SOFTWARE TOOL ARTICLE

REVISED CompoundHetVIP: Compound Heterozygous Variant Identification Pipeline [version 2; peer review: 2 approved]

Dustin B. Miller, Stephen R. Piccolo

Department of Biology, Brigham Young University, Provo, UT, 84602, USA

v2 **First published:** 08 Oct 2020, **9**:1211
https://doi.org/10.12688/f1000research.26848.1
Latest published: 10 Feb 2021, **9**:1211
https://doi.org/10.12688/f1000research.26848.2

Abstract

Compound Heterozygous (*CH*) variant identification requires distinguishing maternally from paternally derived nucleotides, a process that requires numerous computational tools. Using such tools often introduces unforeseen challenges such as installation procedures that are operating-system specific, software dependencies that must be installed, and formatting requirements for input files. To overcome these challenges, we developed Compound Heterozygous Variant Identification Pipeline (CompoundHetVIP), which uses a single Docker image to encapsulate commonly used software tools for file aggregation (*BCFtools* or *GATK4*), VCF liftover (*Picard Tools*), joint-genotyping (*GATK4*), file conversion (*Plink2*), phasing (*SHAPEIT2*, *Beagle*, and/or *Eagle2*), variant normalization (*vt* tools), annotation (*SnpEff*), relational database generation (*GEMINI*), and identification of *CH*, homozygous alternate, and *de novo* variants in a series of 13 steps. To begin using our tool, researchers need only install the Docker engine and download the CompoundHetVIP Docker image. The tools provided in CompoundHetVIP, subject to the limitations of the underlying software, can be applied to whole-genome, whole-exome, or targeted exome sequencing data of individual samples or trios (a child and both parents), using VCF or gVCF files as initial input. Each step of the pipeline produces an analysis-ready output file that can be further evaluated. To illustrate its use, we applied CompoundHetVIP to data from a publicly available Ashkenazim trio and identified two genes with a candidate *CH* variant and two genes with a candidate homozygous alternate variant after filtering based on user-set thresholds for global minor allele frequency, Combined Annotation Dependent Depletion, and Gene Damage Index. While this example uses genomic data from a healthy child, we anticipate that most researchers will use CompoundHetVIP to uncover missing heritability in human diseases and other phenotypes. CompoundHetVIP is open-source software and can be found at https://github.com/dmiller903/CompoundHetVIP; this repository also provides detailed, step-by-step examples.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 10 Feb 2021	✓ report	
version 1 08 Oct 2020	? report	✓ report

1. **Soohyun Lee**, Harvard Medical School, Boston, USA

2. **Sushant Patil**, University of North Carolina, Chapel Hill, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Genetics, compound heterozygous, genome analysis, trio, phasing, reproducibility

Corresponding author: Stephen R. Piccolo (stephen_piccolo@byu.edu)

Author roles: Miller DB: Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Piccolo SR**: Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

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

How to cite this article: Miller DB and Piccolo SR. **CompoundHetVIP: Compound Heterozygous Variant Identification Pipeline [version 2; peer review: 2 approved]** F1000Research 2021, **9**:1211 https://doi.org/10.12688/f1000research.26848.2

First published: 08 Oct 2020, **9**:1211 https://doi.org/10.12688/f1000research.26848.1

REVISED **Amendments from Version 1**

We updated the abstract to get to the point more quickly, to mention the tools used for each major step, and to clarify that MAF and CADD are used as part of variant filtering. We updated the manuscript to mention that whole-exome and targeted exome data can also be used. We changed the way we filter the variants to be more simple but also more strict. Now, CADD and MAF scores must be available for all variants, variants must be in exonic regions, and variants can have either "MED" or "HIGH" putative impact. This update to filtering did not alter the results in our example analyses. There was an error in "identify_homAlt_variants.py" that required a gene to have 2 variants in order to be written to the output file. This has been fixed, and it slightly changed the example results. Originally we had reported only 1 gene with a homozygous alternate variant in the example dataset. However, after updating the script and running it again, we identified 2 genes that have a homozygous alternate variant. The manuscript has been updated accordingly. In response to reviewer comments, we have updated the "gemini_load.py" script with an optional argument to allow users to import a file that was annotated with VEP. Since SnpEff is used as part of the pipeline, the script defaults to snpEff, but if a user chooses to use VEP outside of the pipeline, then they can still use the remaining steps of the pipeline. We have updated Figure 1 to show step numbers. We have updated the manuscript to say "global minor allele frequency" when first referenced. Clarifications about phasing were made to both the main text and example PDF. Other minor updates were made to the example PDF and manuscript for added clarity and explanation.

Any further responses from the reviewers can be found at the end of the article

Introduction

A compound heterozygous (*CH*) variant occurs when a person inherits two alleles, one from each parent, and these alleles are located at different positions within the same gene[1]. The compound effects of these alternate alleles may lead to phenotypic effects as seen in some cases of human disease, including ataxia telangiectasia, NGLY1 deficiency, and various types of pediatric cancer[2–4]. For example, *CH* variants in the mismatch repair gene, *MSH6*, have been identified in pediatric patients with colorectal cancer, medulloblastoma, high-grade glioma, glioblastoma, non-Hodgkin's lymphoma, and acute lymphoblastic leukemia[4]. To detect *CH* variants in next-generation sequencing data, it is necessary to differentiate between paternally and maternally derived nucleotides[1]. Laboratory-based methods such as fosmid-pool-based or linked-read sequencing can be used; however, if DNA libraries are prepared and sequenced without regard to nucleotide inheritance (as is done in most sequencing projects), computational methods can help determine parental inheritance through haplotype estimation ("phasing")[5–7].

Available phasing tools require specific input file types (such as VCF or *Plink* files) and reference files which are not standardized across different phasing software. In addition, many phasing programs require that input files have been aligned to a specific reference genome, do not contain multiallelic positions, are free of repeat positions, and that each chromosome be phased separately[8–10]. Figuring out how to prepare files for phasing can be challenging as passing files from program to program may result in unforeseen incompatibilities. Additionally, installing some programs can be challenging because of operating-system specific installation processes and software dependencies.

We have designed Compound Heterozygous Variant Identification Pipeline (CompoundHetVIP) to help researchers overcome these time-consuming challenges when identifying *CH* variants. CompoundHetVIP encapsulates specific versions of existing tools, required software dependencies, and custom Python scripts into a cohesive computational environment packaged as a Docker image[11]. Accordingly, researchers need only install the Docker software and download the CompoundHetVIP Docker image to begin performing *CH*, homozygous alternate, and *de novo* analyses at the command line. Furthermore, because the source code for CompoundHetVIP is publicly available, other researchers will be able to reproduce the analyses and investigate the specific methodologies used.

Methods

Implementation

The functionality of CompoundHetVIP is divided into a series of 13 steps (Figure 1). For each step, a Python script is executed within a Docker container. These scripts provide logic for processing data files and invoking third-party tools. By breaking the pipeline into 13 steps, users have flexibility to perform the steps that are most relevant to their analysis. For example, researchers can use input data for an affected individual only or for a trio (an affected individual and both parents). If parental data are unavailable and the variant positions within the VCF file correspond to genome build GRCh37, users may skip the first three steps. A detailed, step-by-step guide is available on GitHub and as *Extended data*[12].

Workflow

For **step 1**, the inputs can be either Variant Call Format (VCF)[13] or gVCF[14] files that were generated from whole-genome, whole-exome, or targeted exome sequencing data. VCF files contain variant sites only, whereas gVCF files include non-variant sites, too. For each parent of a trio being evaluated, our script retains nucleotide positions that are in common with the child. When gVCF files are used (whether for trios or individuals), our script removes all non-variant sites for the child (but retains these for the parents to support determination of *CH* status). When applied to trio data, some phasing tools, such as *SHAPEIT2*[8], require a single input file for each trio. Therefore, in **step 2** (used only when working with trios), we combine the variant files for each member of a trio into a single VCF file using either *BCFtools* (VCF input files)[15] or *GATK4* (gVCF input files)[14]. If *GATK4* is used, joint-genotyping is also performed on the trio VCF.

The remaining steps can be applied either to trios or individuals. Some phasing and annotation programs require that data be aligned to genome build GRCh37; thus, we use this reference genome as our standard. For variant files that have been aligned to genome build GRCh38, **step 3** uses *Picard Tools*[16] to convert the data to GRCh37 positions using a lift-over process. During lift-over, some sites may be present in GRCh38, but their exact

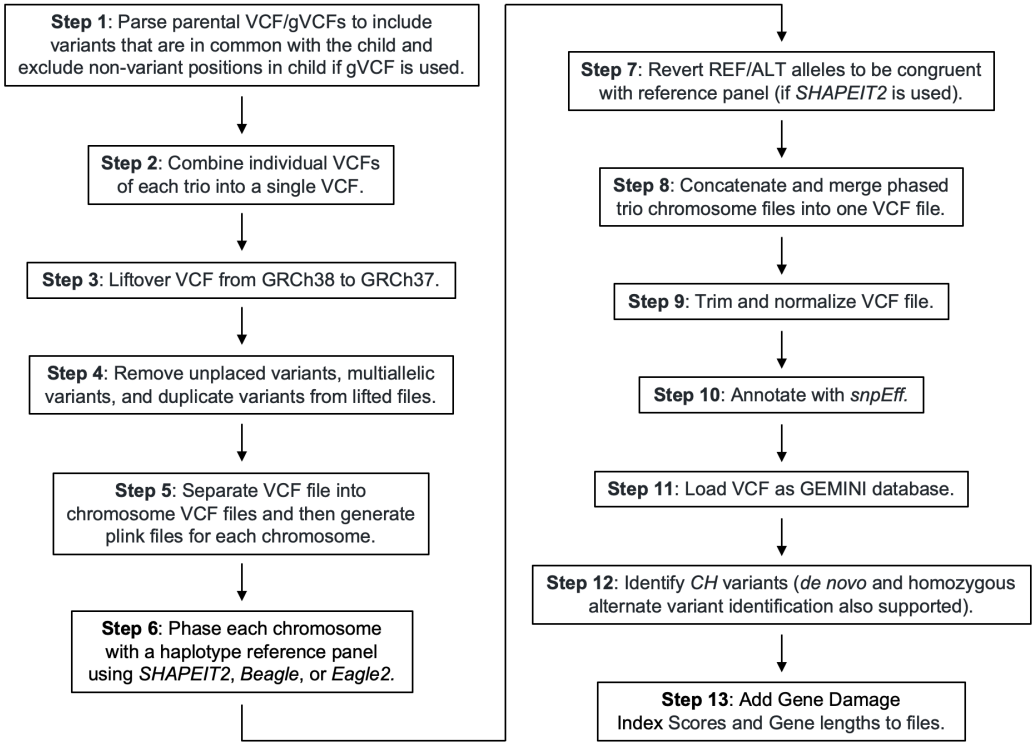

Figure 1. Flow diagram of CompoundHetVIP functionality.

position in GRCh37 is unknown. To avoid ambiguity, these sites are removed during **step 4**. This step also removes positions that are multiallelic or duplicated to maintain compatibility with programs such as *Plink2*[17,18] and *SHAPEIT2* (used in steps 5 and 6, respectively). For trio VCF files, sites that contain missing genotype information (i.e. "./.") for both parents are removed to improve phasing accuracy.

CompoundHetVIP can perform phasing using *SHAPEIT2*, *Eagle2*[10], and/or *Beagle*[9]. Each of these programs requires that each chromosome be phased independently. Additionally, when using *SHAPEIT2,* it is recommended that PLINK files (.bed, .bim, .fam) be used as input for phasing. Therefore, **step 5** divides a VCF file by chromosome into multiple files and creates the necessary PLINK files for each chromosome (when SHAPEIT2 is used for phasing).

Step 6 phases the variants in each chromosome using default parameters for the phasing program chosen by the user. We recommend using *SHAPEIT2* because it can applied either to trios or individuals. When parents' genotypes are available, this program uses Mendelian logic for phasing and a population-based haplotyped reference panel when the phase of the child cannot be determined from Mendelian logic alone (i.e. both parents and child are heterozygous). In addition, if a parent is missing genotype information at a position, *SHAPEIT2* imputes the missing information. All supported phasing programs integrate the 1000 Genomes Project phase 3 haplotype reference panel[19] and do not require sequence alignment files (.bam), such as those required by read-based programs[20,21]. In some scenarios, *SHAPEIT2* switches the REF and ALT alleles. Therefore, **step 7** ensures that the REF/ALT alleles of the phased VCF files are congruent with those of the reference genome. Also, sites with Mendelian errors are removed.

To make subsequent analysis of the phased files easier, **step 8** concatenates all phased chromosomes into a single file. If a user is analyzing multiple trios (or individuals), this script can also merge the data for these trios (or individuals) into a single VCF file.

Step 9 normalizes VCF files as recommended by *GEMINI*[22] (used in step 11). Normalization involves left-alignment and trimming of variants[23]. This process helps ensure that variants are represented at their left-most position, with as few nucleotides as possible, and unambiguously. This step uses *vt tools*[23]. In **step 10,** *SnpEf*[24] provides information about the effects of variants on function for known genes. Then, in **step 11**, *GEMINI*[22] loads the annotated VCF into a relational database (*GEMINI* can also load files annotated with Variant Effect Predictor (*VEP*)[25], although *VEP* is not available as part of our pipeline). **Step 12** uses a custom Python script to extract *CH* variant data from the database. Our provided script identifies *CH* variants and filters the data based on user-set thresholds for global minor allele frequency (MAF) and Combined Annotation Dependent Depletion (CADD) scores[26]. Variants with a MAF less than or equal to the user-set threshold, CADD score greater than or equal to the user-set

threshold, exonic classification, and "HIGH" or "MED" putative impact severity are included in the final output. We consider the variants in the final output as candidates for further evaluation. For step 12, we provide two additional scripts that identify homozygous alternate variants and *de novo* variants using the same user-set thresholds as those described above.

Finally, in **step 13**, we add Gene Damage Index (GDI) scores[27] and gene-length information to the output files. GDI scores quantify accumulated mutational damage in healthy populations as a way to predict whether genes are likely to have disease-causing variants. Genes of longer length (e.g. *TTN*, *MUC5B*) tend to have more total damage but typically less disease-causing damage than shorter genes.

Operation

Because CompoundHetVIP executes all scripts within a Docker container, it can be executed on all major operating systems that are commonly used for scientific computing. Depending on input file sizes, the hardware needed to execute CompoundHetVIP will vary from user to user. Certain tasks, such as phasing (step 6), can be memory intensive. A minimum of 40 GB memory is recommended. When creating a relational database with *GEMINI* (step 11), there is no minimum processing core recommendation, but multiprocessing can significantly speed up the time it takes to load the database. Users can specify how many processing cores *GEMINI* can use when executing step 11.

Results

We applied CompoundHetVIP to high-confidence, VCF data that were generated with whole-genome sequencing data from an Ashkenazim trio available through the Genome in a Bottle Consortium[28]. During step 6, we used *SHAPEIT2* to phase the data. In the child of this trio, we identified a *CH* variant in two genes (*FLNB* and *TTN*) using a MAF threshold of 0.01 and a CADD score threshold of 15. Genes with a GDI score less than or equal to 13.84 are classified as being more likely to have disease-causing damage from variants[27]. *FLNB* (6.2) was lower than this threshold but *TTN* (42.9) was not. *FLNB* has an important role in cytoskeleton development and variations in this gene have been associated with many skeletal disorders[29,30].

In addition, we identified two homozygous alternate variants: one in *TBC1D2* and the other in *TOX2*, using the same MAF and CADD thresholds that we used for *CH* variant identification. *TBC1D2* and *TOX2* had GDI scores of 9.7 and 4.4, respectively. *TBC1D2* codes for a GTPase-activating protein and is involved in E-cadherin degradation[31]. The role of this gene and how it may relate to human disease is not yet fully understood. *TOX2* is a transcription factor that helps drive the development of T follicular helper (Tfh) cells[32]. Tfh cells are an important part of humoral immunity.

Using the same MAF and CADD thresholds described above, we sought to identify *de novo* variants in this trio. However, none passed these thresholds.

Conclusion

CompoundHetVIP provides the necessary tools for *CH* variant identification using VCF or gVCF files as initial input and is executed within a Docker container, which allows for cross-platform compatibility and reproducibility. CompoundHetVIP involves 13 steps (Figure 1) that include a breadth of tasks such as file aggregation, VCF liftover, joint-genotyping, file conversion, phasing, variant normalization, annotating, and variant identification. Our results highlight that potentially damaging *CH* and homozygous alternate variants are observed in seemingly healthy individuals. However, we anticipate that most researchers will use CompoundHetVIP to identify variants in individuals with a known disease.

Data availability

Source data

VCF data used to generate the results were from an Ashkenazim trio, freely-available through the Genome in a Bottle Consortium at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/[27]:

- Child: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh38/supplementaryFiles/HG002_GRCh38_CHROM1-22_v4.1_highconf.vcf.gz

- Mother: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG004_NA24143_mother/latest/GRCh38/HG004_GRCh38_GIAB_highconf_CG-Illfb-IllsentieonHC-Ion-10XsentieonHC_CHROM1-22_v.3.3.2_highconf.vcf.gz

- Father: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG003_NA24149_father/latest/GRCh38/HG003_GRCh38_GIAB_highconf_CG-Illfb-IllsentieonHC-Ion-10XsentieonHC_CHROM1-22_v.3.3.2_highconf.vcf.gz

Extended data

Zenodo: dmiller903/CompoundHetVIP: CompoundHetVIP - v1.1. https://doi.org/10.5281/zenodo.4477686[12].

This project contains the following extended data:

- CompoundHetVIP_example.pdf (detailed step-by-step example)

Software availability

Software available from: https://hub.docker.com/r/dmill903/compound-het-vip

Source code available from: https://github.com/dmiller903/CompoundHetVIP

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4477686[12].

License: MIT

References

1. Kamphans P, Sabri T, Zhu N, *et al.*: **Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees.** *PLoS One.* 2013; **8**(8): e70151.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Piane M, Molinaro A, Soresina A, *et al.*: **Novel compound heterozygous mutations in a child with Ataxia-Telangiectasia showing unrelated cerebellar disorders.** *J Neurol Sci.* 2016; **371**: 48–53.
PubMed Abstract | Publisher Full Text

3. Li R, Pradhan M, Xu M, *et al.*: **Generation of an induced pluripotent stem cell line (TRNDi002-B) from a patient carrying compound heterozygous p.Q208X and p.G310G mutations in the *NGLY1* gene.** *Stem Cell Res.* 2019; **34**: 101362.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Miller D, Piccolo S: **Compound Heterozygous Variants in Pediatric Cancers: A Systematic Review.** *Front Genet.* 2020; **11**: 493.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Duitama J, McEwen GK, Huebsch T, *et al.*: **Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques.** *Nucleic Acids Res.* 2012; **40**(5): 2041–2053.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Zheng GXY, Lau BT, Schnall-Levin M, *et al.*: **Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.** *Nat Biotechnol.* 2016; **34**(3): 303–311.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Choi Y, Chan AP, Kirkness E, *et al.*: **Comparison of phasing strategies for whole human genomes.** *PLoS Genet.* 2018; **14**(4): e1007308.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Delaneau O, Howie B, Cox AJ, *et al.*: **Haplotype estimation using sequencing reads.** *Am J Hum Genet.* 2013; **93**(4): 687–696.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet.* 2007; **81**(5): 1084–1097.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Loh PR, Danecek P, Palamara PF, *et al.*: **Reference-based phasing using the Haplotype Reference Consortium panel.** *Nat Genet.* 2016; **48**(11): 1443–1448.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Piccolo SR, Frampton MB: **Tools and techniques for computational reproducibility.** *GigaScience.* 2016; **5**(1): 30.
PubMed Abstract | Publisher Full Text | Free Full Text

12. dmiller903: **dmiller903/CompoundHetVIP: CompoundHetVIP - v1.1 (Version v1.1).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4477686

13. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–2158.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Poplin R, Ruano-Rubio V, DePristo MA, *et al.*: **Scaling accurate genetic variant discovery to tens of thousands of samples.** *bioRxiv.* 2017; 201178.
Publisher Full Text

15. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics.* 2011; **27**(21): 2987–2993.
PubMed Abstract | Publisher Full Text | Free Full Text

16. **Picard Tools**.
Reference Source

17. Purcell S, Chang C: **PLINK 2.0**.

18. Chang CC, Chow CC, Tellier LC, *et al.*: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *GigaScience.* 2015; **4**: 7.
PubMed Abstract | Publisher Full Text | Free Full Text

19. 1000 Genomes Project Consortium, Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Edge P, Bafna V, Bansal V: **HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies.** *Genome Res.* 2017; **27**(5): 801–812.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Martin M, Patterson M, Garg S, *et al.*: **WhatsHap: fast and accurate read-based phasing.** *bioRxiv.* 2016; 085050.
Publisher Full Text

22. Paila U, Chapman BA, Kirchner R, *et al.*: **GEMINI: integrative exploration of genetic variation and genome annotations.** *PLoS Comput Biol.* 2013; **9**(7): e1003153.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Tan A, Abecasis GR, Kang HM: **Unified representation of genetic variants.** *Bioinformatics.* 2015; **31**(13): 2202–2204.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Cingolani P, Platts A, Wang LL, *et al.*: **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 (Austin).* 2012; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text | Free Full Text

25. 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

26. Rentzsch P, Witten D, Cooper GM, *et al.*: **CADD: predicting the deleteriousness of variants throughout the human genome.** *Nucleic Acids Res.* 2019; **47**(D1): D886–D894.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Itan Y, Shang L, Boisson B, *et al.*: **The human gene damage index as a gene-level approach to prioritizing exome variants.** *Proc Natl Acad Sci U S A.* 2015; **112**(44): 13615–13620.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Zook JM, McDaniel J, Olson ND, *et al.*: **An open resource for accurately benchmarking small variant and reference calls.** *Nat Biotechnol.* 2019; **37**(5): 561–566.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Zhou X, Tian F, Sandzén J, *et al.*: **Filamin B deficiency in mice results in skeletal malformations and impaired microvascular development.** *Proc Natl Acad Sci U S A.* 2007; **104**(10): 3919–3924.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Yang CF, Wang CH, H'ng WS, *et al.*: **Filamin B Loss-of-Function Mutation in Dimerization Domain Causes Autosomal-Recessive Spondylocarpotarsal Synostosis Syndrome with Rib Anomalies.** *Hum Mutat.* 2017; **38**(5): 540–547.
PubMed Abstract | Publisher Full Text

31. Frasa MAM, Maximiano FC, Smolarczyk K, *et al.*: **Armus is a Rac1 effector that inactivates Rab7 and regulates E-cadherin degradation.** *Curr Biol.* 2010; **20**(3): 198–208.
PubMed Abstract | Publisher Full Text

32. Xu W, Zhao X, Wang X, *et al.*: **The Transcription Factor Tox2 Drives T Follicular Helper Cell Development via Regulating Chromatin Accessibility.** *Immunity.* 2019; **51**(5): 821–839.e5.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 24 February 2021

https://doi.org/10.5256/f1000research.54440.r79307

✔ **Soohyun Lee**
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

I thank the authors for addressing all my comments and I have no further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, transcriptomics, genomic variant calling and analysis, software and algorithm development, cloud infrastructure, docker, genomic pipeline

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.

Version 1

Reviewer Report 25 January 2021

https://doi.org/10.5256/f1000research.29647.r76628

✔ **Sushant Patil**
Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

Thank you for the opportunity to review the manuscript "CompoundHetVIP: Compound Heterozygous Variant Identification Pipeline" by Miller *et al.*

I find the published software to be of great utility, even though it may not have scientific novelty as such. The manuscript is neat and well-presented. I recommend this manuscript for publication in the journal. However, I request the authors to address the following points in the final revision:

1. It appears that the program can only be used with whole-genome sequencing VCFs. Is that true? How about VCFs generated from WES or targeted panel?

2. Step 5 - "CompoundHetVIP can perform phasing using SHAPEIT2, Eagle210, and/or Beagle9.". Is there a 'default' mode of execution for the software? If not, I suggest to have one and include one or all of these phasing programs as part of it.

3. In Figure 1, I recommend labeling each box with the corresponding Step # to relate it to the text in the manuscript.

4. Ideally all VCFs should be 1-based. In a hypothetical scenario, if the parents' VCFs are 0-based and the child's 1-based (coming from different workflows), will the program be able to detect and handle such discrepancy?

5. I recommend saying minor allele frequency (MAF) as Global minor allele frequency (MAF) at first instance. It is easy to be mistaken that for mutant allelic frequency in cancer genomics.

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

Yes

Is the description of the software tool technically sound?

Yes

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

Yes

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: Bioinformatics, Computational Biology, NGS/HTS, Cancer/Oncology, Personalized Medicine

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.

Author Response 02 Feb 2021

Dustin Miller, Brigham Young University, Provo, USA

Thank you for taking the time to review our manuscript. We appreciate your thorough and helpful comments. We have addressed these comments below in a point-by-point response. These changes have made the manuscript more informative and insightful.

The reviewers comments are in regular font. Our responses are in **bold**.

1. It appears that the program can only be used with whole-genome sequencing VCFs. Is that true? How about VCFs generated from WES or targeted panel?

Thank you for pointing this out. We designed this pipeline with WGS in mind and did not think about other types of next-gen data. To test the pipelines on WES/targeted exome data, we ran the pipeline on a publicly available dataset available here: https://my.pgp-hms.org/profile/huF85C76. The data we used was the "Helix Exome+ WES" VCF file (https://my.pgp-hms.org/user_file/download/3888). After filtering the data for positions that passed quality filters and had a quality > 20, we started on step 3 and went through all remaining steps. The pipeline worked well. We have made this more clear in the manuscript.

2. Step 5 - "CompoundHetVIP can perform phasing using SHAPEIT2, Eagle210, and/or Beagle9.". Is there a 'default' mode of execution for the software? If not, I suggest to have one and include one or all of these phasing programs as part of it.

We have a separate script for all 3 phasing programs. That way the user can choose which program to use. Within each script, the phasing software is executed under default parameters. Our example PDF uses SHAPEIT2 and we recommend this as the default phaser to use, especially when using trio data. However, we feel that the choice of phasing program is fundamental to our pipeline and should require an explicit decision from the user; so we require the user to select one. We have made this more clear in the example documentation (https://github.com/dmiller903/CompoundHetVIP/blob/master/CompoundHetVIP_example.pdf) and in the manuscript.

3. In Figure 1, I recommend labeling each box with the corresponding Step # to relate it to the text in the manuscript.

Thank you for the recommendation. We have updated the figure to show step numbers.

4. Ideally all VCFs should be 1-based. In a hypothetical scenario, if the parents' VCFs are 0-based and the child's 1-based (coming from different workflows), will the program be able to detect and handle such discrepancy?

This pipeline assumes input VCF files are 1-based. We added this disclaimer to the example documentation referenced in response 2. During step 11, GEMINI 0-bases the

positions. However, during step 12, they are reverted back to 1-based positions for consistency.

5. I recommend saying minor allele frequency (MAF) as Global minor allele frequency (MAF) at first instance. It is easy to be mistaken that for mutant allelic frequency in cancer genomics.

Thank you for this recommendation. We have updated the manuscript to say "global minor allele frequency" when first referenced.

In addition to the changes mentioned above, we have made a few other changes we would like to mention here:

1. We changed the way we filter the variants to be more simple but also more strict. Originally we allowed some variants to be considered if they were of "HIGH" putative impact but were missing CADD or MAF values. To obtain the variants most likely to play a role in disease, we made the requirement more strict. Now, CADD and MAF scores must be available for all variants, variants must be in exonic regions, and variants can have either "MED" or "HIGH" putative impact. Making these changes did not change the number of CH variants identified in the example dataset, but will likely have an effect in other datasets. This change in how variants are filtered may provide the users with a smaller subset of putative variants, but the CompoundHetVIP will provide more information for the users to use in downstream analysis (i.e. CADD and MAF values for all variants rather than just some of the variants). Our description of the filtering method has been updated in the manuscript.

2. There was an error in our script that identified homozygous alternate variants. Some of our previous code required that a gene have 2 variants in order to be written to the output file. This has been fixed, and it slightly changed the results. Originally we reported only 1 gene with a homozygous alternate variant in the example dataset. After updating the script and running it again, there are 2 genes that have a homozygous alternate variant. This result has been modified in the manuscript.

3. Minor updates to the example PDF (https://github.com/dmiller903/CompoundHetVIP/blob/master/CompoundHetVIP_example.pdf) were made for added clarity and explanation.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 January 2021

https://doi.org/10.5256/f1000research.29647.r76619

Soohyun Lee
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

The paper describes a software pipeline that identifies compound heterozygous (CH) variants from the VCF or gVCF files of whole genome sequencing data from either a single sample or a trio. The pipeline has 13 steps and in brief, performs VCF file merging, liftover from hg38 to hg37, haplotype phasing, variant annotation and filtering based on minor allele frequency (MAF) and Combined Annotation Dependent Depletion (CADD) scores. The pipeline is dockerized for portable and reproducible analysis.

The paper addresses an important problem of identifying CH variants. Many researchers seem to end up writing their own custom CH variant detecting tools and it could be a useful addition if that can be avoided.

I have a few comments which hopefully would improve the clarity of the article and the usability of the pipeline. Before we get to that, I think I have to disclose that I myself and my team member designed and wrote our own CH variant caller as a part of our own variant calling/analysis pipeline, with the help of several experts in the genetics field (not necessarily for publication but for our own use). Therefore, my review is based on my experience tackling a similar problem but most likely with a different set of requirements/constraints.

1. For better clarity and flow of the article, it may be helpful to describe relatively early on (e.g. in the title or early in the abstract) what the pipeline does, including the data type (e.g. whole genome sequencing) and starting point (e.g. VCF/gVCF). The steps the pipeline offers could also be mentioned earlier. For example, it is not clear until near the end of the article that filtering based on MAF and CADD is performed as a part of the pipeline, and one could wonder simply reading the abstract, why only two genes were identified to have CH variants, which is much fewer than expected from a whole genome sequencing data. It is also not clear in the beginning that variant annotation and an optional GATK4-based variant calling is part of the pipeline.

2. I think the pipeline usability could be improved if it accommodates previously merged and annotated VCF files. A user may start with a VCF that is annotated with their annotation tool of choice (e.g. VEP which is a more recent tool than SnpEff) that they may want to use for other analysis consistently in addition to CH variant detection. It would be useful to support such a use case. Alternatively, the user may have a pre-merged VCF file generated by GATK4 from joint variant calling with the parameters of their own choice. The pipeline's VCF merging step performs GATK4-based variant calling which may make more sense to be done along with the other GATK4-based upstream steps rather than as part of the compound het detection step, though that can be a matter of preference. Having this kind of modularity could appeal more to bioinformaticians who are setting up a full variant calling and analysis pipeline and just need a CH detection module. It may also be computationally more efficient to first annotate variants with which genes they overlap with and remove intergenic variants before performing CH detection because CH variants by definition would only occur inside a gene.

3. The pipeline performs population-based haplotype phasing based on e.g. SHAPEIT2 to identify CH variants. It is not very convincing though how much this adds compared to simply scanning a VCF file once to collect all pairs of heterozygous variants in a gene whose

parents' genotypes indicate that one comes from the mother and the other from the father. The authors of the latest version of the phasing tool, SHAPEIT4, seems to have used Mendelian logics as the truth set to evaluate the performance of its phasing imputation, suggesting that when the parents' genotypes are directly available, it maybe be more accurate to use those genotypes than using a population-based imputation.[1] It may be helpful for a proband-only case where the parents' genotypes are not available. However, for a trio case, detecting CH variants from a previously annotated VCF file by simply comparing the genotypes of the parents would take one step of scanning the file (rather than 13 steps). I think having a more convincing rationale about why these additional steps are needed for a trio case would be very helpful.

4. I think the pipeline can definitely be useful as it is for specific users who do not have their own variant calling or annotation capacity and took an unannotated VCF or gVCF file from an external source and simply want to perform CH variant identification without parent information. Again, going back to point 1), having a full description of what the pipeline offers (including variant annotation, filtering, even running GATK4) in the title and early on in the abstract could make it easier to reach the right type of potential users.

5. Lastly a minor point, it would be useful to clarify if the pipeline also works for whole exome sequencing data.

References

1. Delaneau O, Zagury J, Robinson M, Marchini J, et al.: Accurate, scalable and integrative haplotype estimation. *Nature Communications*. 2019; **10** (1). Publisher Full Text

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

No

Is the description of the software tool technically sound?

Yes

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?

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: genomics, transcriptomics, genomic variant calling and analysis, software and

algorithm development, cloud infrastructure, docker, genomic pipeline

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.

Author Response 02 Feb 2021

Dustin Miller, Brigham Young University, Provo, USA

Thank you for taking the time to review our manuscript. We appreciate your thorough and helpful comments. We have addressed these comments below in a point-by-point response. These changes have made the manuscript more informative and insightful.

The reviewers comments are in regular font. Our responses are in **bold**.

1. For better clarity and flow of the article, it may be helpful to describe relatively early on (e.g. in the title or early in the abstract) what the pipeline does, including the data type (e.g. whole genome sequencing) and starting point (e.g. VCF/gVCF). The steps the pipeline offers could also be mentioned earlier. For example, it is not clear until near the end of the article that filtering based on MAF and CADD is performed as a part of the pipeline, and one could wonder simply reading the abstract, why only two genes were identified to have CH variants, which is much fewer than expected from a whole genome sequencing data. It is also not clear in the beginning that variant annotation and an optional GATK4-based variant calling is part of the pipeline.

Thank you for these recommendations. We have updated the abstract to get to the point more quickly, to mention the tools used for each major step, and to clarify that MAF and CADD are used as part of variant identification. We also updated the abstract to mention that whole-exome and targeted exome data can also be used. This updated abstract should provide a clearer understanding of what tools are included in the pipeline, and what data can be used as initial input.

2. I think the pipeline usability could be improved if it accommodates previously merged and annotated VCF files. A user may start with a VCF that is annotated with their annotation tool of choice (e.g. VEP which is a more recent tool than SnpEff) that they may want to use for other analysis consistently in addition to CH variant detection. It would be useful to support such a use case. Alternatively, the user may have a pre-merged VCF file generated by GATK4 from joint variant calling with the parameters of their own choice. The pipeline's VCF merging step performs GATK4-based variant calling which may make more sense to be done along with the other GATK4-based upstream steps rather than as part of the compound het detection step, though that can be a matter of preference. Having this kind of modularity could appeal more to bioinformaticians who are setting up a full variant calling and analysis pipeline and just need a CH detection module. It may also be computationally more efficient to first annotate variants with which genes they overlap with and remove intergenic variants before performing CH detection because CH variants by definition would only occur inside a gene.

Thank you for your thoughtful comments about the modularity of this pipeline. This is something we have thought about over the course of this project, and we want this pipeline to be flexible for diverse types of usage. We have updated the "gemini_load.py" script with an optional argument to allow users to import a file that was annotated with VEP. Because SnpEff is used as part of the pipeline, the script defaults to SnpEff; but if a user chooses to use VEP outside of the pipeline, they can still apply the remaining steps of the pipeline. Intergenic regions are not considered during the CH detection step. Within the "identify_CH_variants.py" script, the GEMINI database is converted to a TSV file; and during this process, intergenic regions are removed to speed up the CH detection process. Accordingly, it would be fine if researchers uploaded a VCF file from which intergenic regions have already been removed. In addition, we will be actively monitoring GitHub for any feature recommendations or any issues that arise. For users who are more experienced with coding and aren't afraid of source code, the code is publicly available, and these users can adapt the code to meet their needs.

3. The pipeline performs population-based haplotype phasing based on e.g. SHAPEIT2 to identify CH variants. It is not very convincing though how much this adds compared to simply scanning a VCF file once to collect all pairs of heterozygous variants in a gene whose parents' genotypes indicate that one comes from the mother and the other from the father. The authors of the latest version of the phasing tool, SHAPEIT4, seems to have used Mendelian logics as the truth set to evaluate the performance of its phasing imputation, suggesting that when the parents' genotypes are directly available, it maybe be more accurate to use those genotypes than using a population-based imputation.1 It may be helpful for a proband-only case where the parents' genotypes are not available. However, for a trio case, detecting CH variants from a previously annotated VCF file by simply comparing the genotypes of the parents would take one step of scanning the file (rather than 13 steps). I think having a more convincing rationale about why these additional steps are needed for a trio case would be very helpful.

We considered coding an initial step that identifies all CH variants that can be inferred via Mendelian logic when both parents' genotypes are available. However, we wanted to make the pipeline consistent regardless of whether trio or individual data were used. Furthermore, SHAPEIT2 uses Mendelian logic when the parents' genotypes are available. The only time it uses compact hidden Markov model (CHMM) logic with trios is when the phase of the child cannot be determined using Mendelian logic alone (i.e. both parents are heterozygous, and the child is as well). Therefore, SHAPEIT2 provides the benefit of Mendelian logic and CHMM, when necessary. (The SHAPEIT4 paper also notes that it used the 500 trio children as validation only when they could be "phased using Mendel inheritance logic (i.e., no triple hets and no Mendel inconsistencies).") In addition, there may be scenarios where data is missing for one or both parents for a certain gene region, and population-based phasing with SHAPEIT2 has the added benefit of imputing the missing information. We have updated the manuscript to clarify these points, as suggested.

4. I think the pipeline can definitely be useful as it is for specific users who do not have their own variant calling or annotation capacity and took an unannotated VCF or gVCF file from

an external source and simply want to perform CH variant identification without parent information. Again, going back to point 1), having a full description of what the pipeline offers (including variant annotation, filtering, even running GATK4) in the title and early on in the abstract could make it easier to reach the right type of potential users.

Please see response # 1.

5. Lastly a minor point, it would be useful to clarify if the pipeline also works for whole exome sequencing data.

Thank you for bringing this up. We designed this pipeline with WGS in mind, and did not think about other types of next-gen sequencing data. To test the pipelines on WES/targeted exome data, we ran the pipeline on a publicly available dataset available here: https://my.pgp-hms.org/profile/huF85C76. The data we used was the "Helix Exome+ WES" VCF file (https://my.pgp-hms.org/user_file/download/3888). After filtering the data for positions that passed quality filters and had a quality > 20, we started on step 3 and went through all remaining steps. The pipeline worked well. We have made this more clear in the manuscript.

In addition to the changes mentioned above, we have made a few other changes we would like to mention here:

1. We changed the way we filter the variants to be more simple but also more strict. Originally we allowed some variants to be considered if they were of "HIGH" putative impact but were missing CADD or MAF values. To obtain the variants most likely to play a role in disease, we made the requirement more strict. Now, CADD and MAF scores must be available for all variants, variants must be in exonic regions, and variants can have either "MED" or "HIGH" putative impact. Making these changes did not change the number of CH variants identified in the example dataset, but will likely have an effect in other datasets. This change in how variants are filtered may provide the users with a smaller subset of putative variants, but the CompoundHetVIP will provide more information for the users to use in downstream analysis (i.e. CADD and MAF values for all variants rather than just some of the variants). Our description of the filtering method has been updated in the manuscript.

2. There was an error in our script that identified homozygous alternate variants. Some of our previous code required that a gene have 2 variants in order to be written to the output file. This has been fixed, and it slightly changed the results. Originally we reported only 1 gene with a homozygous alternate variant in the example dataset. After updating the script and running it again, there are 2 genes that have a homozygous alternate variant. This result has been modified in the manuscript.

3. Minor updates to the example PDF (https://github.com/dmiller903/CompoundHetVIP/blob/master/CompoundHetVIP_example.pdf) were made for added clarity and explanation.

Competing Interests: No competing interests were disclosed.