SOFTWARE TOOL ARTICLE

SNPnotes: high-throughput tissue-specific functional annotation of single nucleotide variants [version 1; peer review: 2 approved with reservations]

Shraddha Pai ^{1*}, Michael J. Apostolides ^{2*}, Andrew Jung ³, Matthew A. Moss ^{4,5}

¹The Donnelly Centre, University of Toronto, Toronto, ON, Canada

²The Hospital for Sick Children, Toronto, ON, Canada

³University of Toronto, Toronto, ON, Canada

⁴Cold Spring Harbor Laboratory, Cold Spring Harbor, USA

⁵Zucker School of Medicine, Hempstead, NY, USA

* Equal contributors

V1 First published: 22 Oct 2019, 8:1784
<https://doi.org/10.12688/f1000research.20415.1>

Latest published: 22 Oct 2019, 8:1784
<https://doi.org/10.12688/f1000research.20415.1>

Abstract

A key challenge in the application of whole-genome sequencing (WGS) for clinical diagnostic and research is the high-throughput prioritization of functional variants in the non-coding genome. This challenge is compounded by context-specific genetic modulation of gene expression, and variant-gene mapping depends on the tissues and organ systems affected in a given disease; for instance, a disease affecting the gastrointestinal system would use maps specific to genome regulation in gut-related tissues. While there are large-scale atlases of genome regulation, such as GTEx and NIH Roadmap Epigenomics, the clinical genetics community lacks publicly-available stand-alone software for high-throughput annotation of custom variant data with user-defined tissue-specific epigenetic maps and clinical genetic databases, to prioritize variants for a specific biomedical application. In this work, we provide a simple software pipeline, called SNPnotes, which takes as input variant calls for a patient and prioritizes those using information on clinical relevance from ClinVar, tissue-specific gene regulation from GTEx and disease associations from the NHGRI-EBI GWAS catalogue. This pipeline was developed as part of SVAI Research's "Undiagnosed-1" event for collaborative patient diagnosis. We applied this pipeline to WGS-based variant calls for an individual with a history of gastrointestinal symptoms, using 12 gut-specific eQTL maps and GWAS associations for metabolic diseases, for variant-gene mapping. Out of 6,248,584 SNPs, the pipeline identified 151 high-priority variants, overlapping 129 genes. These top SNPs all have known clinical pathogenicity, modulate gene expression in gut tissues and have genetic associations with metabolic disorders, and serve as starting points for

Open Peer Review

Approval Status ? ?

	1	2
version 1		
22 Oct 2019	view	view
1. Deepti Jain , University of Washington, Seattle, USA		
2. Pavel P. Kuksa , University of Pennsylvania, Philadelphia, USA		
Any reports and responses or comments on the article can be found at the end of the article.		

hypotheses about mechanisms driving clinical symptoms. Simple software changes can be made to customize the pipeline for other tissue-specific applications. Future extensions could integrate maps of tissue-specific regulatory elements, higher-order chromatin loops, and mutations affecting splice variants.

Keywords

bioinformatics, genomics, genetics, GWAS, variant annotation, SNPs, software, clinical genetics, epigenetics, epigenomics

This article is included in the **Research to the People** collection.

Corresponding authors: Shraddha Pai (shraddha.pai@utoronto.ca), Michael J. Apostolides (michael.j.apostolides@gmail.com)

Author roles: **Pai S:** Conceptualization, Investigation, Methodology, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Apostolides MJ:** Conceptualization, Investigation, Methodology, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Jung A:** Software; **Moss MA:** Conceptualization, Formal Analysis, Project Administration, Software, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2019 Pai S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Pai S, Apostolides MJ, Jung A and Moss MA. **SNPnotes: high-throughput tissue-specific functional annotation of single nucleotide variants [version 1; peer review: 2 approved with reservations]** F1000Research 2019, 8:1784 <https://doi.org/10.12688/f1000research.20415.1>

First published: 22 Oct 2019, 8:1784 <https://doi.org/10.12688/f1000research.20415.1>

Introduction

Genome sequencing has become an invaluable tool for clinical diagnostics. Several methodologies exist to look at the human genome, each with own benefits and pitfalls. Whole exome sequencing has been a cost-effective technique to identify structural and nucleotide variants in protein-coding regions of the genome, and has identified variants associated with a number of diseases, including psoriasis¹, Factor V Leiden thrombophilia², and Miller Syndrome³. Genome-wide associations studies, originally based on SNP microarrays, have found that roughly half of genetic associations with disease are located outside gene bodies; this fraction approaches 90% with the inclusion of intronic regions⁴. With dropping costs for DNA sequencing, whole genome sequencing (WGS) promises to extend the ability to identify disease-associated single and structural nucleotide variants in the clinic, to non-coding regions of the genome, including gene and chromatin regulatory sequences. However, the increased size and complexity of the data creates a parallel challenge of annotating non-coding variants, as it requires knowledge of tissue-specific gene regulation (or epigenetics), including regulatory elements such as promoters and enhancers, and features of higher order chromatin organization such as Topological Associated Domains.

Purpose and approach

Several large-scale epigenomics projects have catalogued tissue-specific regulatory elements. This includes genetic modulation of gene expression (GTEx project)⁵, chromatin state (Roadmap Epigenomics)⁶, and enhancer-promoter loops for mapping of distal regulatory elements to genes (FANTOM and individual studies)^{7,8}. A computational workflow that integrated these functional annotation maps to annotate variants from WGS assays would be a valuable resource to prioritize variants with potential functional impact in tissues of interest. Popular high-throughput variant annotation tools, such as BioMart and Variant Effect Predictor^{9,10}, do not provide tissue-specific

annotation. While FUMA¹¹ integrates comprehensive epigenetic annotation, it is a web-based service used to annotate top-ranking variants from GWAS studies, rather than being a standalone tool for variant annotation.

In this work, we describe and provide variant annotation software that starts with output from a WGS assay and prioritizes variants based on epigenomic resources described above, as well as clinical genetic and GWAS catalogs of variant-disease association. This tool will allow users to capture functional and clinical information and to analyze variants simply by providing the commonly-used VCF file format. We demonstrate the software's functionality by prioritizing variants from WGS data for a single patient.

Methods

This work was undertaken as part of SVAI Undiagnosed-1 (<https://sv.ai/undiagnosed-1>), which was a collaborative event with the goal of diagnosing a patient with an unknown genetic condition. As data, participant groups were provided with detailed medical history, genotyping and metabolic data from a 33-year old Caucasian male patient, JCM. The event was hosted in June 2019 by the not-for-profit organization SVAI (<http://sv.ai>), with participants located in the San Francisco Bay Area (USA) as well as in Toronto, Canada.

Pipeline

Figure 1 shows the workflow for the pipeline. Patient genotypes are provided in Variant Call Format as input. Conceptually, the tool compiles prior knowledge about the functional significance of variants from the perspective of tissue-specific regulatory information, known genetic disease associations in the literature, tissue-specific genetic modulation of gene expression, and clinical pathogenicity. The annotation sources are integrated with the genotype calls in the VCF file, and this integration results in a single output table with available annotation for the variant.

Figure 1. Workflow.

The user can then prioritize variants based on the combination of known or predicted functional consequences.

Tissue-specific regulatory regions

The NIH Roadmap Epigenomics project performed comprehensive mapping of noncoding DNA in 111 epigenomes to identify putative regulatory elements in diverse human tissue and cell types (<http://www.roadmapepigenomics.org/>)⁶. The presence of a variant overlapping a tissue-specific promoter or enhancer element signifies that variant alteration could change the regulation of a tissue-specific gene, i.e. that it has regulatory impact on gene expression. Tissue-specific 15-state chromatin state models were downloaded from the Roadmap Epigenomics Portal (downloaded from <https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final.all.mnemonics.bedFiles.tgz>). As the symptoms for John M, hereafter referred to as “the Patient”, included impaired function of the gastrointestinal tract, we limited our annotation to tissues and organs of the digestive system (e.g. esophagus, stomach, intestine). Chromatin states for digestive system tissues were used, including fetal stomach, small and large intestine, sigmoid colon, colonic mucosa, mucosa from stomach, duodenum and rectum, esophagus, rectal mucosa, and stomach mucosa. Regions with open chromatin were included for variant annotation (Table 1; states ≤ 1–7).

Genetic modulation of gene expression in the gut

The GTEx project identified variants that significantly modulate gene expression in each of 44 human tissues⁵. Variants

that modulate transcription in gut tissues (gut eQTLs) were downloaded from the GTEx portal (downloaded from https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz; Table 2).

Variant disease associations

Genome-wide SNP-disease associations were downloaded from the NHGRI-EBI GWAS catalog¹², using the TargetValidation.org API¹³; only those associations mapped to metabolic disorders were included (EFO:0000589). In addition, information on known clinical pathogenicity was downloaded from the ClinVar database¹⁴ (downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz).

Annotation file preparation

The pipeline software was implemented in bash and R 3.4.4. VCFtools v0.1.15¹⁵ was used to filter SNPs from the VCF input file. SNP locations were downloaded from the dbSNP 151 database ((downloaded from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/common_all_20180423.vcf.gz). All coordinates are in GRCh37/hg19 build. SNP coordinates were converted to bed format using awk. dbSNP identifiers were converted to genomic coordinates using dbSNP 151 reference (see above). Bedtools v2.28.0¹⁶ was used to identify overlap of SNP coordinates with individual annotation sources. Finally, R merge was used to join tables by variant location.

Output format

The output file (“final_table.txt”) contains SNP coordinates, positionally-overlapping genes, associated clinical significance from ClinVar, associated GWAS trait and p-value, coordinates and state name of overlapping open chromatin states in gut tissues, and name of tissue and genes for significant eQTL associations. The pipeline then filters this file to report only those SNPs with GWAS hits that achieve genome-wide significance ($p < 5 \times 10^{-8}$); this file is titled “GWASsignificant.txt”, and creates a third file with the list of unique genes that meet this criterion (“GWASsignificant_genes.unique.txt”). The user may filter this data still further to identify SNPs with known clinical pathogenicity, as well as SNPs in functionally annotated non-coding regions.

Results

For our test case, we used the SVAI Undiagnosed-1 Patient whole genome sequencing data provided for the Undisclosed-1 hackathon event, hosted by SVAI/Research to the People. We applied our pipeline to chromosomes 1 to 22 and X, Y chromosomes. Out of 6,248,584 SNPs, we identified 151 high-priority variants, overlapping 129 genes, which demonstrate strong evidence for functional significance (Table 3). Therefore,

Table 1. Samples from the NIH Roadmap Epigenomics project used for gut-specific regulatory region annotation.

Sample Number	Tissue name
E075	Colonic mucosa
E077	Duodenum Mucosa
E079	Esophagus
E084	Fetal large intestine
E085	Fetal small intestine
E092	Fetal stomach
E094	Gastric
E101	Rectal Mucosa Donor 29
E102	Rectal Mucosa Donor 31
E106	Sigmoid colon
E109	Small intestine
E110	Stomach Mucosa

Table 2. Samples from the GTEx dataset used to obtain significant gut expression QTLs (eQTLs).

File names
Colon_Sigmoid.v7.signif_variant_gene_pairs.txt.gz
Colon_Transverse.v7.signif_variant_gene_pairs.txt.gz
Esophagus_Muscularis.v7.signif_variant_gene_pairs.txt.gz
Esophagus_Mucosa.v7.signif_variant_gene_pairs.txt.gz
Small_Intestine_Terminal_Ileum.v7.signif_variant_gene_pairs.txt.gz
Stomach.v7.signif_variant_gene_pairs.txt.gz

Table 3. Examples of variants prioritized by SNPnotes. Information includes mapped genes, disease associations from genetic studies, eQTL associations from GTEx, and overlap with open chromatin regions in gut tissues.

Variant	Gene	GWAS ($p < 5 \times 10^{-6}$)	eQTL	Open chromatin
rs964184	ZPR1	Diabetes mellitus ($p < 10^{-108}$)	Esophagus	Transcribed
rs9268645	HLA-DRA	Type I diabetes mellitus ($p < 10^{-100}$)	Esophagus muscularis and mucosa	Weak transcription
rs4148325	UGT1A8, UGT1A9, UGT1A5, UGT1A4, UGT1A6, UGT1A3	Obesity ($p < 5 \times 10^{-93}$)	Esophagus mucosa	Enhancer
rs964184	ZPR1	Metabolic syndrome (3×10^{-31})	Esophagus muscularis	Transcribed
rs2292239	ERBB3	Type 1 diabetes ($p < 3 \times 10^{-27}$)	Sigmoid colon, Transverse colon, Small intestine terminal ileum, Stomach	Transcribed

this pipeline allows the user to prioritize certain variants for further downstream analyses, or clinical follow up.

Conclusions and next steps

This pipeline will allow easy integration of several epigenomic functional annotation maps that could assist in SNV prioritization for clinical and basic research applications. Our provided software can be customized by someone with basic bioinformatics or scripting expertise to generalize to other tissues profiled in the GTEx and NIH Roadmap Epigenomics project.

One beneficial extension would be the identification of variants predicted to affect gene splicing. Such predictions are available in databases of splicing variants, such as dbSNV^{17,18}, or simple splice site prediction algorithms such as MaxEntScan^{19,20}; the latter uses sequence motifs and maximum entropy calculations for its predictions. This approach is limited by the quality of variant databases, and by models that are limited to predicting only in instances that follow canonical rules for splice site regulation. Another promising avenue to predict aberrant splicing is SpliceAI²¹, a deep learning-based model that predicts splice junctions from an arbitrary pre-mRNA transcript sequence. Another valuable addition would be that of using higher-order chromatin interaction maps, which allow the mapping of distal regulatory elements, such as enhancers, to genes (e.g. 8).

Ethical statement and consent

This article is based on research that occurred at the Undisclosed-1 hackathon event, hosted by SVAI/Research to the People.

The patient provided written informed consent for data release of their medical records, including genetic results, blood work and clinical laboratory reports, to the organisers of the event (SVAI/Research to the People). This consent included data release to participants of Undisclosed-1 during the event, and subsequently for this data to be hosted by SVAI/Research to the People in an online data repository with restricted access (see details in “Data Availability”).

The patient provided written informed consent for the publication of all articles based on the research that was carried out at Undisclosed-1 and any accompanying images.

Since the medical records are the patient’s property, the patient was fully informed of what data release would entail and written informed consent was obtained for the release of the medical records. No ethical approval was sought for the Undisclosed-1 event or publication of articles relating to this event.

Data availability

Synapse repository: SVAI, Undiagnosed-1 (syn:20554923), <https://doi.org/doi:10.7303/syn20554923>²².

License information: Since the data contains detailed medical records, access is restricted in order to protect the identity of the patient. Intermediary data is provided throughout the article. In order to access the data, applicants must be registered users of synapse.org and must provide a proposal detailing what the data will be used for. Applicants will also be required to sign a statement that ensures that the data are not shared with others who have not applied to use the data from the SVAI. Please submit applications for data access to hello@sv.ai.

Software availability

Software for this pipeline is available at: <https://github.com/shraddhapai/SNPNotes>

Archived release at time of publication is located at: <http://doi.org/10.5281/zenodo.3352276>²³.

License: MIT

Data in this project contains:

- data/final_table.txt.informative.txt contains the list of informative genes obtained by running this pipeline on WGS data from the patient at the hackathon.
- data/NHGRI_GWAS contains precompiled GWAS associations by disease category.

References

1. Zuo X, Sun L, Yin X, *et al.*: **Whole-exome SNP array identifies 15 new susceptibility loci for psoriasis.** *Nat Commun.* 2015; **6**: 6793.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Kujovich JL: **Factor V Leiden thrombophilia.** *Genet Med.* 2011; **13**(1): 1–16.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Ng SB, Buckingham KJ, Lee C, *et al.*: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet.* 2010; **42**(1): 30–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Hindorf LA, Sethupathy P, Junkins HA, *et al.*: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A.* 2009; **106**(23): 9362–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group, *et al.*: **Genetic effects on gene expression across human tissues.** *Nature.* 2017; **550**(7675): 204–13.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Roadmap Epigenomics ConsortiumKundaje A, Meuleman W, *et al.*: **Integrative analysis of 111 reference human epigenomes.** *Nature.* 2015; **518**(7539): 317–30.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Andersson R, Gebhard C, Miguel-Escalada I, *et al.*: **An atlas of active enhancers across human cell types and tissues.** *Nature.* 2014; **507**(7493): 455–61.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Schmitt AD, Hu M, Jung I, *et al.*: **A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.** *Cell Rep.* 2016; **17**(8): 2042–59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. McLaren W, Gil L, Hunt SE, *et al.*: **The Ensembl Variant Effect Predictor.** *Genome Biol.* 2016; **17**(1): 122.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Smedley D, Haider S, Durinck S, *et al.*: **The BioMart community portal: an innovative alternative to large, centralized data repositories.** *Nucleic Acids Res.* 2015; **43**(W1): W589–98.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Watanabe K, Taskesen E, van Bochoven A, *et al.*: **Functional mapping and annotation of genetic associations with FUMA.** *Nat Commun.* 2017; **8**(1): 1826.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Buniello A, MacArthur JAL, Cerezo M, *et al.*: **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.** *Nucleic Acids Res.* 2019; **47**(D1): D1005–D12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al.*: **Open Targets Platform: new developments and updates two years on.** *Nucleic Acids Res.* 2019; **47**(D1): D1056–D65.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Landrum MJ, Lee JM, Riley GR, *et al.*: **ClinVar: public archive of relationships among sequence variation and human phenotype.** *Nucleic Acids Res.* 2014; **42**(Database issue): D980–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Liu X, Jian X, Boerwinkle E: **dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions.** *Hum Mutat.* 2011; **32**(8): 894–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Liu X, Wu C, Li C, *et al.*: **dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.** *Hum Mutat.* 2016; **37**(3): 235–41.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Eng L, Coutinho G, Nahas S, *et al.*: **Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths.** *Hum Mutat.* 2004; **23**(1): 67–76.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.** *J Comput Biol.* 2004; **11**(2–3): 377–94.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. **SpliceAI.**
[Reference Source](#)
22. SVAI Research: **“SVAI Undiagnosed-1: WGS”.** 2019.
<http://www.doi.org/10.7303/syn20554923>
23. Pai S, Apostolides MJ, Moss Matthew A, *et al.*: **SNPnotes - initial release (Version v1.0.0).** *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.3352276>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 04 March 2020

<https://doi.org/10.5256/f1000research.22437.r60483>

© 2020 Kuksa P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pavel P. Kuksa

Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA

In this contribution authors propose SNPnotes pipeline for high-throughput tissue-specific functional annotation of single-nucleotide variants.

This work is motivated by a need for a stand-alone software capable of high-throughput annotation of custom WGS variant data.

Several types of data are integrated into the pipeline to enable functional annotation including known genetic associations (GWAS catalog), data on clinical significance of variants (ClinVar), expression QTL (GTEx) and others.

The pipeline is made available as a Github repository.

Major comments:

1. Functionality of the pipeline overlaps with existing tools such as WGS¹ and FAVOR (<http://favor.genohub.org>). The authors make no comparisons with these other tools. The manuscript will benefit greatly from clear comparisons of features, highlighting novel functionalities, benefits, etc.
2. No data on the pipeline performance (running time, scalability) is provided. It is also not clear if pipeline can benefit from multi-core or HPC cluster environments to speed-up annotation.
3. Current implementation of the pipeline requires modification of the source code in order to apply to the custom user data and specify parameters. The pipeline scripts should be modified in order to accept custom data, target output directory, analysis parameters, etc as command-line arguments. e.g., main script could be parametrized `vcfFile=${1:-input.vcf}`, `outdir=${2:-snpnotes_output}` instead of using hard-coded absolute paths to the data and output directories. Same applies to the scripts that set up/preprocess annotation data repository.
4. Pipeline documentation is very limited and should be expanded and provide description of

the command-line interface (see also point 3).

Minor comments:

1. One suggestion is to include a small test data, corresponding reference output along with a test script that executes the pipeline on the test data. This can help prospective users 1) with making sure their set up/settings are correct if they can reproduce results on their systems, and 2) with making progress in using pipeline on their own data.
2. Another suggestion is to include in the documentation a description of software requirements (dependencies, versions, etc) and brief instructions on installing them.

References

1. Liu X, White S, Peng B, Johnson AD, et al.: WGSa: an annotation pipeline for human genome sequencing studies. *J Med Genet*. 2016; **53** (2): 111-2 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Algorithms, bioinformatics, genomics, sequence modeling and analysis, high-throughput sequencing data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 November 2019

<https://doi.org/10.5256/f1000research.22437.r56414>

© 2019 Jain D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Deepti Jain

Department of Biostatistics, University of Washington, Seattle, WA, USA

The authors report the workflow and pipeline code they developed to annotate variants. The authors were motivated to develop this pipeline specifically so that a user could use tissue- and disease-specific resources of annotation and annotate a large set of variants, such as from a whole genome sequencing (WGS) analysis, provided in a VCF format. As a use case, their provided code is suppose to download annotations linked with gut related tissues and metabolic disorders, format them and use them to annotate and filter variants provided in a VCF format. The authors applied the provided code to annotate 6,248,584 variants from a patient with history of gastrointestinal symptoms and generated a filtered set of 151 variants using the annotations.

The manuscript reports the pipeline-code for a workflow which is very important and useful for generating a filtered set of variants that likely have biological function and which could be followed-up in more detail after a WGS study. However I have major concerns about the practical feasibility and interests of others using this code because it is tailored for a very specific use case and lacks documentation. For these reasons I hesitate to endorse its acceptance at the present stage.

Major concerns:

- The features for which the authors developed the pipeline i.e ability to use tissue- and disease-specific annotations, and to provide VCF as an input is already available in the Whole Genome Sequence Annotator (WGSA, <https://sites.google.com/site/jpopgen/wgsa>)¹. In addition, WGSA has a large selection of a recently updated annotation resources that a user can choose to annotate the single nucleotide variants as well as indels.
- The code provided is very specific to a use case. A user wanting another set of annotations or more than one set of annotations might end up having multiple versions of the code. In general, such an approach is not considered a good practice as it leads to duplication of lot of code. This can be avoided if the code is updated to handle user specifications provided through a config file.
- A user's ability to use tissue specific annotations in a format different than the currently used resources (example annotations from long range chromatin interactions experiments) will be limited.
- I did not come across any documentation associated with the code. A documentation and vignette will be very helpful so that user can use the code as is, as well as modify it if they desire to use other resources.
- It would be helpful to add a description about how the code handles and reports multiple annotations for a given variant from a given resource. For example, if a variant was found as an eQTL for two different tissues is that variant reported twice or the information from

the two tissues combined and reported in a specific format

Minor suggestion:

- Annotation of a large set of variants can have a huge computational burden. A user would find it useful if the authors could provide software performance benchmarks.

References

1. Liu X, White S, Peng B, Johnson AD, et al.: WGSa: an annotation pipeline for human genome sequencing studies. *J Med Genet*. 2016; **53** (2): 111-2 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular biology and Human genetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research