and leverage genomic (WGS/WES) data processing and analysis in clinical settings, and support decision-making. Furthermore, processed variant data through these pipelines is not available in Artificial Intelligence and Machine Learning (AI/ML) ready formats. So, it can be directly used for integrated and predictive analysis, and deep phenotyping.

DISCUSSION

The first DNA sequencer came out in 1977 by the Sanger (*Sanger, Nicklen & Coulson, 1977*), and later new NGS technologies and methods emerged with time. The chemical structure of the genome is double-stranded DNA, and the smallest unit of genetic information is the base pair (bp), which is two nucleotides paired by hydrogen bonds across the double helix (*Langridge et al., 1957*; *Chargaff, 1979*; *Laird, 1971*). DNA stands (polynucleotides) are composed of four smaller chemical molecules called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T) (*Chaffey, 2003*). The total information content in a haploid set of chromosomal DNA is approximately 3.2 billion bp (*Chial, 2008*). The human genome consists of about 3 × 109 base pairs of DNA. The distance between bases along the DNA strand is ≈ 0.3 nm (1 billionth of a m) (*Marvin et al., 1961*), the length of the fully stretched out human DNA molecule is ≈3 Giga-bp (billion bp), therefore, each cell harbors roughly 1 m of DNA, and remarkably, each cell in our body must compress this one-meter-long DNA into a nuclear volume with a radius of only a few microns. The majority (≈62%) of the human genome comprises of intergenic regions (*Nelson, Hersh & Carroll, 2004*; *Bartonicek et al., 2017*), the non-protein coding (*Clamp et al., 2007*) parts of the genome that lie between genes (*Djebali et al., 2012*). These intergenic sequences used to be called "junk DNA", (*Pennisi, 2012*; *Wright & Bruford, 2011*; *Palazzo & Lee, 2015*) but now genome research over the last few years has revealed functions associated with these regions, suggesting that every part of the genome may have some importance. The bulk of the intergenic DNA is composed of transposable elements, randomly distributed repeated sequences called interspersed repeats (*Smit, 1999*; *Smit, 1996*), and closely spaced repeat units called tandemly repeated repeats (*Quilez et al., 2016*; *Press, Carlson & Queitsch, 2014*) DNA. Intergenic DNA may also include gene regulatory sequences (*Levo & Segal, 2014*; *Sheffield & Furey, 2012*) such as promoters (*Levine, 2010*; *Williamson, Hill & Bickmore, 2011*), enhancers (*Tirosh & Barkai, 2008*; *Aow et al., 2013*), and silencers (*Bushey, Dorman & Corces, 2008*) that have yet to be characterized. The functional elements of the genome are discrete DNA segments that encode a defined product or display a reproducible biochemical signature

(*e.g.*, protein-binding or a specific chromatin structure). Segments of DNA that carry genetic information are called genes (*Maglott et al., 2005*; *Brown et al., 2015*; *Gerstein et al., 2007*; *Durmaz et al., 2015*; *Portin, 2002*; *Cavalli-Sforza, Menozzi & Piazza, 1996*). A gene is a hereditary unit of DNA sequence transferred from parent to offspring that defines a biological function. Alteration to gene is known as mutation/variant, and WGS/WES are most adopted sequencing methodologies to investigate it.

Even with immense molecular and computational advancements, still the goals of completely discovering and understanding human genome functions have not been fulfilled (*Chen et al., 2019*). There are quite a few publicly and commercially available tools that support interpretation of sequence variants (*Zeeshan et al., 2020*; *Ahmed et al., 2019*; *Ahmed et al., 2020b*). Many predictive algorithms (*Ahmed et al., 2020a*) have been proposed and implemented to determine effect of the variant on the primary and alternative transcripts, and impact on protein (*Richards et al., 2015*). These algorithms are mainly divided among three types: Missense (*Thusberg, Olatubosun & Vihinen, 2011*), Splice site (*Jian, Boerwinkle & Liu, 2014*), and Nucleotide conservation (*Savas et al., 2004*). Missense prediction includes but not limited to: ConSurf (*Ashkenazy et al., 2016*), Functional Analysis Through Hidden Markov Models (FATHMM) (*Shihab et al., 2013*), MutationAssessor (*Gnad et al., 2013*), PANTHER (*Thomas et al., 2003*), PhD-SNP (*Capriotti & Fariselli, 2017*), Sorting Intolerant From Tolerant (SIFT) (*Ng & Henikoff, 2003*), SNPs&GO (*Capriotti et al., 2013*), Align GVGD (*Hicks et al., 2011*), MAPP (*Chao et al., 2008*), MutationTaster (*Hombach et al., 2019*), MutPred (*Pienaar, Howell & Elson, 2017*), PolyPhen-2 (*Adzhubei et al., 2010*), PROVEAN (*Choi & Chan, 2015*), nsSNPAnalyzer (*Bao, Zhou & Cui, 2005*), Condel (*González-Pérez & López-Bigas, 2011*), and CADD (*Rentzsch et al., 2019*). Splice site prediction includes, GeneSplicer (*Leman et al., 2018*), Human Splicing (*Spurdle et al., 2008*), Finder (*Tang, Prosser & Love, 2016*), MaxEntScan, NetGene2, NNSplice, FSPLICE. When, Nucleotide conservation predication are made using Genomic Evolutionary Rate Profiling (GERP) (*Shamsani et al., 2019*), PhastCons (*Anna & Monika, 2018*), PhyloP (*Moles-Fernández et al., 2018*). To support variant interpretation and annotation, there are some commercial and freely available reference databases (*e.g.*, Ensembl, GenCode, ClinVar, GeneCards, DISEASES, HGMD, OMIM, GTR, CNVD, GWAS Catalog, COSMIC, dbGaP etc.). These have been designed to support the storage and sharing of sequence data of different types (*e.g.*, genes, somatic and germline mutations etc.), species (*e.g.*, human, mouse, canine, fish, etc.) and size (*Koboldt et al., 2010*; *Cavalleri & Delanty, 2012*; *Pepin et al., 2016*).

Processed high quality WGS/WES data (*e.g.*, generated by Illumina HiSeq) concludes with, if not millions then over hundred thousand variants (*Abecasis et al., 2010*). Downstream analysis of dataset including few samples can be well managed by the small team of bioinformaticians. However, investigating susceptibility of multiple samples (*e.g.*, hundred/thousands) is cumbersome, tedious and time consuming. It is still a challenging task today to perform automatic downstream analysis, which includes gene-variant discovery, annotation, prediction, and genotyping. Furthermore, it is difficult to timely detect *D. Novo* Single-Nucleotide Variants (DNSNVs) (*Liang et al., 2019*) and minimize the number of false negatives (*Hwang et al., 2019*). Implementing platforms

dealing big data analytic challenges require manpower (*e.g.*, bioinformaticians, biostatisticians), computational resources (*e.g.*, HPC and cloud computing environments), and bioinformatics applications (*e.g.*, data inspection, mapping to reference genomes, expression analysis and variant calling). In this manuscript, we have reported our analysis of different genomics applications designed and developed to process and analyze WGS/WES data. We have discussed all applications individually, as well as, performed features based comparative analysis, which includes quality checking (QC), alignment to reference genome, removing duplicates, variant calling, and gene-variant annotation. Most popular bioinformatics applications include FastQC and Fastx are for quality checking, BWA for alignment a reference genome, Picard for removing duplicates, GATK for variant calling, and ANNORVAR and VEP for annotation. We have also assessed the user friendliness, open source and freely availability, deployment environment types, multi sample processing at one time, and debugging and troubleshooting support. We were interested to find if applications can be used by the scientists without much computer science knowledge and bioinformatics skills; allow processing of multiple WGS/WES samples at the same time; report results and log each computational step involved in pipeline.

We need to implement personalized approaches to improve the traditional symptom-driven medical practice with disease-causal genetic variant discovery *Ahmed et al. (2021)*. Major challenges need to be addressed include but not limited to: developing disease-specific cohorts based on patients' genomics profiles; searching for the underlying immunity genes, and common and the rare disease-causal variants, imputations, and haplotype resolutions; finding variant's agnostic of certain predetermined pathologies – like, similarities between disease states not classically considered related at a molecular level; and predicting genetic variants not only affecting targeted but other disorders in particular subjects, which can then be extended to the overall population. Furthermore, it is also important to address ethical issues related to genomics data and those include, forensic identification, and personal, family, and ethnic identities; rights of lost privacy; genetic discrimination; misused human genomics; state-enforced eugenics; and ownership and control of genetic information of participants (*Takashima et al., 2018*; *Roche, 2009*; *Niemiec & Howard, 2016*; *Ahmed & Shabani, 2019*). We need to implement Findable, Accessible, Intelligent, and Reproducible (FAIR) methodologies to efficiently process and analyze high volume - heterogeneous genomics data with multiple data structures in real-time.

CONCLUSIONS

In this manuscript, we have evaluated and reported different genomics pipelines developed to investigate susceptibility of WGS/WES data for variant discovery, annotation, prediction, and genotyping. During our evaluation of discussed pipelines, we found that most of the pipelines are not user friendly and require programming and scripting skills. Furthermore, none of these supports efficient high-volume variant data management and visualization to support timely downstream analysis with less manual and computational power.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AML	Acute Myeloid Leukemia
AMI	Amazon Machine Image
AWS	Amazon Web Services
BWA	Burrows-Wheeler Alignment
DB	Database
DBMS	Database Management System
DNSNVs	*De novo* single-nucleotide variants
DNA	Deoxyribonucleic acid
DRD	Dihydroxyphenylalanine Responsive Dystonia
ERD	Entity Relationship Diagram
ETL	Extract Transform Load
FAIR	Findable, Accessible, Intelligent, and Reproducible
FATHMM	Functional Analysis Through Hidden Markov Models
GEO	Gene Expression Omnibus
GATK	Genome Analysis Toolkit
HPC	High-Performance Computing
INDEL	Insertion or deletion
IGV	Integrative Genomics Viewer
ML	Machine Learning
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PE	Paired End
QC	Quality Check
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SAM	Sequence Alignment Map
SRA	Sequence Read Architecture
SIFT	Sorting Intolerant From Tolerant
SE	Single End
SNP	Single Nucleotide Polymorphism
SMS	Single Molecule Sequencing
SMRT	Single Molecule Real Time
SaaS	Software-as-a-Service
VCF	Variant Call Format file
WES	Whole Exome Sequencing
WEP	Whole Exome Sequencing Pipeline
WGS	Whole Genome Sequencing

ACKNOWLEDGEMENTS

We appreciate great support by the Rutgers Institute for Health, Health Care Policy and Aging Research (IFH); Department of Medicine, Rutgers Robert Wood Johnson Medical School (RWJMS); and Rutgers Biomedical and Health Sciences (RBHS), at Rutgers, The State University of New Jersey.

We thank members and collaborators of Ahmed Lab at the Rutgers IFH and RWJMS for their active participation and contribution to this study.

We acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Institute for Health, Health Care Policy and Aging Research, and Robert Wood Johnson Medical School, at Rutgers, The State University of New Jersey. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Institute for Health, Health Care Policy and Aging Research.
Robert Wood Johnson Medical School, at Rutgers.
State University of New Jersey.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Zeeshan Ahmed conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Eduard Gibert Renart conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Saman Zeeshan conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a review article. Data and source code are not applicable. However, the source code of discussed applications (where applicable) is already made available online by the developers.

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA, 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467(7319)**:1061–1073 DOI 10.1038/nature09534.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7(4)**:248–249 DOI 10.1038/nmeth0410-248.

Ahmed Z, Mohamed K, Zeeshan S, Dong X. 2020a. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database: the Journal of Biological Databases and Curation* **2020(104)**:baaa010 DOI 10.1093/database/baaa010.

Ahmed Z, Renart EG, Zeeshan S, Dong XQ. 2021. Advancing clinical genomics and precision medicine with GVViZ: FAIR bioinformatics platform for variable gene-disease annotation, visualization, and expression analysis. *Human Genomics* **15(37)** DOI 10.1186/s40246-021-00336-1.

Ahmed E, Shabani M. 2019. DNA Data Marketplace: An analysis of the ethical concerns regarding the participation of the individuals. *Frontiers in Genetics* **10**:1107 DOI 10.3389/fgene.2019.01107.

Ahmed Z, Ucar D. 2017. I-ATAC: interactive pipeline for the management and pre-processing of ATAC-seq samples. *PeerJ* **5(Suppl. 14)**:e4040 DOI 10.7717/peerj.4040.

Ahmed Z, Zeeshan S, Mendhe D, Dong X. 2020b. Human gene and disease associations for clinical-genomics and precision medicine research. *Clinical and Translational Medicine* **10(1)**:297–318.

Ahmed Z, Zeeshan S, Xiong R, Liang BT. 2019. Debutant iOS app and gene-disease complexities in clinical genomics and precision medicine. *Clinical and Translational Medicine* **8(1)**:26.

Anna A, Monika G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics* **59(3)**:253–268 DOI 10.1007/s13353-018-0444-7.

Ansorge W, Sproat BS, Stegemann J, Schwager C. 1986. A non-radioactive automated method for DNA sequence determination. *Journal of Biochemical and Biophysical Methods* **13(6)**:315–323 DOI 10.1016/0165-022X(86)90038-2.

Aow JS, Xue X, Run JQ, Lim GF, Goh WS, Clarke ND. 2013. Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Research* **41(9)**:4877–4887 DOI 10.1093/nar/gkt210.

Ardagna D. 2015. Cloud and multi-cloud computing: current challenges and future applications. In: *2015 IEEE/ACM 7th International Workshop on Principles of Engineering Service-Oriented and Cloud Systems*. 1–2.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research* **44(W1)**:W344–W350 DOI 10.1093/nar/gkw408.

Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel GD, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43(1110)**:11.10.1–11.10.33 DOI 10.1002/0471250953.bi1110s43.

Bao R, Hernandez K, Huang L, Kang W, Bartom E, Onel K, Volchenboum S, Andrade J. 2015. ExScalibur: a high-performance cloud-enabled suite for whole exome germline and somatic mutation identification. *PLOS ONE* **10(8)**:e0135800 DOI 10.1371/journal.pone.0135800.

Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Research* **33(Web Server)**:W480–W482 DOI 10.1093/nar/gki372.

Bartonicek N, Clark MB, Quek XC, Torpy JR, Pritchard AL, Maag JLV, Gloss BS, Crawford J, Taft RJ, Hayward NK, Montgomery GW, Mattick JS, Mercer TR, Dinger ME. 2017. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biology* **18(1)**:241 DOI 10.1186/s13059-017-1363-3.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLOS Computational Biology* **9(4)**:e1003031 DOI 10.1371/journal.pcbi.1003031.

Brouard J, Schenkel F, Marete A, Bissonnette N. 2019. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *Journal of Animal Science and Biotechnology* **10(1)**:72 DOI 10.1186/s40104-019-0359-0.

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Scids Research* **43(D1)**:D36–D42 DOI 10.1093/nar/gku1055.

Brown J, Pirrung M, McCue LA. 2017. FQC dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33(19)**:3137–3139.

Bushey AM, Dorman ER, Corces VG. 2008. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Molecular Cell* **32(1)**:1–9 DOI 10.1016/j.molcel.2008.08.017.

Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. 2013. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* **14(Suppl 3)**:S6 DOI 10.1186/1471-2164-14-S3-S6.

Capriotti E, Fariselli P. 2017. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Research* **45(W1)**:W247–W252 DOI 10.1093/nar/gkx369.

Causey JL, Ashby C, Walker K, Wang Z, Yang M, Guan Y, Moore J, Huang X. 2018. DNAp: a pipeline for DNA-seq data analysis. *Scientific Reports* **8**:6793.

Cavalleri GL, Delanty N. 2012. Opportunities and challenges for genome sequencing in the clinic. *Advances in Protein Chemistry and Structural Biology* **89**:65–83.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1996. *The history and geography of human genes*. Princeton, NJ: Princeton University Press.

Chaffey N. 2003. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell—4th edn.. *Annals of Botany* **91(3)**:401 DOI 10.1093/aob/mcg023.

Chao EC, Velasquez JL, Witherspoon MS, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennert G, Anton-Culver H, Lynch H, Lipkin SM. 2008. Accurate classification of *MLH1/MSH2* missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). *Human Mutation* **29(6)**:852–860 DOI 10.1002/humu.20735.

Chargaff E. 1979. How genetics got a chemical education. *Annals of the New York Academy of Sciences* **325**:344–360 DOI 10.1111/j.1749-6632.1979.tb14144.x.

Chen J, Li X, Zhong H, Meng Y, Du H. 2019. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Scientific Reports* **9(1)**:9345 DOI 10.1038/s41598-019-45835-3.

Chial H. 2008. DNA sequencing technologies key to the human genome project. *Nature Education* **1(1)**:219.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31(16)**:2745–2747 DOI 10.1093/bioinformatics/btv195.

Cibulskis K, Lawrence M, Carter S, Sivachenko AY, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander E, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31(3)**:213–219 DOI 10.1038/nbt.2514.

Cingolani P, Platts A, Wang l, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6(2)**:80–92 DOI 10.4161/fly.19695.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104(49)**:19428–19433 DOI 10.1073/pnas.0709013104.

D'Antonio M, Meo PD, Paoletti D, Elmi B, Pallocca M, Sanna N, Picardi E, Pesole G, Castrignanò T. 2013. WEP: a high-performance analysis pipeline for whole-exome data. *BMC Bioinformatics* **14**:S11.

Dander A, Pabinger S, Sperk M, Fischer M, Stocker G, Trajanoski Z. 2014. SeqBench: integrated solution for the management and analysis of exome sequencing data. *BMC Research Notes* **7(1)**:43 DOI 10.1186/1756-0500-7-43.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43(5)**:491–498 DOI 10.1038/ng.806.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. 2012. Landscape of transcription in human cells. *Nature* **489(7414)**:101–108 DOI 10.1038/nature11233.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327(5961)**:78–81.

Durmaz AA, Karaca E, Demkow U, Toruner G, Schoumans J, Cogulu O. 2015. Evolution of genetic techniques: past, present, and beyond. *BioMed Research International* **2015(7434)**:461524 DOI 10.1155/2015/461524.

Elshazly H, Souilmi Y, Tonellato P, Wall D, Abouelhoda M. 2016. MC-GenomeKey: a multicloud system for the detection and annotation of genomic variants. *BMC Bioinformatics* **18**:49.

Evani US, Challis D, Yu J, Jackson AR, Paithankar S, Bainbridge MN, Jakkamsetti A, Pham P, Coarfa C, Milosavljevic A, Yu F. 2012. Atlas2 cloud: a framework for personal genome analysis in the cloud. *BMC Genomics* **13(Suppl. 6)**:S19.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260(5551)**:500–507 DOI 10.1038/260500a0.

Fischer M, Snajder R, Pabinger S, Dander A, Schossig A, Zschocke J, Trajanoski Z, Stocker G. 2012. SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. *PLOS ONE* **7(8)**:e41948 DOI 10.1371/journal.pone.0041948.

Franke KR, Crowgey EL. 2020. Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. *Genomics & Informatics* **18(1)**:e10 DOI 10.5808/GI.2020.18.1.e10.

Gao X, Xu J, Starmer J. 2015. Fastq2vcf: a concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Research Notes* **8(1)**:491 DOI 10.1186/s13104-015-1027-x.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*. *Available at https://arxiv.org/abs/1207.3907*.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17(6)**:669–681 DOI 10.1101/gr.6339607.

Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. 2013. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14(Suppl. 3)**:S7 DOI 10.1186/1471-2164-14-S3-S7.

González-Pérez A, López-Bigas N. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics* **88(4)**:440–449 DOI 10.1016/j.ajhg.2011.03.004.

Gut IG. 2013. New sequencing technologies. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the. *National Cancer Institute of Mexico* **15(11)**:879–881.

Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. 2013. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics* **29(12)**:1498–1503 DOI 10.1093/bioinformatics/btt183.

Heather JM, Chain B. 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics* **107(1)**:1–8 DOI 10.1016/j.ygeno.2015.11.003.

Heldenbrand JR, Baheti S, Bockol MA, Drucker TM, Hart S, Hudson M, Iyer R, Kalmbach M, Kendig KI, Klee E, Mattson NR, Wieben E, Wiepert M, Wildman D, Mainzer LS. 2019. Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinformatics* **20(1)**:31 DOI 10.1186/s12859-019-3169-7.

Hicks S, Wheeler DA, Plon SE, Kimmel M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutation* **32(6)**:661–668 DOI 10.1002/humu.21490.

Holley RW, Apgar J, Merrill SH, Zubkoff PL. 1961. Nucleotide and oligonucleotide compositions of the alanine-, valine-, and tyrosine-acceptor soluble ribonucleic acids of yeast. *Journal of the American Chemical Society* **83(23)**:4861–4862 DOI 10.1021/ja01484a040.

Hombach D, Schuelke M, Knierim E, Ehmke N, Schwarz JM, Fischer-Zirnsak B, Seelow D. 2019. MutationDistiller: user-driven identification of pathogenic DNA variants. *Nucleic Acids Research* **47(W1)**:W114–W120 DOI 10.1093/nar/gkz330.

Hunkapiller T, Kaiser RJ, Koop BF, Hood L. 1991. Large-scale and automated DNA sequence determination. *Science* **254(5028)**:59–67 DOI 10.1126/science.1925562.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology* **8(7)**:R143.

Hwang KB, Lee IH, Li H, Won DG, Hernandez-Ferrer C, Negron JA, Kong SW. 2019. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Scientific Reports* **9(1)**:3219 DOI 10.1038/s41598-019-39108-2.

Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research* **42(22)**:13534–13544 DOI 10.1093/nar/gku1206.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, De Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24(24)**:2938–2939 DOI 10.1093/bioinformatics/btn564.

Kanehisa M, Goto S, Kawashima S, Nakaya A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research* **30(1)**:42–46 DOI 10.1093/nar/30.1.42.

Karczewski KJ, Fernald GH, Martin A, Snyder M, Tatonetti N, Dudley J. 2014. STORMSeq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud. *PLOS ONE* **9(1)**:e84860 DOI 10.1371/journal.pone.0084860.

Kim S, Jeong K, Bhutani K, Lee J, Patel A, Scott E, Nam H, Lee H, Gleeson JG, Bafna V. 2013. Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biology* **14(8)**:R90 DOI 10.1186/gb-2013-14-8-r90.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25(17)**:2283–2285.

Koboldt DC, Ding L, Mardis ER, Wilson RK. 2010. Challenges of sequencing human genomes. *Briefings in Bioinformatics* **11(5)**:484–498.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22(3)**:568–576 DOI 10.1101/gr.129684.111.

Laird CD. 1971. Chromatid structure: relationship between DNA content and nucleotide sequence diversity. *Chromosoma* **32(4)**:378–406 DOI 10.1007/BF00285251.

Langridge R, Seeds WE, Wilson HR, Hooper CW, Wilkins MHF, Hamilton LD. 1957. Molecular structure of deoxyribonucleic acid (DNA). *The Journal of Biophysical and Biochemical Cytology* **3(5)**:767–778 DOI 10.1083/jcb.3.5.767.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28(3)**:311–317 DOI 10.1093/bioinformatics/btr665.

Leggett RM, Ramírez-González RH, Clavijo B, Waite D, Davey R. 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in Genetics* **4**:288 DOI 10.3389/fgene.2013.00288.

Leman R, Gaildrat P, Gérald LG, Ka C, Fichou Y, Audrezet M-P, Caux-Moncoutier V, Caputo SM, Boutry-Kryza N, Mélanie L, Mazoyer S, Fçoise B-D, Sevenet N, Guillaud-Bataille M, Rouleau E, Bressac-de Paillerets B, Wappenschmidt B, Rossing M, Muller D, Bourdon V, Fçoise R, Parsons MT, Rousselin A, Gégoire D, Castelain G, Castéra L, Sokolowska J, Coulet F, Delnatte C, Férec C, Spurdle AB, Martins A, Krieger S,

Houdayer C. 2018. Novel diagnostic tool for prediction of variant spliceogenicity derived from a set of 395 combined in silico/in vitro studies: an international collaborative effort. *Nucleic Acids Research* **46(15)**:7913–7923 DOI 10.1093/nar/gky372.

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Current Biology: CB* **20(17)**:R754–R763 DOI 10.1016/j.cub.2010.06.070.

Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* **15(7)**:453–468 DOI 10.1038/nrg3684.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27(21)**:2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26(5)**:589–595 DOI 10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079.

Liang Y, He L, Zhao Y, Hao Y, Zhou Y, Li M, Li C, Pu X, Wen Z. 2019. Comparative analysis for the performance of variant calling pipelines on detecting the de novo mutations in humans. *Frontiers in Pharmacology* **10**:358 DOI 10.3389/fphar.2019.00358.

Liu X, Han S, Wang Z, Gelernter J, Yang BZ. 2013. Variant callers for next-generation sequencing data: a comparison study. *PLOS ONE* **8(9)**:e75619.

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM. 1990. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research* **18(15)**:4417–4421 DOI 10.1093/nar/18.15.4417.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2005. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research* **33(Suppl. 1)**:D54–D58 DOI 10.1093/nar/gki031.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics* **39(10)**:1181–1186.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* **9(1)**:387–402 DOI 10.1146/annurev.genom.9.081307.164359.

Marvin DA, Spencer M, Wilkins MHF, Hamilton LD. 1961. The molecular configuration of deoxyribonucleic acid III—X-ray diffraction study of the C form of the lithium salt. *Journal of Molecular Biology* **3(5)**:547–565 DOI 10.1016/S0022-2836(61)80021-1.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74(2)**:560–564 DOI 10.1073/pnas.74.2.560.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella KV, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20(9)**:1297–1303 DOI 10.1101/gr.107524.110.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR,

Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19(9)**:1527–1541.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome Biology* **17(1)**:122 DOI 10.1186/s13059-016-0974-4.

Menon R, Patel NV, Mohapatra A, Joshi C. 2016. VDAP-GUI: a user-friendly pipeline for variant discovery and annotation of raw next-generation sequencing data. *3 Biotech* **6(1)**:68.

Min Jou W, Haegeman G, Ysebaert M, Fiers W. 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237(5350)**:82–88 DOI 10.1038/237082a0.

Moles-Fernández A, Duran-Lozano L, Montalban G, Bonache S, López-Perolio I, Menéndez M, Santamariña M, Behar R, Blanco A, Carrasco E, López-Fernández A, Stjepanovic N, Balmaña J, Capellá G, Pineda M, Vega A, Lázaro C, De la Hoya M, Diez O, Gutiérrez-Enríquez S. 2018. Computational tools for splicing defect prediction in breast/ovarian cancer genes: how efficient are they at predicting RNA alterations? *Frontiers in Genetics* **9**:366 DOI 10.3389/fgene.2018.00366.

Nelson CE, Hersh BM, Carroll SB. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology* **5(4)**:R25 DOI 10.1186/gb-2004-5-4-r25.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* **42(9)**:790–793.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31(13)**:3812–3814 DOI 10.1093/nar/gkg509.

Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. 2011. Landscape of next-generation sequencing technologies. *Analytical chemistry* **83(12)**:4327–4341 DOI 10.1021/ac2010857.

Niemiec E, Howard HC. 2016. Ethical issues in consumer genome sequencing: Use of consumers' samples and data. *Applied & translational genomics* **8**:23–30 DOI 10.1016/j.atg.2016.01.005.

Nyrén P, Lundin A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry* **151(2)**:504–509 DOI 10.1016/0003-2697(85)90211-8.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* **15(2)**:256–278 DOI 10.1093/bib/bbs086.

Palazzo AF, Lee ES. 2015. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics* **6(R61)**:2 DOI 10.3389/fgene.2015.00002.

Patel RK, Jain M. 2012. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLOS ONE* **7(2)**:e30619 DOI 10.1371/journal.pone.0030619.

Pennisi E. 2012. Genomics: ENCODE project writes eulogy for junk DNA. *Science* **337(6099)**:1159–1161 DOI 10.1126/science.337.6099.1159.

Pepin MG, Murray ML, Bailey S, Leistritz-Kessler D, Schwarze U, Byers PH. 2016. The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* **18(1)**:20–24.

Pienaar IS, Howell N, Elson JL. 2017. MutPred mutational load analysis shows mildly deleterious mitochondrial DNA variants are not more prevalent in Alzheimer's patients, but may be under-represented in healthy older individuals. *Mitochondrion* **34(9)**:141–146 DOI 10.1016/j.mito.2017.04.002.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell T, Carneiro M, Auwera GA, Kling D, Gauthier L, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly M, Neale B, MacArthur D, Banks E. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. Epub ahead of print 14 November 2017. *bioRxiv* DOI 10.1101/201178.

Portin P. 2002. Historical development of the concept of the gene. *The Journal of Medicine and Philosophy* **27(3)**:257–286 DOI 10.1076/jmep.27.3.257.2980.

Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics: TIG* **30(11)**:504–512 DOI 10.1016/j.tig.2014.07.008.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35(Database Issue)**:D61–D65 DOI 10.1093/nar/gkl842.

Puri GS, Tiwary R, Shukla S. 2019. A review on cloud computing. In: *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. 63–68.

Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ. 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Research* **44(8)**:3750–3762 DOI 10.1093/nar/gkw219.

Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Källberg M, Kumar SA, Liao A, Little KM, Strömberg MP, Tanner SW. 2013. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29(16)**:2041–2043 DOI 10.1093/bioinformatics/btt314.

Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G. 2015. Oncotator: cancer variant annotation tool. *Human Mutation* **36(4)**:E2423–E2429 DOI 10.1002/humu.22771.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47(D1)**:D886–D894 DOI 10.1093/nar/gky1016.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, ACMG Laboratory Quality Assurance Committee. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* **17(5)**:405–424.

Riehle K, Coarfa C, Jackson A, Ma J, Tandon A, Paithankar S, Raghuraman S, Mistretta TA, Saulnier D, Raza S, Diaz MA, Shulman R, Aagaard K, Versalovic J, Milosavljevic A. 2012. The genboree microbiome toolset and the analysis of 16S rRNA microbial sequences. *BMC Bioinformatics* **13(Suppl. 13)**:S11.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg S, Wilkie A, McVean G, Lunter G, WGS500 Consortium. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46(8)**:912–918.

Roche PA. 2009. Ethical challenges encountered in genomic research circulation. *Cardiovascular Genetics* **2(3)**:293–297 DOI 10.1161/CIRCGENETICS.108.846758.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475(7356)**:348–352.

Rumale AS, Chaudhari D. 2017. Cloud computing: software as a service. In: *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 1–6.

Sanger F, Brownlee GG, Barrell BG. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology* **13(2)**:373–398 DOI 10.1016/S0022-2836(65)80104-8.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94(3)**:441–448 DOI 10.1016/0022-2836(75)90213-2.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74(12)**:5463–5467 DOI 10.1073/pnas.74.12.5463.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28(14)**:1811–1817 DOI 10.1093/bioinformatics/bts271.

Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. 2004. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* **13(5)**:801–807.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Human Molecular Genetics* **19(R2)**:R227–R240.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27(6)**:863–864.

Shamsani J, Kazakoff SH, Armean IM, McLaren W, Parsons MT, Thompson BA, O'Mara TA, Hunt SE, Waddell N, Spurdle AB. 2019. A plugin for the ensembl variant effect predictor that uses MaxEntScan to predict variant spliceogenicity. *Bioinformatics* **35(13)**:2315–2317 DOI 10.1093/bioinformatics/bty960.

Sheffield NC, Furey TS. 2012. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. *Genes* **3(4)**:651–670 DOI 10.3390/genes3040651.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29(1)**:308–311.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation* **34(1)**:57–65 DOI 10.1002/humu.22225.

Smit AF. 1996. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development* **6(6)**:743–748 DOI 10.1016/S0959-437X(96)80030-X.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development* **9(6)**:657–663 DOI 10.1016/S0959-437X(99)00031-3.

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic acids research* **13(7)**:2399–2412 DOI 10.1093/nar/13.7.2399.

Spurdle AB, Couch FJ, Hogervorst FB, Radice P, Sinilnikova OM, IARC Unclassified Genetic Variants Working Group. 2008. Prediction and assessment of splicing alterations: implications for clinical testing. *Human Mutation* **29(11)**:1304–1313 DOI 10.1002/humu.20901.

Takashima K, Maru Y, Mori S, Mano H, Noda T, Muto K. 2018. Ethical concerns on sharing genomic data including patients' family members. *BMC Medical Ethics* **19(1)**:61 DOI 10.1186/s12910-018-0310-5.

Tang CY, Hung CL, Zheng H, Lin CY, Jiang H. 2015. Novel computational technologies for next-generation sequencing data analysis and their applications. *International Journal of Genomics* **2015**:254685 DOI 10.1155/2015/254685.

Tang R, Prosser DO, Love DR. 2016. Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. *Advances in Bioinformatics* **2016**:5614058 DOI 10.1155/2016/5614058.

Tawfik DS, Griffiths AD. 1998. Man-made cell-like compartments for molecular evolution. *Nature Biotechnology* **16(7)**:652–656 DOI 10.1038/nbt0798-652.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research* **13(9)**:2129–2141 DOI 10.1101/gr.772403.

Thusberg J, Olatubosun A, Vihinen M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation* **32(4)**:358–368 DOI 10.1002/humu.21445.

Tirosh I, Barkai N. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Research* **18(7)**:1084–1091 DOI 10.1101/gr.076059.108.

Voelkerding KV, Dames SA, Durtschi JD. 2009. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry* **55(4)**:641–658 DOI 10.1373/clinchem.2008.112789.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38(16)**:e164 DOI 10.1093/nar/gkq603.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171(4356)**:737–738 DOI 10.1038/171737a0.

Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. 2011. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research* **39(19)**:e132 DOI 10.1093/nar/gkr599.

Williamson I, Hill RE, Bickmore WA. 2011. Enhancers: from developmental genetics to the genetics of common human disease. *Developmental Cell* **21(1)**:17–19 DOI 10.1016/j.devcel.2011.06.008.

Wright MW, Bruford EA. 2011. Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Human genomics* **5(2)**:90–98 DOI 10.1186/1479-7364-5-2-90.

Zallen DT. 2003. Despite Franklin's work, Wilkins earned his Nobel. *Nature* **425(6953)**:15 DOI 10.1038/425015b.

Zeeshan S, Xiong R, Liang BT, Ahmed Z. 2020. 100 Years of evolving gene-disease complexities and scientific debutants. *Briefings in Bioinformatics* **21(3)**:885–905.