Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:_____ _____

Kevin Johnson Date

Web Tool for Clinic Trio Based Sequence Data Analysis to Identify Potential Pathogenic Variants for Rare Genetic Diseases

By

Kevin Johnson
Master of Public Health

Epidemiology

_____ [Chair's signature]

Dr. Anke Huels
Committee Chair

_____ [Member's signature]

Dr. Jingjing Yang
Committee Member

Web Tool for Clinic Trio Based Sequence Data Analysis to Identify Potential Pathogenic Variants for Rare Genetic Diseases

By

Kevin Johnson

Bachelor of Science
University of Minnesota
2017

Thesis Committee Chair: Dr. Anke Huels PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2022

Abstract

Web Tool for Clinic Trio-Based Sequence Data Analysis to Identify Potential Pathogenic Variants for Rare Genetic Diseases

By Kevin Johnson

An important tool in the diagnosis of rare genetic diseases is whole exome or genome sequencing (WES/WGS). It is believed that many rare genetic diseases are caused by rare variations in the patients' genome. Typically, clinical WES/WGS is done for patients and their parents in an attempt to find potential pathogenic variations. Advanced bioinformatics skills are needed to analyze these WES/WGS data. Variant annotations about the variant position in a gene, biological functions, and possible pathogenicity need to be referred to from genomic databases. WES/WGS data with variant annotation need to be cleaned, filtered, and presented in a meaningful way. An automated web tool for the processing and analysis of trio-based WES/WGS data could help clinicians diagnose rare genetic diseases. I have developed a workflow process that automatically, which will create a table of possibly pathogenic variants and display the output table alongside the input WES/WGS data in a webtool. The tool makes the process of variant identification and visual review easier and quicker in the diagnosis of rare genetic diseases.

Web Tool for Clinic Trio Based Sequence Data Analysis to Identify Potential Pathogenic Variants for Rare Genetic Diseases

By

Kevin Johnson

Bachelor of Science
University of Minnesota
2017

Thesis Committee Chair: Dr. Anke Huels PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2022

Introduction

Rare diseases are typically defined as those that affect around 1 in 2,000 individuals (0.05%)[1]. It is estimated that up to 80% of these rare diseases have an important genetic component[2]. Determining the exact cause of a genetic disorder can improve disease management and lead to better patient outcomes. It is observed that in general clinical settings using next generation sequencing (NGS) for whole exome or genome sequencing (WES/WGS), the diagnostic rate for rare genetic diseases varies between 20% and 50%[3]. Providing convenient tools to help clinicians analyze WES/WGS data could improve the diagnostic rate and patient outcomes.

Genetic diseases can be mono-genetic (primarily caused by a single gene), poly-genetic (caused by multiple genes), or caused by a structural chromosomal abnormality. Mono-genetic disorders can be further classified as due to Mendelian inheritance, de novo mutation, or both. Mendelian means the mutation in question follows Mendelian inheritance and is present in one or both of the parents, at least in carrier form. De novo mutations are spontaneous changes to the genome, typically occurring early in embryonic development. Sometimes a disorder requires two copies of an allele, and one can be inherited and one can be a de novo mutation. De novo mutations are of special concern, as they are rarer and subject to less evolutionary pressure and could be more likely than inherited variants to cause rare genetic diseases[4]. Structural chromosomal abnormalities are where large pieces or entire chromosomes are duplicated or missing and depending on the change can determined before birth or will need sequencing to determine. Polygenic disorders can have many causal mutations, each with small effects that combine to a pathogenic phenotype, these polygenetic diseases are more common than rare

diseases[5]. Rare genetic diseases are more likely to be related to mono-genetic or small structural chromosomal changes than to polygenic effects[6].

Nowadays, the process for determining a genetic cause for a rare disease would use the sequencing tools of WES/WGS. WGS data provides information on the entire genome while WES focuses on the exome, or protein coding regions. A typical workflow involves sequencing patients and their parents, aligning sequencing data to a reference genome, determining variants present in the family, obtaining annotation information about the variants, filtering variants, and finally outputting a report for the identification of potentially pathogenic variants and genes.

On average, an individual will have around 5 million total variants and 500,000 variants on known regulatory regions[7]. The vast majority of these variants are benign. To be useful to clinicians, variant data needs to be cleaned, filtered, and presented in a meaningful way. Before Variants can be filtered, information about the variants needs to be compiled. Using that information, a filtering scheme needs to be able to identify possibly pathogenic variants. Various methods exist to attempt to quantify how likely a variant is to be pathogenic but there is not a formalized test or process. Instead, variants are typically ranked or filtered according to their differing attributes with common, or likely benign variants removed.

These filtered variants along with their functional annotations can then be included in a report of possibly pathogenic variants to provide clinicians additional information, especially for the cases no known pathogenic variants are identified. To help clinicians make diagnosis with the report, visual review of the variant annotation information along with their raw WES/WGS variant call files or aligned sequence read files is needed. Variant review can help check the region of a variant to see nearby genes or check the sequence depth to confirm a variant. A few

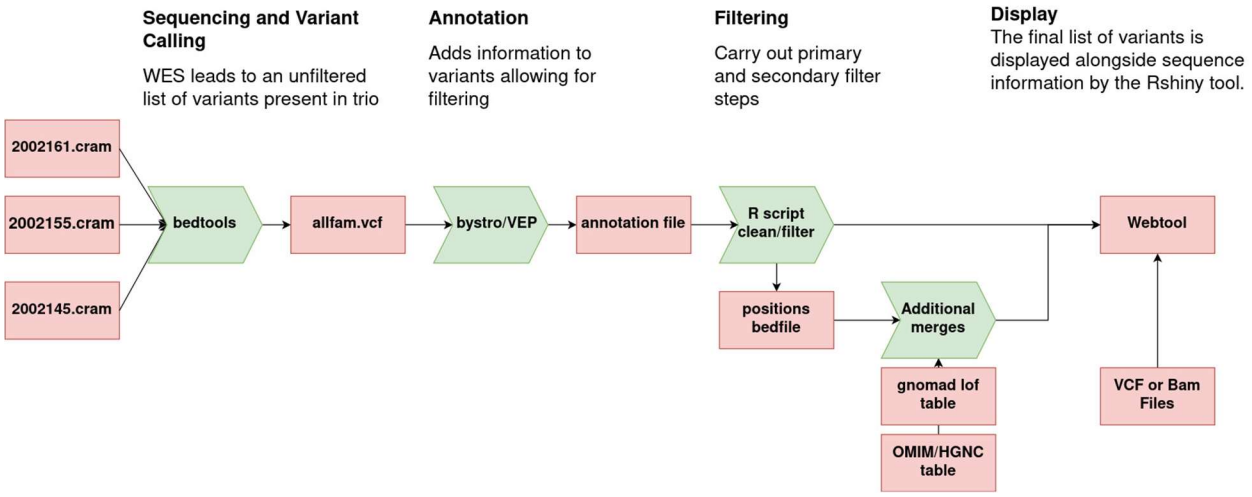
tools already exist to allow clinicians and researchers to visually inspect sequencing data and review variants, including the Integrative Genomics Viewer (IGV)[8].

Here we introduce a web tool that simplifies the process of variant filtering and visual review. The webtool combines the functionality of IGV with an interactive table that allows the user to explore possibly pathogenic variants and their functional annotations. The webtool displays filtered, annotated variants and presents those variants along with a viewing instance of the genome. The table of variant locations and attributes is linked to the viewing instance allowing for quicker review of variants and their attributes.

Methods

A pipeline was developed to annotate and filter genetic variants called from WES/WGS raw read data and saved as variant call files (VCFs). The current process of annotating and filtering the data in the developed pipeline is outlined in figure 1. An output file containing a subset of potential pathogenic variants and their functional annotations is generated by the pipeline. A webtool developed using R Shiny then visualizes the output file of the pipeline, along with the corresponding VCF and aligned sequence files. The process was used on a trial data set received from a clinical setting.

Figure 1 Data pipeline for Pathogenic Variant Identification



Sequencing and Variant Calling

WGS/WES on the proband (patient) and close family members is the first step of most NGS workflows. The data is then aligned to a reference genome. Variant calling is the process of determining the differences in the target sequence from a reference DNA sequence. Sequencing, alignment, and variant calling yield the candidate variants that could be related to the rare disease.

The current process for the webtool is functional for both WGS and WES on a single individual or multiple related individuals. The data is expected in the form of a variant call format file. A VCF file only stores variations from a reference genome rather than an entire sequence[9]. Using a bioinformatics package, Samtools, multiple samples can be merged into a single file for use in the annotation process[10].

Annotation

Annotation is the process of labeling known biological function information to the target data set of genetic variants. Often this includes the variants' locations and if they are on or near known coding regions. Common annotation software will categorize variant locations as intergenic regions, introns, exons, splice donors and splice acceptors, three and five prime

untranslated sites, and ncRNA coding sites. Many of the variants will be on intergenic regions and will not directly code for genes. However, some intergenic variants have been known to cause changes in gene expression and effect disease phenotypes[11]. Variants can also be located on genes. Variants can be located on introns, where they will not code for amino acids and on exons, where they will code amino acids. Variants can also be in regions where non-coding RNA (ncRNA) is coded. ncRNA is RNA that will ultimately not be translated into protein. Most ncRNA has no known function, but some ncRNA can affect gene expression[12]. Variants can also occur on splice acceptor and donor sites. These are areas at the beginning and ends of exons and introns where transcripts will be spliced together. Variants can also be located on three prime untranslated regions or five prime untranslated regions. These are areas right before or after start and stop codons. Where a variant is located can be important to determine the possible effects a variant could have. Variants located in coding regions or that affect the starting or stopping of transcription may be likelier to cause disease. However, variants in these locations may not always have deleterious effects.

Annotation sources will often provide the coding effect a variant will have. Annotations will describe a variant as synonymous, non-synonymous, indel-frameshift, indel-nonframeshift, startLoss, stopLoss, and stopGain. Synonymous variants will not change the amino acid coded for on a codon, while non-synonymous variants will. Changing an amino acid may change the protein and lead to loss of function for a gene. Indel-frameshift variants are insertions or deletions that will change the reading frame during transcription, often leading to a non-functional protein. StartLoss and stopLoss variants are single variations on start or stop codons. StopGain variants are single nucleotide variations that lead to a new stop codon being created.

Variants that interfere with the starting and stopping of transcription are more likely to lead to non-functional proteins.

Annotation sources can add more information in addition to variant sites and possible coding effects of variants. The exact amino acid coded for, the exact codon and transcript the variant is on, and outside databases entries on variants can be included. Outside databases can give frequency of a variant among certain populations or if a variant has been flagged as pathogenic in any clinical databases. Often annotation will add genome wide variant scores that attempt to predict conservation and deleteriousness. These scores are based on algorithms that attempt to predict the conservation or deleterious effects of variants. Deleterious scores try to measure harmful variants. Conservation scores try to measure how quickly or slowly a region of the genome evolves, with the assumption that slowly evolving regions are evolutionary important and variants in these regions could be harmful.

In this project, the variants were annotated using Bystro and the Variant Effect Predictor (VEP) tool from the Ensembl Genomes Project [13] [14]. The annotation process will refer to databases from the NCBI Reference Sequence (refSeq) project, the Single Nucleotide Polymorphism Database (dbSNP), the Clinvar database, and the genome Aggregation Database (gnomAD).

The refSeq database was used for information about the structural effect of variants and the location in the genome[15]. The dbSNP was used for additional information about single nucleotide polymorphisms including unique identifiers[16]. Clinvar was important to access for information on disease causing or pathogenic variants[17]. GnomAD gave information on allele frequency in different populations needed for filtering[18].

Filtering

Annotated variants need to be filtered to identify the possibly pathogenic variants. It is generally agreed upon which variables to use as filters, but the exact thresholds used can vary. Filtering typically takes place in two stages, primary and secondary. Primary filtering typically accounts for read quality, minor allele frequency, coding effects of variants, and if available family segregation. Not all variants are of the same quality and sequencing tools will give some sort of quality score reflecting the probability that a variant exists a location. Minor allele frequency refers to how often a variant appears in a given sub-population. Rarer variants are more likely to be disease causing and a minor allele frequency of 1% is often the upper limit of filtering[19]. Family segregation is powerful for identifying de novo variants and relies on parental sequencing data being available. Secondary filtering is based off genome wide variant scores.

Before the webtool can display data, output from both Bystro and VEP annotation are read into and filtered by an R script. The webtool displays the resulting table. The filtering scheme is based on existing literature and was separated into primary and secondary filtering. The primary filter was based on VCF quality scores, allele frequency, family segregation, allele function, allele site type, and the variants presence in the clinvar database. The secondary filter used cadd, phastCons, and phyloP scores.

The first filter was removing variants below a certain quality score threshold. The VCF quality scores were based on a PHRED scaling[20].

$$\text{PHRED quality} = -10log(Basecallingerrorprobability)$$

A higher quality implies a lower error probability. A quality score of 10 corresponds to 90% accuracy and 20 to 99%. Variants with a quality score below 20 were excluded. Family segregation means variants were excluded if they were not present in the target individual, only

present in their parents. Variants could exist on multiple transcripts and therefore could have multiple values for certain fields. Variants were excluded if they did not have a listed function in refSeq and were only located on intergenic sites. Novel variants or variants with allele frequency of less than 1% in gnomad were included in the final output table.

For secondary filtering, three different genome wide variant scores were used; cadd, phastCons, and phyloP. Cadd is the Combined Annotation Dependent Depletion score for determining the deleteriousness of simple insertions, deletions, or single nucleotide variants[21]. Cadd scores are scaled similarly to PHRED scores. A cadd score of 20 implies a variant is in the 1% most deleterious variants and 30 would be among the 0.1%. A cutoff of 15 was used in line with recommendations from cadd's creators, but pathogenic cutoffs are understood to be fairly arbitrary. Both phastCons and phyloP are algorithms that give conservation scores[22]. PhastCons accounts for the effects of neighboring bases while phyloP does not. PhyloP is scaled 0 to 1 with values nearer to 1 evolving slowly. Phylop is scaled from -20 to 30 with positive values evolving slower than expected and negative values evolving faster. Variants were included if they had phyloP scores above 3 or PhastCons scores greater than or equal to 0.9. A variant was filtered out only if was below all three score thresholds.

Additionally, any variant with a clinvar entry with a clinical significance rating other than "benign" was automatically included in the final outcome, regardless of any other field values. Clinvar designates the clinical significance of variants in line with recommendations of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology[23]. Clinvar also includes information about number of submitters and the review status of the variant.

Additional Merges

After filtering, some additional cleaning and merges take place. This only applies to filtered variants. Gnomad Loss of Function (lof) tables are designed to give additional information about frameshift, splice donor, splice acceptor, and stop-gain variants[24]. Specifically, these tables aim to give more information about whether variants cause loss of function for the proteins they code for. These tables added probability of loss of function intolerance (pLI) scores and observed expectation ratios (oe). PLI scores attempt to identify genes that are cannot tolerate truncating mutations. A gene that changes phenotypes after a single loss of function mutation is known as a haploinsufficient gene, and pLI can be interpreted as the probability of a gene being haploinsufficient. Oe ratios show the difference in the amount of observed LoF variants to the expected amount of LoF varaints if the rate of LoF variants was solely governed by chance.

Information from the Human Gene Nomenclature Committee was added to get gene descriptions and allow links to be created to outside sources. Links were created to the Gene Cards website (https://www.genecards.org), pubmed (https://pubmed.ncbi.nlm.nih.gov), the Online Mendelian Inheritance in Man data base (OMIM)(https://www.omim.org), the Human Gene Mutation Database(HGMD)(www.hgmd.cf.ac.uk),the dbsnp database (www.ncbmi.nlm.nig.gov/snp), and gnomAD (https://gnomad.broadinstitute.org).

Visualization using the Webtool

The webtool takes as inputs, the filtered variant table obtained from the annotation and filtering process described above and files containing sequencing information. The webtool was developed using the shiny package for R[25 26]. The webtool has two major components, an IGV instance and an interactive table of possibly pathogenic variants. The table is searchable and sort-able across all columns and includes five different options to control the variables displayed. The

table is also linked to the IGV instance, and clicking on an entry in the table will focus the IGV instance around that variant's coordinates. The webtool can incorporate sequencing data from VCF and BAM files. The IGV instance allows the user to set coordinates to view, customize the display of sequencing data, and save images of the current instance. IGV functionality was incorporated using the ivgshiny package[27]. Genetic data was handled using functions from the GenomicAlignments and VariantAnnotation packages[28] [29]. The interactive data table was created using functions from the DT package and functions from the tidyverse package were used in general data preparation[30] [31].

Results

From the trial data set, it was believed the patient may have a CTCF related disorder. CTCF disorders are still the subject of ongoing research but variants on the CTCF gene have been linked with rare genetic diseases[32]. These results are not a formal test of sensitivity or specificity but more of a proof of concept trial project. This data was made available in the form of compressed reference alignment map (CRAM) file. Variant calling was done by the lab that sequenced the data and VCF files were also made available. The filtering process started with 606,315 candidate variants.

The final output table had 3,276 variants and 46 columns of information. A data dictionary with field descriptions and some possible output values is included in appendix A.

The variants could be associated with multiple transcripts, so could be associated with various site types and could have various coding effects. The difference in site type between the filtered variants and all variants is shown in table 1. The difference in coding effect between the filtered variants and all variants is displayed in table 2. Filtered results contain fewer variants on

exonic sites and more variants on all other sites. Filtered results contain fewer variants with no known effect and more variants with any other coding effect.

Table 1. Site Type of Filtered Variants

Site Type	Filtered Variants	All Variants
exonic	1,171 (30.17%)	112,045 (4.99%)
intronic	1,805 (46.5%)	1,743,514 (77.45%)
intergenic	3 (0.08%)	288,614 (12.82%)
ncRNA	561 (14.46%)	20,655 (0.92%)
spliceAcceptor	14 (0.03%)	468 (0.02%)
spliceDonor	17 (0.03%)	529 (0.02%)
UTR3	190 (4.89%)	59,947 (2.66%)
UTR5	120 (3.09%)	25,372 (1.13%)
Total Transcripts	3881	2,251,144

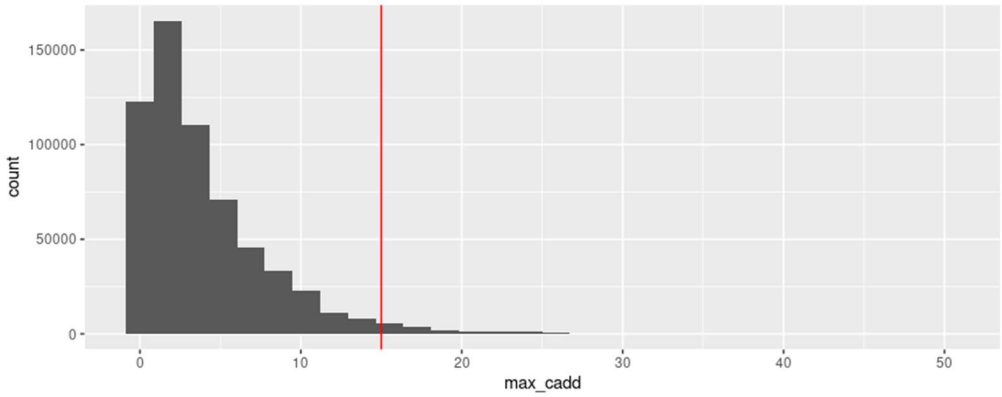
Table 2 Coding Effects from Variants

Coding Effect	Filtered Variants	All Variants
! (No known effect)	2,435 (67.47%)	744,688 (86.94%)
Indel-Frameshift	54 (1.50%)	2,539 (0.30%)
Indel-nonFrameshift	54 (1.50%)	10,421 (1.22%)
nonSynonymous	740 (20.50%)	45,656 (5.33%)
Synonymous	295 (8.17%)	5,276 (0.62%)
startLoss	1 (0.03%)	56 (0.0065%)
stopGain	28 (0.78%)	326 (0.038%)
stopLoss	2 (0.06%)	129 (0.015%)
Total Transcripts	3,609	856,577

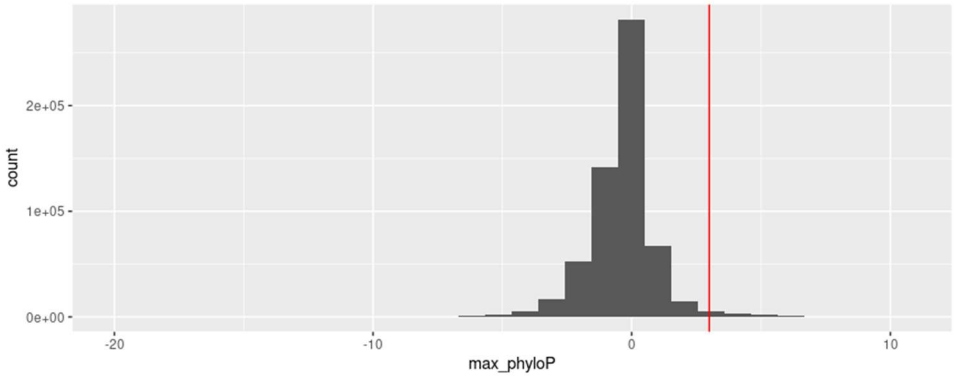
Variants on multiple transcripts were associated with multiple scores. Max scores were used for filtering. Cadd, phastCons, and phyloP scores are displayed on figure 2 and table 3. Filtered variants had higher values for all scores.

Figure 2. Maximum Genome Wide Scores in Unfiltered Variants. Red lines represent filter levels. A) cadd, B) phyloP, and C) phastCons

A



B



C

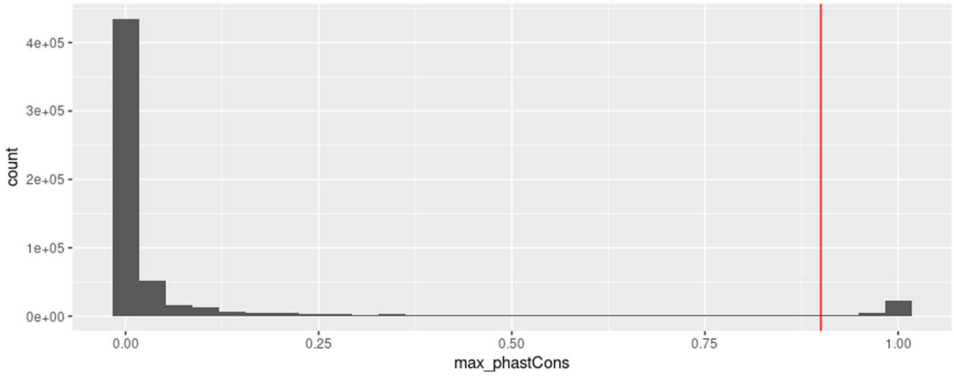


Table 4 Means in Filtered and All Variants for cadd, phyloP, and phastCons.

Score	**Filtered Variants**	**All Variants**
Cadd	12.36(8.37)	3.73(4.02)
phyloP	2.07(2.45)	-0.31(1.25)
phastCons	0.79(0.37)	0.09(0.23)

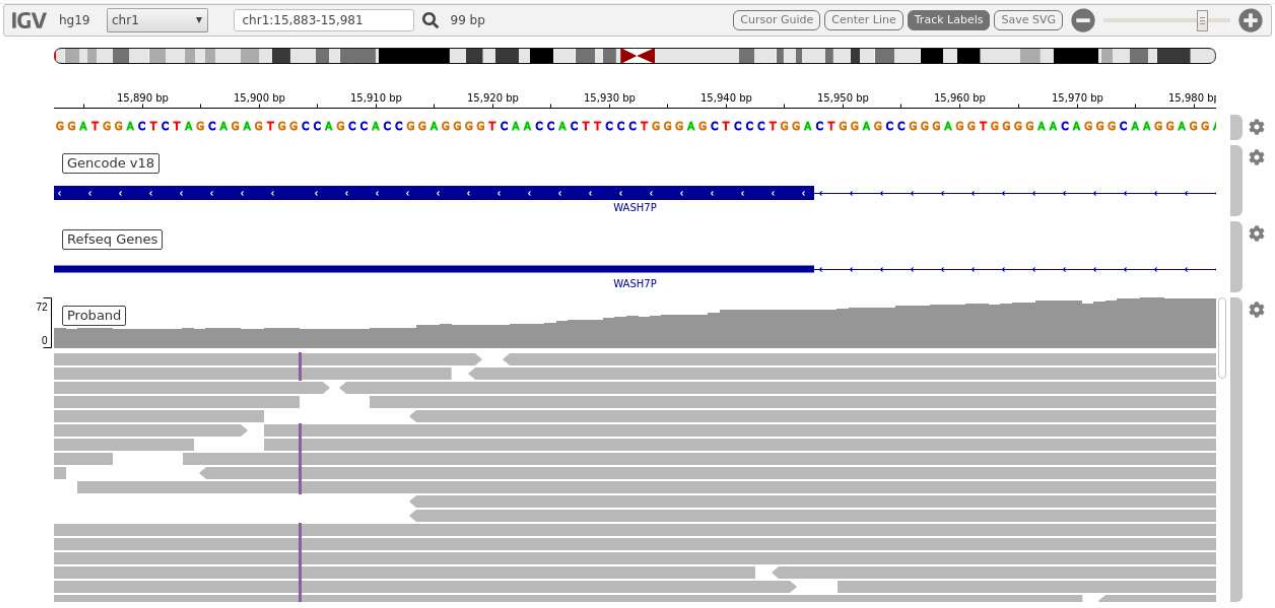
Variants were only included if they were present in the proband. Proband variants shown by parental status at displayed in table 5.

Table 5 Proband Variants by Parental Alleles.

	Father		
Mother	Not Present	Heterozygote	Homozygote
Not Present	845	635	44
Heterozygote	581	889	60
Homozygote	62	44	116

The webtool correctly functions and displays both a table of variants and an IGV instance with sequencing data. These aspects of the webtool can be seen on Figure 3.

Figure 3. Webtool with Bam file loaded



Display Amino Acid and Transcript Information (ref and alt amino acid, cDNA and codon)

Display Gene info and Trio status

Display Genetic Algorithms (phasCons, pLI, oe, phyloP, cadd)

Display Allele Frequency Information

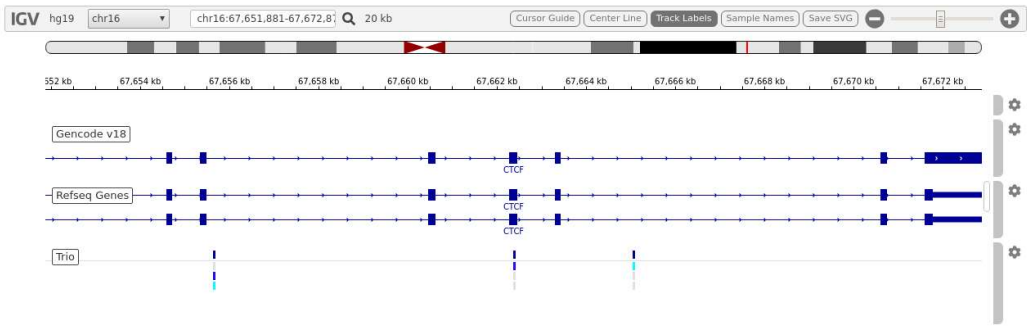
Display Refseq, Clinvar, and other links

Show 5 entries Search:

chrom	chromStart	chromEnd	ID	ref	alt	QUAL
chr1	15903	15904	!	G;C	+C	1623.86
chr1	16977	16978	rs879121945	G	A	189.16

Using the webtool, it was explored if there were any variants related to CTCF or in the region of the CTCF gene. Using the webtool a nonSynonymous variant on an exonic coding region for the CTCF gene was identified (figure 4). This variant had a phastCons score of 1, phyloP score of 3.3, and a cadd score of 27. The variant was novel to the gnomad database but had an entry in Clinvar. The clinvar entry was “likely pathogenic” with only 1 submitter and no assertion criteria provided. The allele was included in a submission without evidence but an interpretation was included.

Figure 4 Example of variant identification process.



Display Amino Acid and Transcript Information (ref and alt amino acid, cDNA and codon)

Display Gene info and Trio status

Display Genetic Algorithms (phasCons, pLI, oe, phyloP, cadd)

Display Allele Frequency Information

Display Refseq, Clinvar, and other links

Show 5 entries Search: chr16

chrom	chromStart	chromEnd	ID	ref	alt	QUAL	gene	hgnc.gene.description	Mother	Father	Proband	pubmed_links
chr16	68719183	68719184	rs35232945	C	T	1628.13	CDH3	cadherin-3	Heterozygote	Not.Present	Heterozygote	https://pubmed.ncbi.nlm.nih.gov/?term=CDH3
chr16	67662377	67662378	!	C	G	1523.13	CTCF	CCCTC-binding-factor	Not.Present	Not.Present	Heterozygote	https://pubmed.ncbi.nlm.nih.gov/?term=CTCF

Discussion

I have developed a pipeline to annotate and filtered WES/WGS data, and a webtool to visualize the output annotated file of a subset of potential pathogenic variants along with their VCF and sequence read files. Source code for the replication of the variant viewer is publicly available as is the R code for the filtering scheme. The command line codes for running the VEP annotation is also included in a read-me file. Currently the source code and instructions for the tool are available at the github page https://github.com/joh11045/variant_viewer/.

The filtering strategy did change the distribution of site types and coding functions of the variants. The variants included in the final table had a lower proportion of intergenic and intronic sites and a higher proportion of every other site type. Similarly, the proportion of variants with no known coding effect fell while all other coding effects rose. Filtered variants had higher

genome wide scores for the three algorithms used compared to all variants. The final table created from the filtering strategy represents about 0.5% of the variants present in the family trio.

So far this thesis report has outlined a method for filtering variants in an attempt to highlight likely pathogenic variants and described the accompanying webtools for variant exploration. It stays mostly in line with recommended practices but diverges in a few key places. It does not explicitly remove synonymous variants. Synonymous variants have been shown to cause rare genetic diseases in some cases[33] [34]. Overall though, most identified variants were non-synonymous. Additionally, using a soft secondary filter was different, the total number of included variants using hard filters would have been 710. The goal of the filtering was to be sensitive enough to capture possibly damaging variants rather than more specific about which variants to include.

The webtool allows for quickly searching and ranking filtered variants and immediately seeing their location on the genome. This can allow researchers and clinicians to get a better sense of the variants that could be pathogenic. Additionally, the tool allows for the examination of the sequencing data. This can allow for better understanding of possible sequencing errors and speed up the process of visually confirming variants.

This method has a few limitations. This is not a tool that can ultimately determine if a variant is casual for a disease, it is only a tool designed to help clinicians and researchers better understand their data. This tool is also sensitive to upstream data errors. It is reliant on accurate sequencing data. Especially with the potential identification of de novo variants. Of the 845 variants not present in either parent, the proband was homogeneous for 83 and heterogeneous for the rest. It is unlikely that there would be double de novo mutations on the same site. "Not present" was re-coded from the inherited "missing" value from the Bystro annotation. Missing

could represent areas where the genome matches the reference genome and there is no variant or areas where there was not enough coverage to have an accurate call. This is problematic for the detection of de novo variants. This is an issue that can be resolved by visual inspection of the sequencing data, but may not be practical at scale. The returned table is searchable, sortable, and allows for different sets data to be displayed on screen. However, the returned table may be difficult to navigate as 3,276 variants are returned with over 40 columns of information for each variant. It may be necessary to tweak the filtering scheme to return fewer variants. Future offerings for the webtool could allow custom filtering of the variants, but currently no such functionality exists.

Future functionality of the webtool will likely include merging the filtering script and webtool display to allow real time filtering of the variants. In addition, the annotation data will be reduced to only include VEP annotation data to simplify the filtering script. Updated versions will be available on Github.

In case of the trial data with the CTCF trio, the webtool was able to identify a nonSynonymous variant on an exonic coding region. The variant was not in the gnomad database, an indication that it has a very low frequency in the population. The genome wide scores associated with the variant indicated that it was likely deleterious and located on a conserved region of the genome where variations could be harmful. The webtool also showed that the variant was not present in either parent and located directly on the CTCF gene. Sequencing data showed it was truly not present in either parent, rather than missing. The proband was heterogeneous, meaning it is likely a de novo mutation. This example illustrates the potential benefits and use case for the webtool as an aid in the review of possibly pathogenic variants. Better identification of possibly pathogenic variants could lead to better understanding

of the underlying biology of the disease. If many possibly pathogenic variants are clustered around a single gene, it could mean that gene is important to the disease. Additionally, it could be used to identify novel variants on known disease-causing genes. Better understanding around the causes of rare genetic diseases will lead to better patient outcomes.

In conclusion, I have developed a tool for the viewing and identification of potentially pathogenic variants out of free, open-source software. The source code can be freely downloaded and the tool can be replicated. This tool could be useful for physicians or geneticists working in a clinical setting to identify and treat rare genetic diseases.

Appendix A

Variable	Description	Values
chrom	Chromosome	
chromEnd	Chromosome End Position	
chromStart	Chromosome Strart Position	
ID	dbSNP id, rs and number	
ref	Reference allele (HG 37)	
alt	The alternative or nonference allele	
QUAL	Quality score from original sequencing	
cDNA	Relative cDNA value from VEP annotation	
refSeq.refAminoAcid	Amino Acid code for reference allele	
refSeq.altAminoAcide	Amino Acid code for the alternative allele	
refSeq.codonNumber	Codon number	
Gene	Gene name	
hgnc.gene.description	Gene description from HUGO Gene Nomenclature Committee	
Father	Allele type of the father	Heterozygote
		Homozygote
		Not Present
Mother	Allele type of the Mother	
Proband	Allele type of proband (cannot be not present)	
pubmed_links	link to pubmed pubmed.ncbi.nlm.nih.gov	
omim_link	Link to OMIM www.omim.org	
PhastCons	conservation score that includes neighboring bases	
pLi	Probability of loss-of-function intolerance; probability that transcript falls into distribution of haplo insufficient genes	
oe_lof	Observed over expected ratio for pLoF variants in transcript	
phyloP	conservation score that does not include neighboring bases	
cadd	score for the deleteriousness of a variant	
gnomad.genomes.af	Non-reference allele frequency across all populations	
gnomad_links	link to gnomad gnomad.broadinstitute.org	
refSeq.nearest.name		

refSeq.siteType	Effect of the alt allele	intronic
		exonic
		UTR3
		UTR5
		spliceAcceptor
		SplicDonor
		ncRNA
		intergenic
refSeq.exonicAlleleFunction	Coding effect of variant	Synonymous
		NonSynonymous
		indel-nonFrameshift
		indel-frameshift
		stopGain
		stopLoss
		StartLoss
clinvar.alleleID	Clinvar unique identifier for a variant	
clinvar.clinicalSignificance	Significance of a variant in clinvar	Benign
		Pathogenic
		Likely Benign or Pathogenic
		Conflicting interprtations
		Uncertain significance
clinvar.type	Type of variant	Single nucleotide variant
		Insertion
		Deletion
		NT expansion
		Duplication
		Indel
clinvar.phenotypeList	Associated phenotypes for variants at this position	
clinvar.numberSubmitters	Number of submissions in clinvar overlapping this position	
clinvar.origin	Orgin tissue of clinical sample	germline
		somatic
		unknow; not provided
clinvar.referenceAllele	reference allele for this position in clinvar	
clinvar.alternateAllele	alternative alleles for this position in clinvar	
clinvar.reviewStatus	Level of review supporting clinical significance values	Reviewed by expert panel

		criteria provided multiple submiters no conflicts
		criteria provided single submiter
		criteria provided conflicting interpretations
		no assertion criteria provided
clinvar.structure.alleleID	Clinvar unique identifier for a variant	
clinvar.structure.clinicalSignificance	Significance of a variant in clinvar	Benign
		Pathogenic
		Likely Benign or Pathogenic
		Conflicting interprtations
		Uncertain significance
clinvar.structure.type	Type of variant	Single nucleotide variant
		Insertion
		Deletion
		NT expansion
		Duplication
		Indel
clinvar.structure.phenotypeList	Associated phenotypes for variants at this position	
clinvar.structure.numberSubmitters	Number of submissions in clinvar overlapping this position	
clinvar.structure.origin	Orgin tissue of clinical sample	germline
		somatic
		unknow; not provided
clinvar.structure.referenceAllele	reference allele for this position in clinvar	
clinvar.structure.alternateAllele	alternative alleles for this position in clinvar	
clinvar.structure.reviewStatus	Level of review supporting clinical significance values	Reviewed by expert panel
		criteria provided multiple submiters no conflicts
		criteria provided single submiter

		criteria provided conflicting interpretations
		no assertion criteria provided
refSeq.description	Gene Descritpition from refSeq database	
HGMD_link	link to Human Gene Mutation Database	
genecard_link	link to genecards.org	

References

[1] European Organisation for Rare Diseases. Rare Diseases: Understanding this Public Health Priority. (Eurodis, 2005)

[2] Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018 May;19(5):253-268. doi: 10.1038/nrg.2017.116. Epub 2018 Feb 5. Erratum in: Nat Rev Genet. 2018 Feb 19;: PMID: 29398702.

[3] Shashi, Vandana et al. "The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders." Genetics in medicine : official journal of the American College of Medical Genetics vol. 16,2 (2014): 176-82. doi:10.1038/gim.2013.99

[4] Veltman, Joris A, and Han G Brunner. "De novo mutations in human genetic disease." Nature reviews. Genetics vol. 13,8 565-75. 18 Jul. 2012, doi:10.1038/nrg3241

[5] Lvovs, D et al. "A Polygenic Approach to the Study of Polygenic Diseases." Acta naturae vol. 4,3 (2012): 59-71.

[6] Rahit, K M Tahsin Hassan, and Maja Tarailo-Graovac. "Genetic Modifiers and Rare Mendelian Disease." Genes vol. 11,3 239. 25 Feb. 2020, doi:10.3390/genes11030239

[7] Auton, Adam, et al. "A Global Reference for Human Genetic Variation." Nature, vol. 526, no. 7571, Oct. 2015, pp. 68–74. www.nature.com, https://doi.org/10.1038/nature15393.

[8] Robinson, James T et al. "Variant Review with the Integrative Genomics Viewer." Cancer research vol. 77,21 (2017): e31-e34. doi:10.1158/0008-5472.CAN-17-0337

[9] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics (2011) 27(15) 2156-8

[10] Li, Heng et al. "The Sequence Alignment/Map format and SAMtools." Bioinformatics (Oxford, England) vol. 25,16 (2009): 2078-9. doi:10.1093/bioinformatics/btp352

[11] Thein, Swee Lay et al. "Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults." Proceedings of the National Academy of Sciences of the United States of America vol. 104,27 (2007): 11346-51. doi:10.1073/pnas.0611393104

[12] Mattick, John S, and Igor V Makunin. "Non-coding RNA." Human molecular genetics vol. 15 Spec No 1 (2006): R17-29. doi:10.1093/hmg/ddl046

[13] Kotlar, Alex V et al. “Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale.” Genome biology vol. 19,1 14. 6 Feb. 2018, doi:10.1186/s13059-018-1387-3

[14] McLaren, William et al. “The Ensembl Variant Effect Predictor.” Genome biology vol. 17,1 122. 6 Jun. 2016, doi:10.1186/s13059-016-0974-4

[15] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

[16] Sherry, S T et al. “dbSNP: the NCBI database of genetic variation.” Nucleic acids research vol. 29,1 (2001): 308-11. doi:10.1093/nar/29.1.308

[17] Landrum, Melissa J et al. “ClinVar: public archive of interpretations of clinically relevant variants.” Nucleic acids research vol. 44,D1 (2016): D862-8. doi:10.1093/nar/gkv1222

[18] Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). https://doi.org/10.1038/s41586-020- 2308-7

[19] Seaby, Eleanor G., et al. “Exome Sequencing Explained: A Practical Guide to Its Clinical Application.” Briefings in Functional Genomics, vol. 15, no. 5, Sept. 2016, pp. 374–84. DOI.org (Crossref), https://doi.org/10.1093/bfgp/elv054.

[20] Ewing, B et al. “Base-calling of automated sequencer traces using phred. I. Accuracy assessment.” Genome research vol. 8,3 (1998): 175-85. doi:10.1101/gr.8.3.175

[21] Rentzsch, Philipp et al. “CADD: predicting the deleteriousness of variants throughout the human genome.” Nucleic acids research vol. 47,D1 (2019): D886-D894. doi:10.1093/nar/gky1016

[22] Ramani, Ritika et al. “PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP.” Bioinformatics (Oxford, England) vol. 35,13 (2019): 2320-2322. doi:10.1093/bioinformatics/bty966

[23] Richards, Sue et al. “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.” Genetics in medicine : official journal of the American College of Medical Genetics vol. 17,5 (2015): 405-24. doi:10.1038/gim.2015.30

[24] Koch, Linda. “Exploring Human Genomic Diversity with GnomAD.” Nature Reviews Genetics, vol. 21, no. 8, Aug. 2020, pp. 448–448. www.nature.com, https://doi.org/10.1038/s41576-020-0255-7.

[25] Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny

[26] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[27] Paul Shannon (2022). igvShiny: igvShiny. R package version 1.4.3.

[28] Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013) Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9(8): e1003118. doi:10.1371/journal.pcbi.1003118

[29] Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M (2014). “VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.” _Bioinformatics_, *30*(14), 2076-2078. doi: 10.1093/bioinformatics/btu168 (URL:https://doi.org/10.1093/bioinformatics/btu168).

[30] Yihui Xie, Joe Cheng and Xianying Tan (2022). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.21. https://CRAN.R-project.org/package=DT

[31] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[32] Konrad, Enrico D H et al. “CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum.” Genetics in medicine : official journal of the American College of Medical Genetics vol. 21,12 (2019): 2723-2733. doi:10.1038/s41436-019-0585-z

[33] Courage, Carolina et al. “Novel synonymous and missense variants in FGFR1 causing Hartsfield syndrome.” American journal of medical genetics. Part A vol. 179,12 (2019): 2447-2453. doi:10.1002/ajmg.a.61354

[34] Da Palma, Mariana Matioli et al. “Synonymous Variant in the CHM Gene Causes Aberrant Splicing in Choroideremia.” Investigative ophthalmology & visual science vol. 61,2 (2020): 38. doi:10.1167/iovs.61.2.38