Varstation: a complete and efficient tool to support NGS data analysis

Faria, ACO¹; Caraciolo, MP²; Minillo, RM³; Almeida, TF^{1,3}; Pereira, SM¹; Cervato, MC^{1,3}; Filho, JBO²

¹ Innovation Lab, Incubadora Eretz.bio, Sociedade Beneficente Israelita Brasileira Albert Einstein, São Paulo, SP, Brazil.

² Genomika Diagnostics, Sociedade Beneficente Israelita Brasileira Albert Einstein, São Paulo, SP, Brazil.

³ Clinical laboratory, Ambulatory and Diagnostic Center, Sociedade Beneficente Israelita Brasileira Albert Einstein, São Paulo, SP, Brazil.

Abstract

Summary: Varstation is a cloud-based NGS data processor and analyzer for human genetic variation. This resource provides a customizable, centralized, safe and clinically validated environment aiming to improve and optimize the flow of NGS analyses and reports related with clinical and research genetics.

Availability and implementation: Varstation is freely available at <http://varstation.com>, for academic use.

Contact: contact@varstation.com

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

In recent years, advances in next-generation sequencing (NGS) technologies and applications, such as whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted-sequencing have enabled the generation of large-scale data and identification of millions of genetic variants both for research and clinical diagnostic purposes (Yang *et al.*, 2014; Xue *et al.*, 2015). Identifying, interpreting, classifying, and associating human genetics variants with phenotypic particularities among individuals or diseases are important parts of the process to accomplish personalized medicine (Gaspar *et al.*, 2014; Geoffroy *et al.*, 2015).

Data workflow in NGS includes several bioinformatics steps from the analysis of raw sequencing data transforming the signal from the sequencers to raw sequences that are further aligned and compared with the reference genome (Geoffroy *et al.*, 2015). In general, a typical workflow analysis consists of the following steps: raw data quality control (QC), preprocessing, alignment, post-alignment processing, variant calling, annotation, and prioritization (Bao *et al.*, 2014).

There is a pressing need for the development of optimized workflows and variant analysis toolkits that assist clinical and laboratory geneticists and researchers in the identification, classification and ranking of variants in a timely fashion.

We describe a new powerful tool named Varstation (www.varstation.com, Figure 1) to support the human NGS data analysis workflows. Varstation is a cloud-based solution for computational processing and clinical support of NGS genetic testing that provides a customizable, centralized, safe and clinically validated environment. This resource was developed by a multidisciplinary team composed by bioinformaticians, biologists, geneticists, software engineers, and designers. The goal was to design a resource that could be used for different analysis workflows being both scientifically sound and easy to use.

Figure 1: Layout and main features of Varstation.

2 Software description

2.2 The Varstation

Varstation comprises three main components of bioinformatics pipelines: alignment of the sequenced readings to the reference genome; variant calling and annotation, which entails capturing and summarizing all existing information about the variant across multiple public databases.

The strength of Varstation workflow is based on (i) flexibility: files from any DNA sequencing platform, in any format (.fastq, .bam, and .vcf) are processed to generate results; (ii) end-to-end and automated processing: evaluation of quality parameters, mapping, multiple variant callers, database annotation and automatic variant pre-classification according recommendations and guidelines of the American College of Medical Genetics and Genomics (ACMG), the Association of Molecular Pathology (AMP), the College of American Pathologists (CAP) (Richards *et al.*, 2015) and the Clinical Laboratory Accreditation Program (*Programa de Acreditação de Laboratórios Clínicos*, PALC) (<http://www.sbpc.org.br/programas-da-qualidade/documentos-do-palc/>); (iii) support for clinical interpretation: more than 100 genetic mutations databases are incorporated, including data from germline, somatic and structural variants; (iv) robust filters: filtering engine based on all annotated mutation data; (v) clear and structured results: relevant clinical information to support the medical report. In addition, Varstation provides visualization of data features to share results with other institutions. Information from the external databases and filters provides in Varstation are described in [supplementary table S1](#).

2.3 Workflow optimization: main features

To access Varstation, the user must register using a valid e-mail address. After an e-mail confirmation, the new user will be able to access the platform and the new users onboarding tutorial (raw clinical exome data sample are available). Users can update and personalize the workflow depending on the type of the starting file and adjust parameters each component.

QC of raw data and their preprocessing can be performed using the FastQC (Wingett and Andrews, 2018), BEDTools (Quinlan and Hall, 2010), BamTools (Barnett *et al.*, 2011) and VCFtools (Danecek *et al.*, 2011) toolkits. After raw data QC and preprocessing, the next step

is to map readings to the reference genome and process the variant calling with high efficiency and accuracy. Two of the main mapping and alignment tools are available in the platform: Burrows–Wheeler Aligner (BWA) (Li, 2013) and Torrent Mapping Alignment Program (TMAP, <https://github.com/iontorrent/TS/tree/master/Analysis/TMAP>).

Programs available for germline variant calling include GATK - UnifiedGenotyper and HaplotypeCaller (Van der Auwera *et al.*, 2013), SAMtools (Li *et al.*, 2009), FreeBayes (Garrison and Marth, 2012), Atlas (Challis *et al.*, 2012) and smCounter (Xu *et al.*, 2018). For somatic variant detection, Varstation provides GATK (Van der Auwera *et al.*, 2013), SAMtools (Li *et al.*, 2009), VarScan (Koboldt *et al.*, 2012), Vardict (Lai *et al.*, 2016) and smCounter (Xu *et al.*, 2018).

After variant identification, the annotation attributes such as genomic components, gene symbol, amino acid change, and functional consequences are attached to the variant list. For annotation, Varstation offers one in-house annotation algorithm that was based on ANNOVAR, the most commonly used variant annotation programs (Wang *et al.*, 2010).

The last step is variant filtration and prioritization. Users can filter the variants by gene name, ACMG (Richards *et al.*, 2015) or ClinVar (Landrum *et al.*, 2018) classification, protein coding consequence, variant allele frequency (VAF), depth coverage, frequency in public databases, specifically: 1000g, ESP6500, gnomAD, ABraOM, and ExAC (1000 Genomes Project Consortium *et al.*, 2010; Tennessen *et al.*, 2012; Karczewski *et al.*, 2019; Naslavsky *et al.*, 2017; Lek *et al.*, 2016), presence in dbNSP (Sherry *et al.*, 2001), protein-damage prediction tools (VEST, FATHMM, SIFT, Polyphen, PROVEAN, CADD) (Douville *et al.*, 2016; Carter *et al.*, 2013; Shihab *et al.*, 2014; Shihab, Gough, Cooper, Day, *et al.*, 2013; Shihab, Gough, Cooper, Stenson, *et al.*, 2013; Sim *et al.*, 2012; Adzhubei *et al.*, 2013; Choi and Chan, 2015; Rentzsch *et al.*, 2019), disease associated by OMIM (McKusick, 2007) and UniProt (UniProt Consortium, 2019). Variants are also linked to any associated phenotypes in the Human Phenotype Ontology (HPO) (Köhler *et al.*, 2019). Users can also easily visualize the variant of interest, in their context of the .bam files, using the Integrative Genomics Viewer (Robinson *et al.*, 2017).

One of the key features of Varstation is that users can build their organization's internal database, customizing their own filter's pipelines, link variants to phenotypes, diseases or

articles, and can make their own pathogenicity assessments, potentially allowing for the rapid identification of recurrent mutations, and periodic variant reevaluation or reanalysis, following the ACMG guidelines (Deignan *et al.*, 2019). This feature is extremely relevant mainly for clinical laboratories that need policies and protocols established for variant re-analyzes. Besides that, Varstation also provides multiple ready-to-use standardized pipelines.

Finally, Varstation enables the analysis of each sample by up to 3 experts to guarantee accuracy of results, and then it is possible to generate a customized report with a list of candidate variants that will be reported.

4 Conclusion

The process of analyzing NGS data in Varstation improves the performance, traceability, and safety of analyses, in addition the resource complies with the good international practices for analysis of human genetic variants.

Varstation is a complete, simple, and powerful tool to process and analyze NGS data that was already adopted by leading laboratories in Brazil. The initial project of Varstation was developed in June 2015 and became available for use only for the *Sociedade Beneficiente Hospital Israelita Albert Einstein* hospital and the *Genomika* laboratory. In 2018, Varstation began to become available to other laboratories, hospitals, and research centers. Since then, Varstation has being used by approximately 500 organizations and has processed more than 10,000 samples.

As one of Varstation best features, the software has an online chat service that helps users to solve problems found during the analysis process. In addition, a multi-disciplinary team performs hands-on courses to help the community to further improve the quality of their genomic analysis.

Acknowledgments

We would like to thank all colleagues from the Clinical Laboratory of Hospital Israelita Albert Einstein for the partnership and the enriching discussions. We would like to also thank to all developers from the Innovation Laboratory that did contribute and keep working in Varstation, offering an incredible atmosphere of work.

Conflict of Interest: none declared.

References

- 1000 Genomes Project Consortium *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
- Adzhubei,I. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, **Chapter 7**, Unit7.20.
- Bao,R. *et al.* (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.*, **13**, 67–82.
- Barnett,D.W. *et al.* (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692.
- Carter,H. *et al.* (2013) Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, **14 Suppl 3**, S3.
- Challis,D. *et al.* (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*, **13**, 8.
- Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.
- Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
- Deignan,J.L. *et al.* (2019) Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.*, **21**, 1267–1270.
- Douville,C. *et al.* (2016) Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum. Mutat.*, **37**, 28–35.
- Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*.
- Gaspar,P. *et al.* (2014) Variobox: automatic detection and annotation of human genetic variants. *Hum. Mutat.*, **35**, 202–207.
- Geoffroy,V. *et al.* (2015) VaRank: a simple and powerful tool for ranking genetic variants. *PeerJ*, **3**, e796.
- Karczewski,K.J. *et al.* (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes.

BioRxiv.

- Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
- Köhler,S. *et al.* (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.*, **47**, D1018–D1027.
- Lai,Z. *et al.* (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.*, **44**, e108.
- Landrum,M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
- Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
- Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.
- Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
- Naslavsky,M.S. *et al.* (2017) Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum. Mutat.*, **38**, 751–763.
- Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
- Richards,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.
- Robinson,J.T. *et al.* (2017) Variant Review with the Integrative Genomics Viewer. *Cancer Res.*, **77**, e31–e34.
- Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
- Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., *et al.* (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.
- Shihab,H.A., Gough,J., Cooper,D.N., Day,I.N.M., *et al.* (2013) Predicting the functional

- consequences of cancer-associated amino acid substitutions. *Bioinformatics*, **29**, 1504–1510.
- Shihab,H.A. *et al.* (2014) Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics*, **8**, 11.
- Sim,N.-L. *et al.* (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452-7.
- Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
- UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
- Van der Auwera,G.A. *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, **11**, 11.10.1-11.10.33.
- Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
- Wingett,S.W. and Andrews,S. (2018) FastQ Screen: A tool for multi-genome mapping and quality control. [version 2; peer review: 4 approved]. *F1000Res.*, **7**, 1338.
- Xue,Y. *et al.* (2015) Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet. Med.*, **17**, 444–451.
- Xu,C. *et al.* (2018) smCounter2: an accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. *BioRxiv*.
- Yang,Y. *et al.* (2014) Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, **312**, 1870–1879.

varstation

ROUTINES PATIENTS ANALYSIS LIBRARY DATABASE

R30_QA01_005_20180923_567_562_test2

DETAILS VARIANTS METRICS

SHOW FILTER SEARCH VARIANTS 28 SOFT FILTER 12 GENES 5

ORDER BY Population frequency ARTIFACTS PREVIEW

	OPEN	CLASS	VARIANT	CLINVAR	FUNCTIONAL	RAF	DEPTH	ZYG	FREQ	DISNP	IMPACT	DISEASE	HPD
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	5/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	4/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	3/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	0/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	0/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	4/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	3/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	0/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	0/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	4/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	3/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	4/5
✓	ON	B	BRCA1-NM_007294.exon12:c.4327C>T;p.(Arg1443Kaa) [P] [B] [B]	1.0 0.0 Common	exonic nonsynonymous SNV	0.550	206	het	0.0% [all databases]	rs206076	Deleterious [MIT]	OMIM [Unreviewed]	3/5

AUTOMATIC
CLASSIFICATION
•
ACMG/AMP
guidelines

ANNOTATION & OTHER
RESOURCES

•
IGV access
HGVS annotation
PubMed
ClinVar
OMIM/UniProt
Read Depth
Population
frequency
Splicing predictors
Domain predictors
Impact predictors

...and more!

✓ DELIVERED TO
THE DOCTOR

DATABASE

ACMG LEVELS OF EVIDENCE

- P Pathogenic
- P? Prob. Pathogenic
- ? Uncertain (VUS)
- B? Prob. Benign
- B Benign
- A Artifact

P

VARIANTS TRACKING

- ✓ REVIEW OF PAST SAMPLES
- ✓ BETTER RESULTS FOR
FUTURE SAMPLES