Practical session: identification of disease-causative candidate(s) from a VCF file

Silvia Salatino, PhD - 2018.10.24

Today we will learn how to get from the long list of mutations contained in a Variant Call Format (VCF) file to a much smaller number of putative candidates that might be causal for a given disease or phenotype.

To do this, we will use the approach employed in a study published in 2016 by *Hastings et al.* (**Figure 1**), in which variant prioritisation played an important role in reducing the total number of variants identified from whole-genome sequencing (~ 5 millions) to just a handful of them:

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Format: Abstract

Send to

Circ Cardiovasc Genet. 2016 Oct;9(5):426-435. Epub 2016 Sep 13.

Combination of Whole Genome Sequencing, Linkage, and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy With Features of Left Ventricular Noncompaction.

Hastings R, de Villiers CP, Hooper C, Ormondroyd L, Pagnamenta A, Lise S, Salatino S, Knight SJ, Taylor JC, Thomson KL, Arnold L, Chatzifthimiou SD, Konarev PV, Wilmanns M, Ehler E, Ghisleni A, Gautel M, Blair E, Watkins H, Gehmlich K.

Abstract

BACKGROUND: High throughput next-generation sequencing techniques have made whole genome sequencing accessible in clinical practice; however, the abundance of variation in the human genomes makes the identification of a disease-causing mutation on a background of benign rare variants challenging.

METHODS AND RESULTS: Here we combine whole genome sequencing with linkage analysis in a 3-generation family affected by cardiomyopathy with features of autosomal dominant left ventricular noncompaction cardiomyopathy. A missense mutation in the giant protein titin is the only plausible disease-causing variant that segregates with disease among the 7 surviving affected individuals, with interrogation of the entire genome excluding other potential causes. This A178D missense mutation, affecting a conserved residue in the second immunoglobulin-like domain of titin, was introduced in a bacterially expressed recombinant protein fragment and biophysically characterized in comparison to its wild-type counterpart. Multiple experiments, including size exclusion chromatography, small-angle x ray scattering, and circular dichroism spectroscopy suggest partial unfolding and domain destabilization in the presence of the mutation. Moreover, binding experiments in mammalian cells show that the mutation markedly impairs binding to the titin ligand telethonin.

CONCLUSIONS: Here we present genetic and functional evidence implicating the novel A178D missense mutation in titin as the cause of a highly penetrant familial cardiomyopathy with features of left ventricular noncompaction. This expands the spectrum of titin's roles in cardiomyopathies. It furthermore highlights that rare titin missense variants, currently often ignored or left uninterpreted, should be considered to be relevant for cardiomyopathies and can be identified by the approach presented here.

© 2016 The Authors.

KEYWORDS: cardiomyopathy; left ventricular noncompaction; missense mutation; telethonin; titin; whole genome sequencing

Comment in

Wrestling the Giant: New Approaches for Assessing Titin Variant Pathogenicity. [Circ Cardiovasc Genet. 2016]

Letter by Finsterer and Zarrouk-Mahjoub Regarding Article, "Combination of Whole Genome Sequencing, Linkage, and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy With Features of Left Ventricular Noncompaction". [Circ Cardiovasc Genet. 2016]

PMID: 27625337 PMCID: PMC5068189 DOI: 10.1161/CIRCGENETICS.116.001431

Figure 1

However, since the sequencing data was not made public (due to obvious ethical concerns), we created an

example dataset of 1000 variants, 994 of which are randomly-generated. The remaining 6 variants were taken from Table S2 of the manuscript's supplementary material. Each of the 1000 variants was annotated using VEP and the same set of databases used in the manuscript.
Let's start this practical by downloading the filtering tool we will use for the prioritisation analysis:

1. Download the VCF file named "BGM_course_variants.vcf" from the following website:
http://www.well.ox.ac.uk/bioinformatics/training/BGM_2018/BGM_course_variants.vcf

2. Go to the following webpage https://github.com/BSGOxford/BrowseVCF/releases and download the version of BrowseVCF for your operating system (***Windows users***: "BrowseVCF_win7_v2.8.zip", ***Mac users***: BrowseVCF_osx_v2.8.tar.gz", ***GNU/Linux users***: "BrowseVCF_gnu_2.8.tar.gz")

Depending on your operating system, launch BrowseVCF as follows:

- ***Windows users***: extract BrowseVCF from the .zip bundle, open the folder of BrowseVCF, go to the "web" directory and double-click on "launcher-windows.bat"

- ***Mac users***: extract BrowseVCF from the .tar.gz bundle, open a terminal, go to the "web" directory and type "./launcher-osx.sh"

- ***GNU/Linux users***: extract BrowseVCF from the .tar.gz bundle, open a terminal, go to the "web" directory and type "./launcher-gnu.sh".

This will open a small dialogue box containing two buttons (**Figure 2**). Please keep this window open throughout the whole analysis. For windows users: the launcher will also open a terminal that will run BrowseVCF in the background; please keep that open as well.

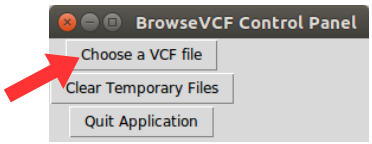


Figure 2

3. Click on the button "Choose a VCF file". This will open a window to navigate through your file system; select the VCF file you just downloaded and click "Open" (**Figure 3**).

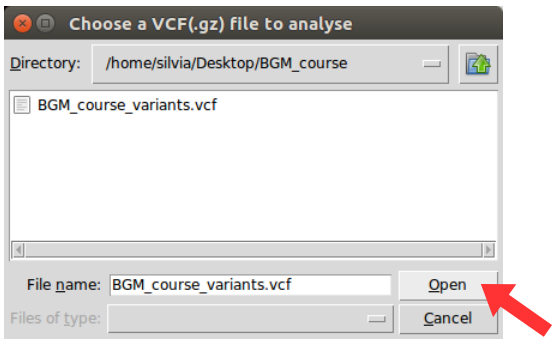


Figure 3

A new tab of BrowseVCF will automatically open on your default web browser. As you can see from the icons at the centre of the page, the analysis consists of four phases, the first of which is to load your VCF file. The sample VCF file should now appear on the blue drop-down box. To use that file, click on "LOAD VCF" (**Figure 4**).

Figure 4

4. You are now at the second phase of the analysis, which will create the wormtables (i.e. the indexes) for your annotation fields of interest. Your working directory will be displayed just below the analysis icons; this is where your intermediate results are stored (**Figure 5**). Please do not delete the content of that folder until the analysis is completed! Select the following annotation field from the blue drop-down menu (you can also search the field name by typing it in the white search box):

- FILTER
- INFO.1000G
- INFO.ESP6500
- INFO.EXAC
- INFO.REPMASK
- INFO.SD
- INFO.UK10K
- INFO.WGS500
- INFO.CSQ_CONSEQUENCE
- SAMPLE1.GT
- SAMPLE2.GT

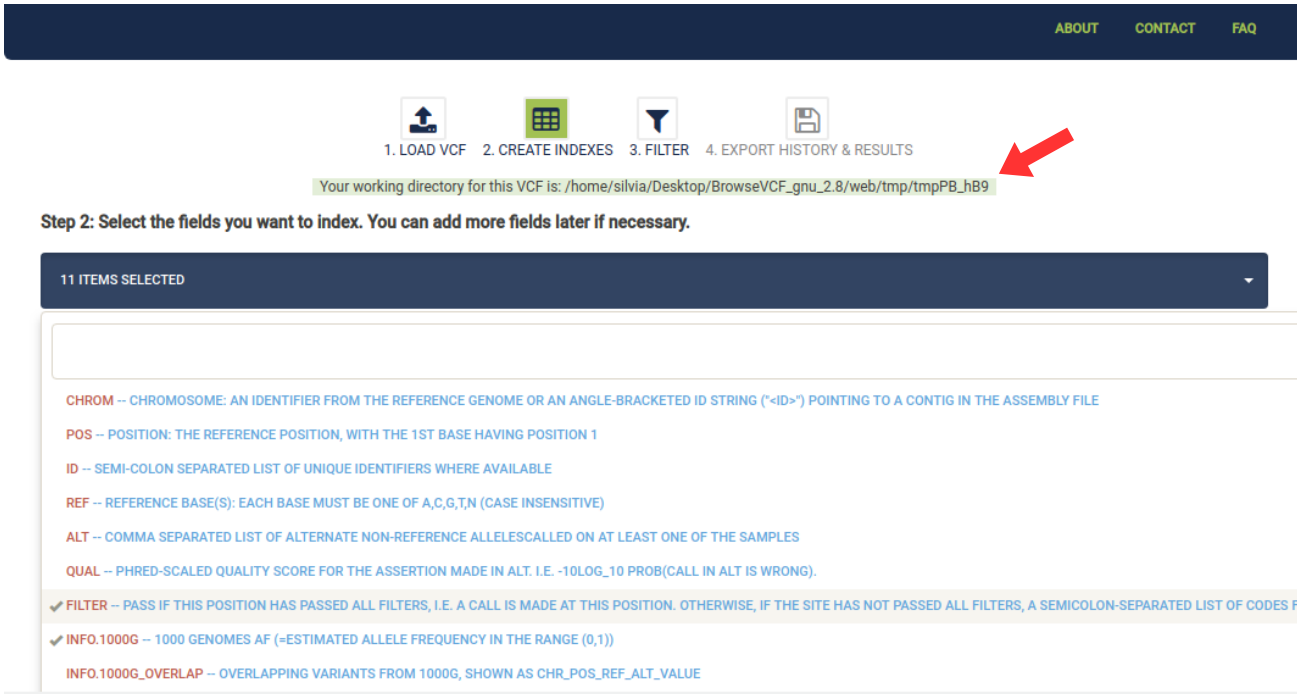


Figure 5

Please note that at the bottom of this page there are links to two of the most commonly used annotation tools (Ensembl VEP and SnpEff), that might help you to choose one or more variant effect or consequence, shown in order of severity. Optionally, you can select how many CPUs to use to run this step. The more you use, the faster the process will be. By default, the program will use all the available CPUs in your computer (**Figure 6**). Simply click on the green button "INDEX SELECTED FIELDS" and wait for the indexing step to complete.

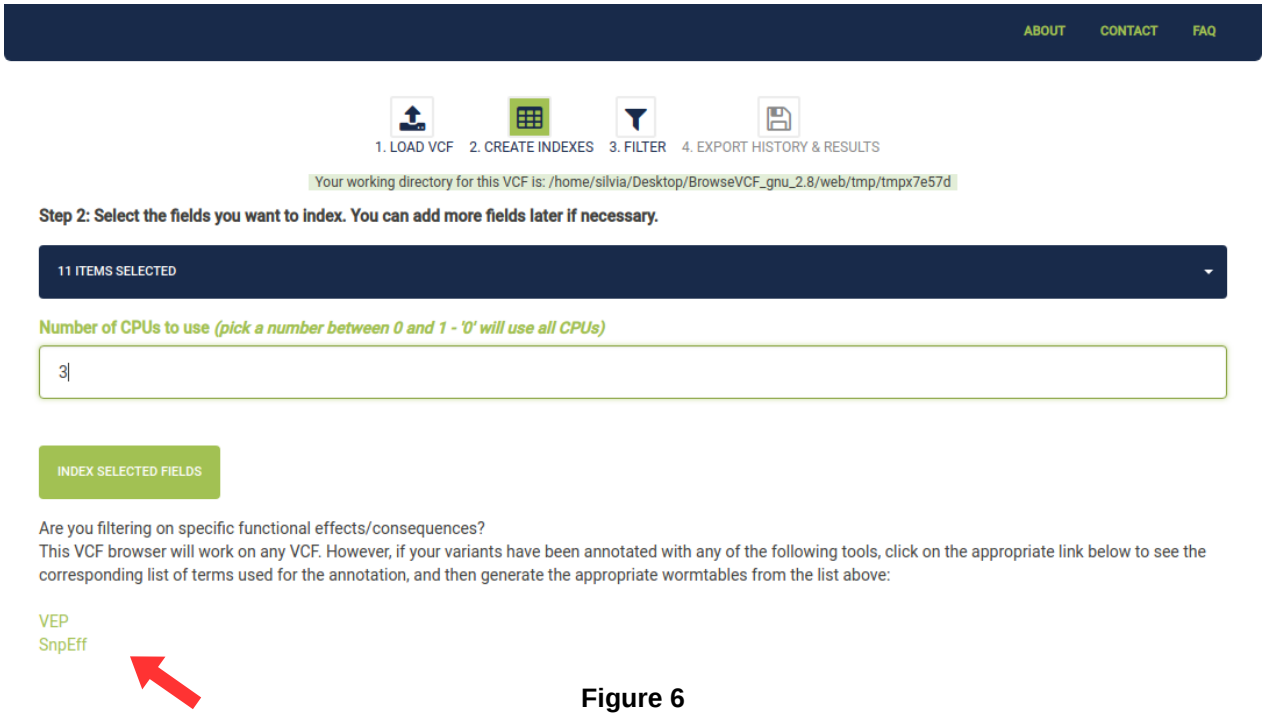


Figure 6

5. You just reached the third phase of the analysis, indicated by the green box labeled “3.FILTER”. On the left panel there are five different type of queries that you can use to filter your variants (**Figure 7**). The “Filter History” section shows you that there were (as we already know) 1000 variants in the input file.

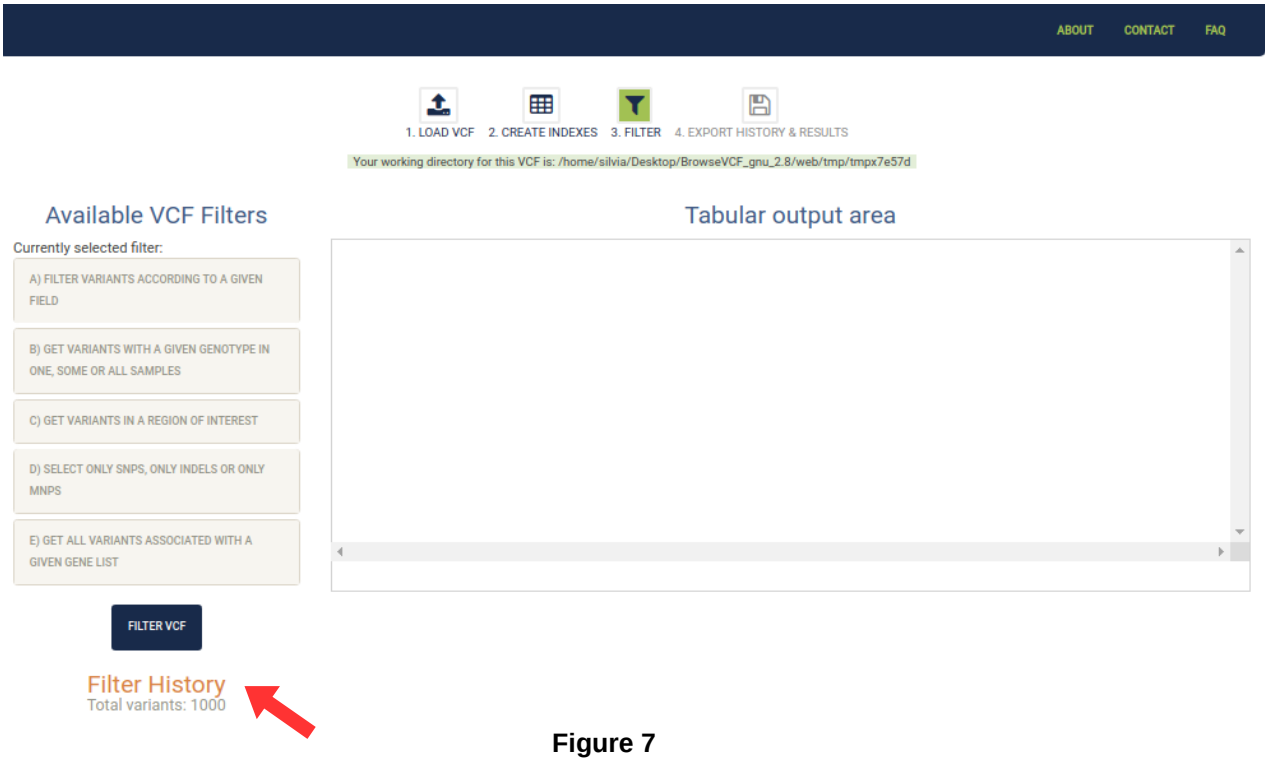


Figure 7

SOME BACKGROUND ABOUT THE STUDY...

As described in the paper, this study focuses on a 3-generational family with history of a particular type of cardiomyopathy, called "***left ventricular noncompaction cardiomyopathy***" (***LVNC***). As indicated in the pedigree below (**Figure 8**), males are depicted by squares, females by circles, deceased individuals by slanted symbols. Affected individuals are marked in grey and the proband is indicated by the black arrow.

The proband (II-3) was a 20-years old male who died suddenly in hospital having presented with rapidly decompensating congestive heart failure. His brother (II-4) was later found to have an enlarged heart with wall thickness at the upper limit of normal and marked hypertrabeculation. The proband's sister (II-2) had a myocardial infarct because of coronary embolus at the age of 61 years and was disgnosed with LVNC. Several screenings identified the same condition in further family members with consistent clinical features of adult onset cardiomyopathy with features of LVNC.

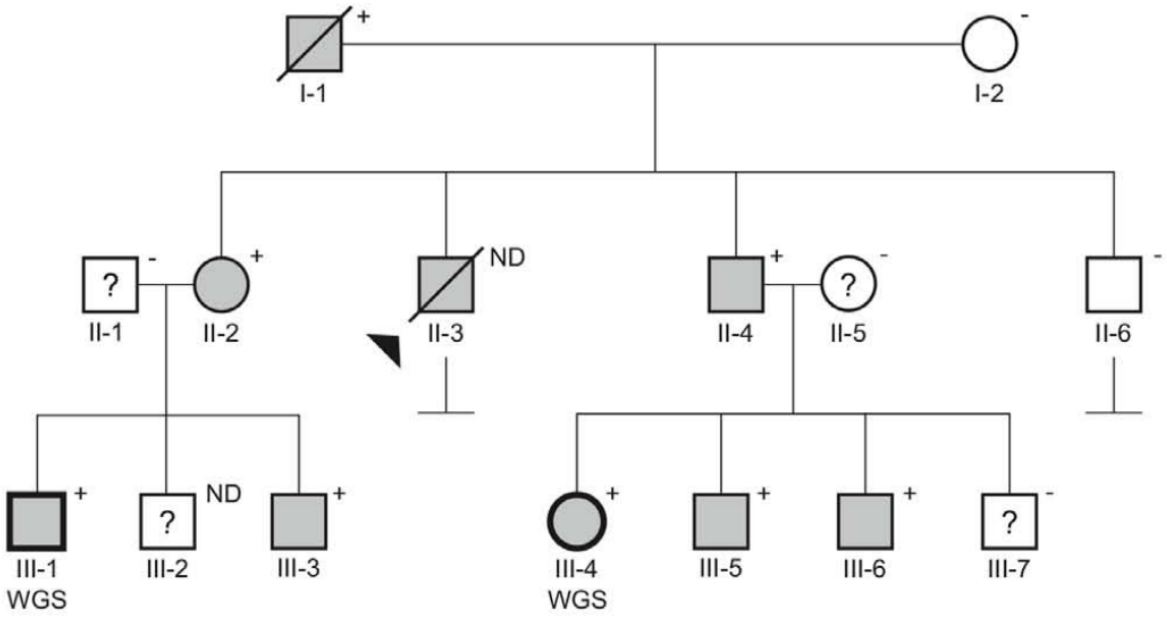


Figure 8

By performing whole-genome sequencing (WGS) in 2 family members (circled in black), filtering against variants seen in normal population cohorts and using linkage information derived from single nucleotide polymorphism (SNP) arrays of 13 family members, researchers could identify a missense variant in the ***titin gene*** (***TTN***) as the most plausible cause of disease in the family. Functional data generated from biophysical and protein-binding experiments on this titin missense variant provided further support of a causative role in cardiomyopathy through domain misfolding and destabilization, resulting in impaired binding to the ligand telethonin (also known as t-cap).

...AND HERE'S OUR GOAL FOR TODAY:

In this practical, we'll try to identify the disease-causative mutation for LVNC by doing sequential filters on the input set of 1000 variants.

Feel free to try applying different filters on the annotation fields defined at step 4 to get to a smaller set of variants, or simply continue reading this tutorial to follow the same steps of the analyst who analysed these data.

6. Usually, it is common practice to filter all variants which are not flagged as “PASS” in the **filter** field by the variant caller. This will excluse any anomalies identified by the software, like for example “badReads” (i.e. variant supported only by reads with low quality bases close to variant position, and not present on both strands) or “hp10” (i.e. “flanking sequence contains homopolymer of length 10 or greater”). However, keep in mind that this is not a fixed rule and can be adapted to different situations. For example, when analysing cancer samples, we tend to keep also all variants flagged as “alleleBias”, because often cancer samples might present aneuploidy and, therefore, an allele bias.

To apply this filter, select the first dropdown menu, you will see the fields that have just been indexed. If you forgot any of them, no problem! You can go back to the indexing step and add more annotation fields at any time point of the analysis. For our first query, select "FILTER" from the top drop-down menu, "equal_to" from the second, and "PASS" from the third. Then, click on the blue button "FILTER VCF" and wait for the results to appear in the right tabular output area. After a few seconds, you will see in the bottom-left corner of the page a light-green box under the "Filter History" panel, which will indicate you the filter applied and the resulting number of variants, that is 991 (**Figure 7**). On the right panel the first 100 variants left after applying this filter will appear. Usually this filter removes a big portion of the reads, but in this case the “random” variants contained in the input VCF file were generated to have ~90% PASS values.

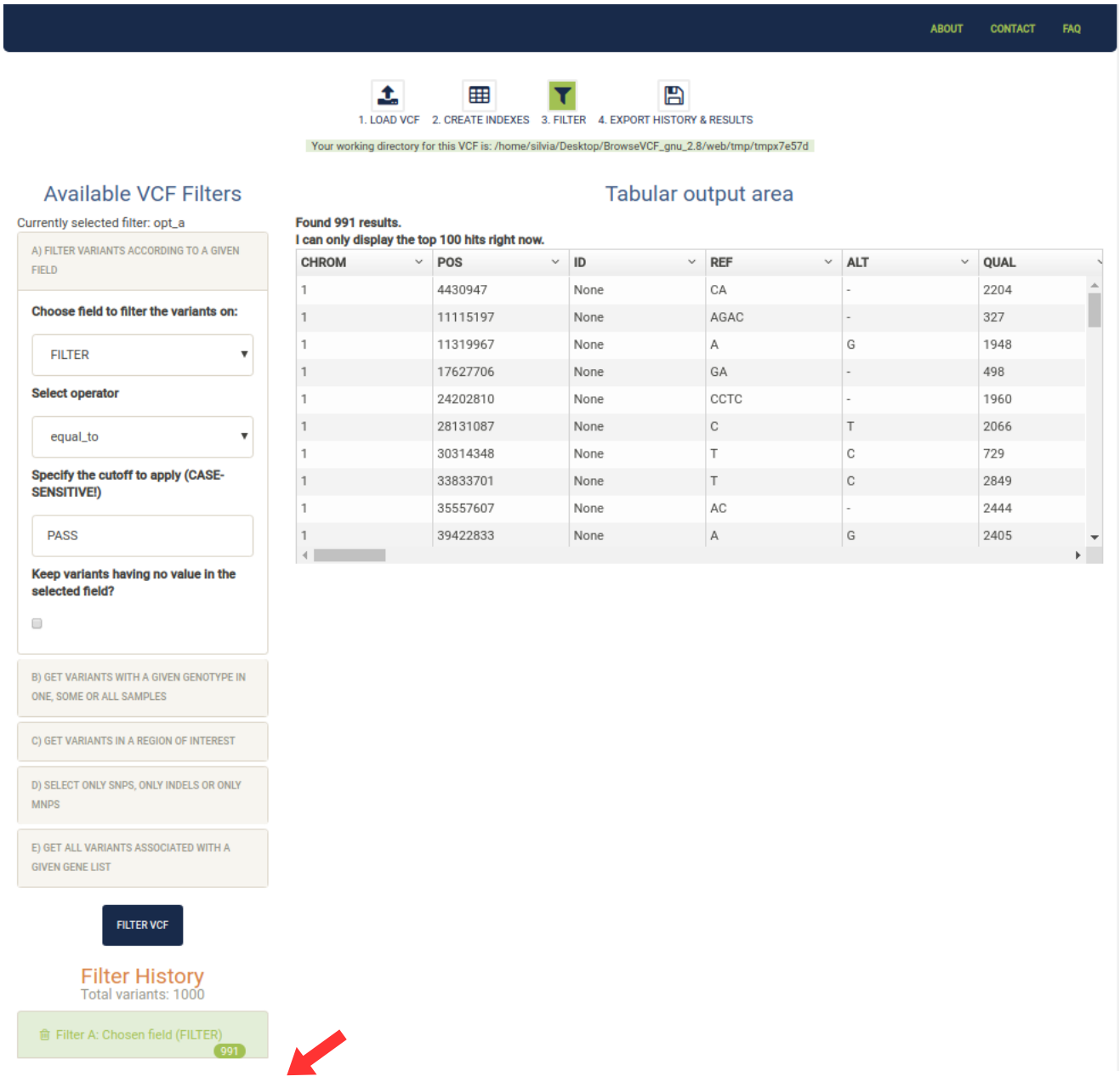


Figure 9

7. Another filter commonly applied is on **segmental duplications** and **repeats**. This is because, particularly with Illumina sequencers, base calling is not reliable in these genomic segments. Using once more filter A, choose “INFO.SD” from the drop-down menu and the operator “is_absent”. Then click on the “FILTER VCF” blue button and you’ll see our variants reducing to 929 (**Figure 10**).

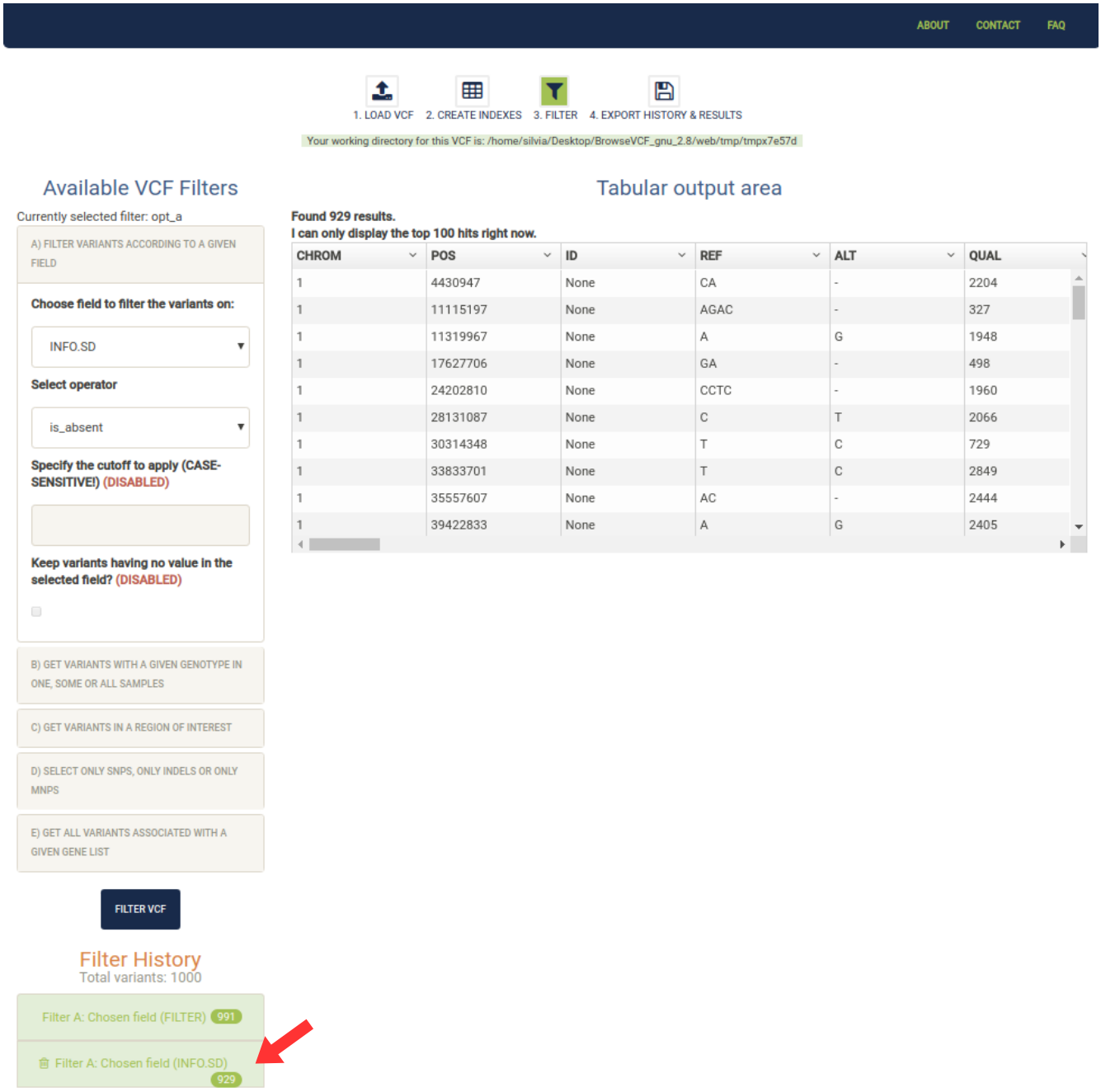


Figure 10

Doing the same for the “INFO.REPMASK” annotation, which reports overlap of the variants with any type of repeats (e.g. LINE, LTR, etc.) will result in 439 variants left (**Figure 11**)

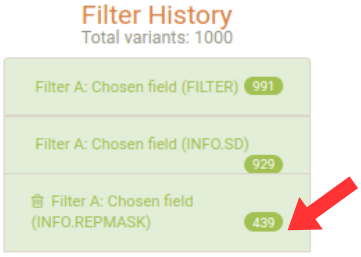


Figure 11

8. The next step is, generally, to discard variants found too often in **large cohort projects**, i.e. with a low allele frequency. Setting this cut-off is arbitrary, although a commonly used threshold is 1%. Select again filter A and choose “INFO.1000G” from the drop-down menu. The operator must be “less_than” and the cut-off “0.01”. Tick the box to keep variants having no annotation for the selected field (because some variant could not be present at all in the 1000 Genomes project and, therefore, have no annotation for the selected field) and click the "FILTER VCF" button once more. The number of variants passing this second filter has reduced only to 438, as you can see from the "Filter History" panel (**Figure 12**).

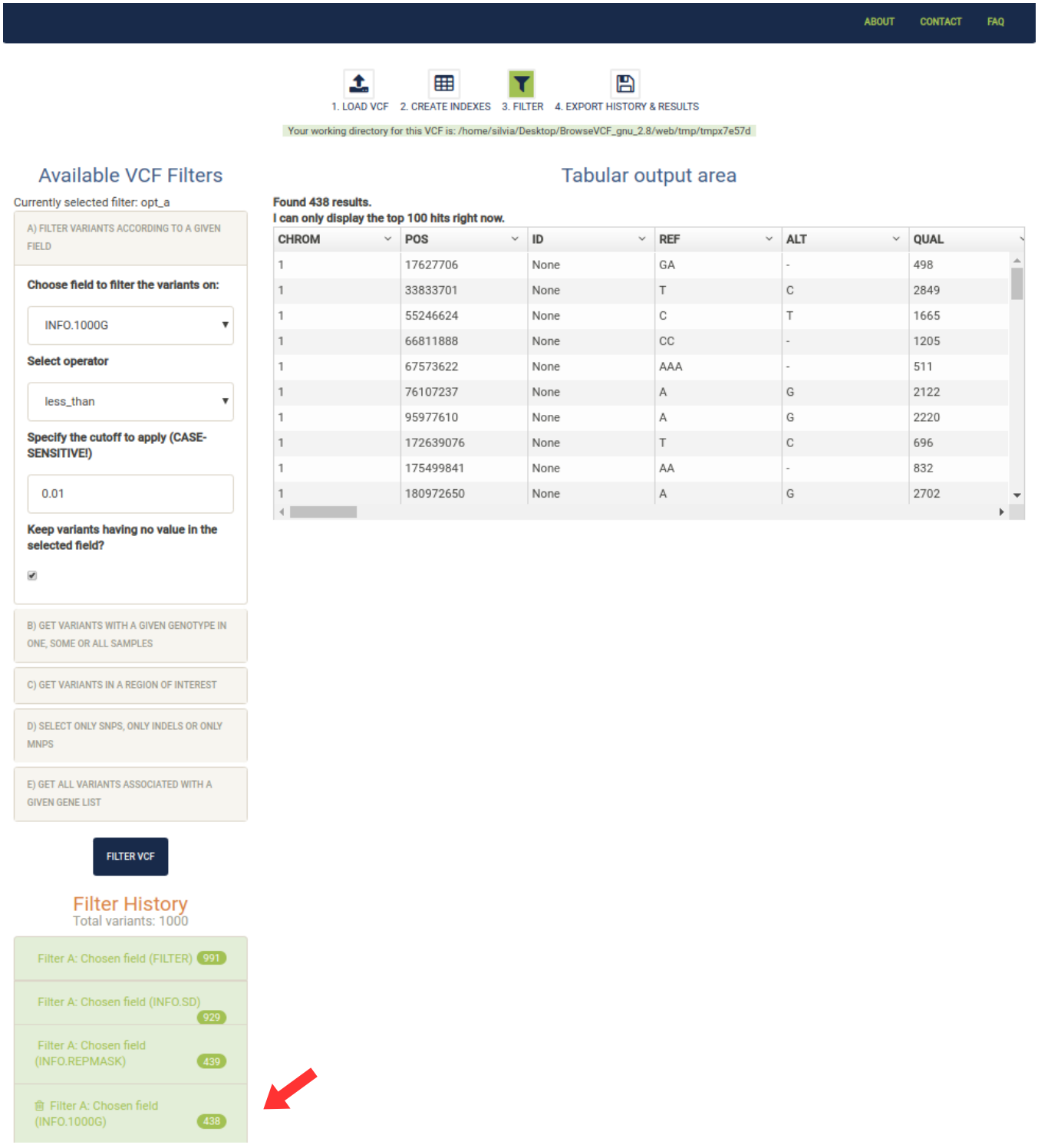


Figure 12

Applying the same filter on other large genomic projects, like the Exome Sequencing Project 6500

(corresponding annotation field “INFO.ESP6500”), the WGS500 Project (corresponding annotation field “INFO.WGS500”), the UK10K Project (corresponding annotation field “INFO.UK10K”), or the Exome Aggregation Consortium (corresponding annotation field “INFO.ExAC”), will return the same number of variants. This is somehow expected, as it is very common to find (or not to find) the same variants across these major genomic projects.
However, if you want to give them a try, since -in this case- the allele count was reported rather than the allele frequency, you might need to use a cut-off expressed in integers instead of decimals.

9. The next step could be filtering variants on the basis of sample genotypes. Since LVNC is inherited and transmitted in **autosomal dominant** patterns, and since the two sequenced individuals were affected, we will select only mutations that are shared between the two cousins and **heterozygous**. To do this, use filter B and select “Heterozygous” from the drop-down menu. Since we want both samples to be heterozygous, select both “SAMPLE1” and “SAMPLE2” from the second drop-down menu and launch the filter. As you will notice from the Filter History panel, there are now only 112 variants left (**Figure 13**), although that’s still quite a lot to manually look at each of them.

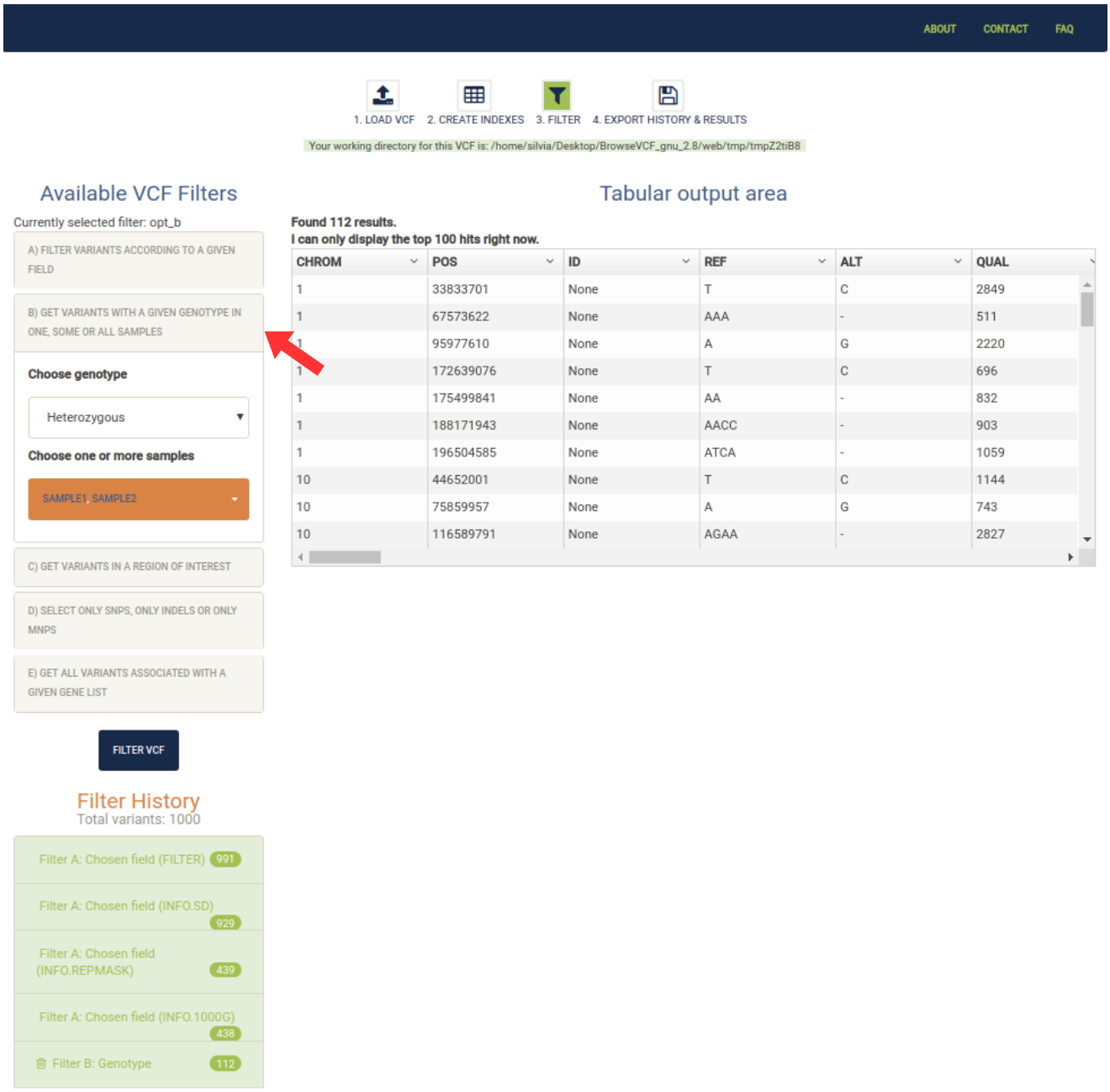


Figure 13

10. How can we reduce the number of variants further? An important filter to apply is the predicted variant consequence or **deleteriousness**. This information was obtained by annotating variants with the Ensembl tool Variant Effect Predictor (VEP) and can be found in the INFO.CSQ_Consequence field. Select this field from filter A together with the operator "contains_keyword". If you open a new tab in your internet browser and type the following website address, you'll see all the possible consequence effects that can be predicted by VEP: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

The "IMPACT" column of the table on that website gives you an indication of each sequence ontology term deleteriousness. In this step, we will use the following terms; please add them to the cut-off field in BrowseVCF, taking care of not introducing white spaces:

transcript_ablation,splice_donor_variant,splice_acceptor_variant,stop_gained,frameshift_variant,stop_lost,initiator_codon_variant,transcript_amplification,inframe_insertion,inframe_deletion,missense_variant,splice_region_variant

This query should return 7 variants, located on four different chromosomes (**Figure 14**).

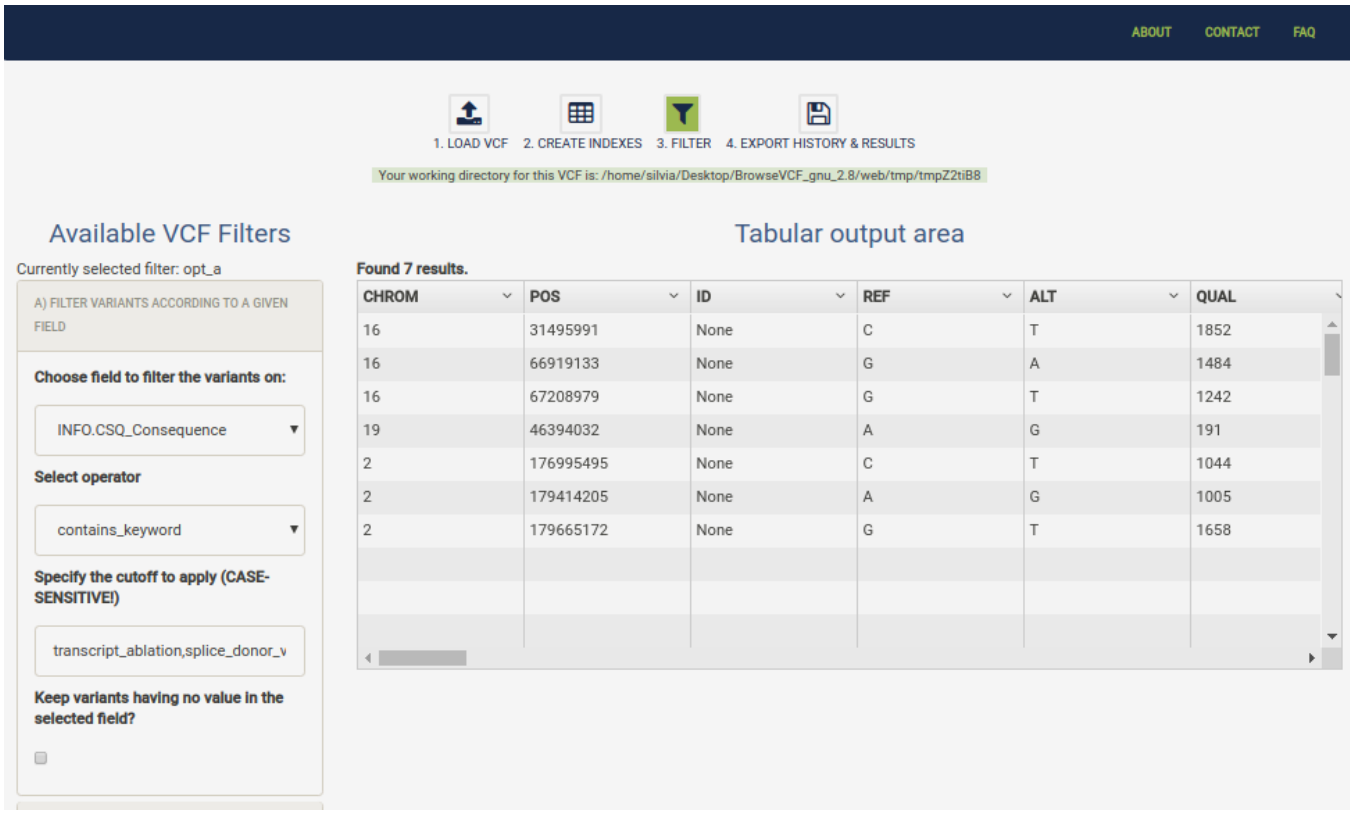


Figure 14

11. Thankfully, this study also included SNP array genotyping, which was performed on all family members (excluding II-3 and III-2) using nearly 300000 genetic markers. A refined subset of roughly 24000 SNPs in approximate linkage equilibrium was generated using the software PLINK and data from HapMap. **Linkage analysis** of the SNP subset was performed using MERLIN and specifying an autosomal dominant disease model. Genomic intervals with logarithm of the odds (LOD) scores >0, compatible with segregation of variants in these regions, were selected for downstream analyses. From Figure S2 of the supplementary material we can see the three intervals that were identified on chromosomes 2, 9, and 16 (**Figure 14**).

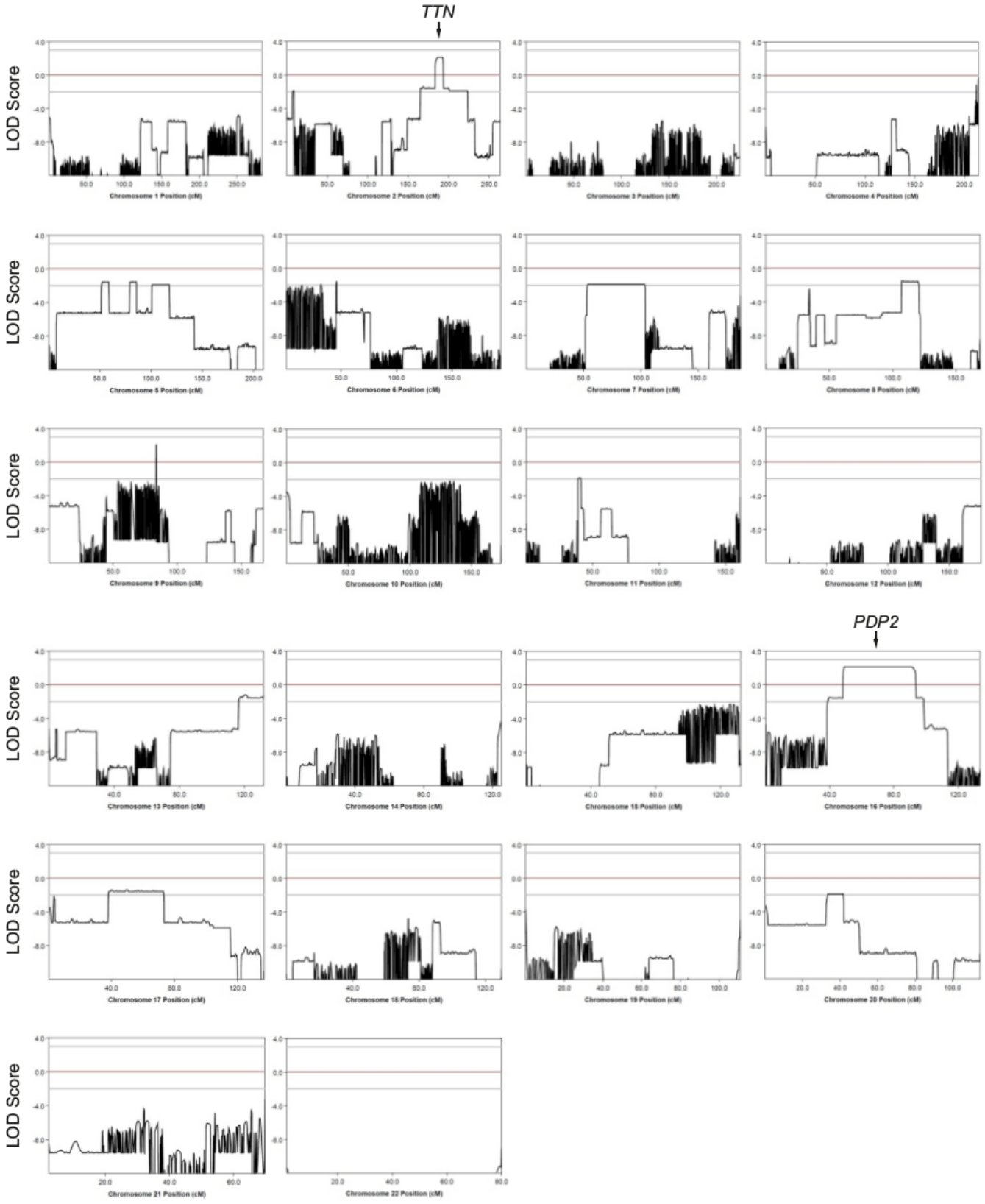


Figure 14

Let's use just one of these intervals for the scope of this practical. Use filter C by typing chromosome 2 and the (approximate) coordinates 156 000 000 – 193 000 000, as in Figure 15. You should have only 3 variants now.

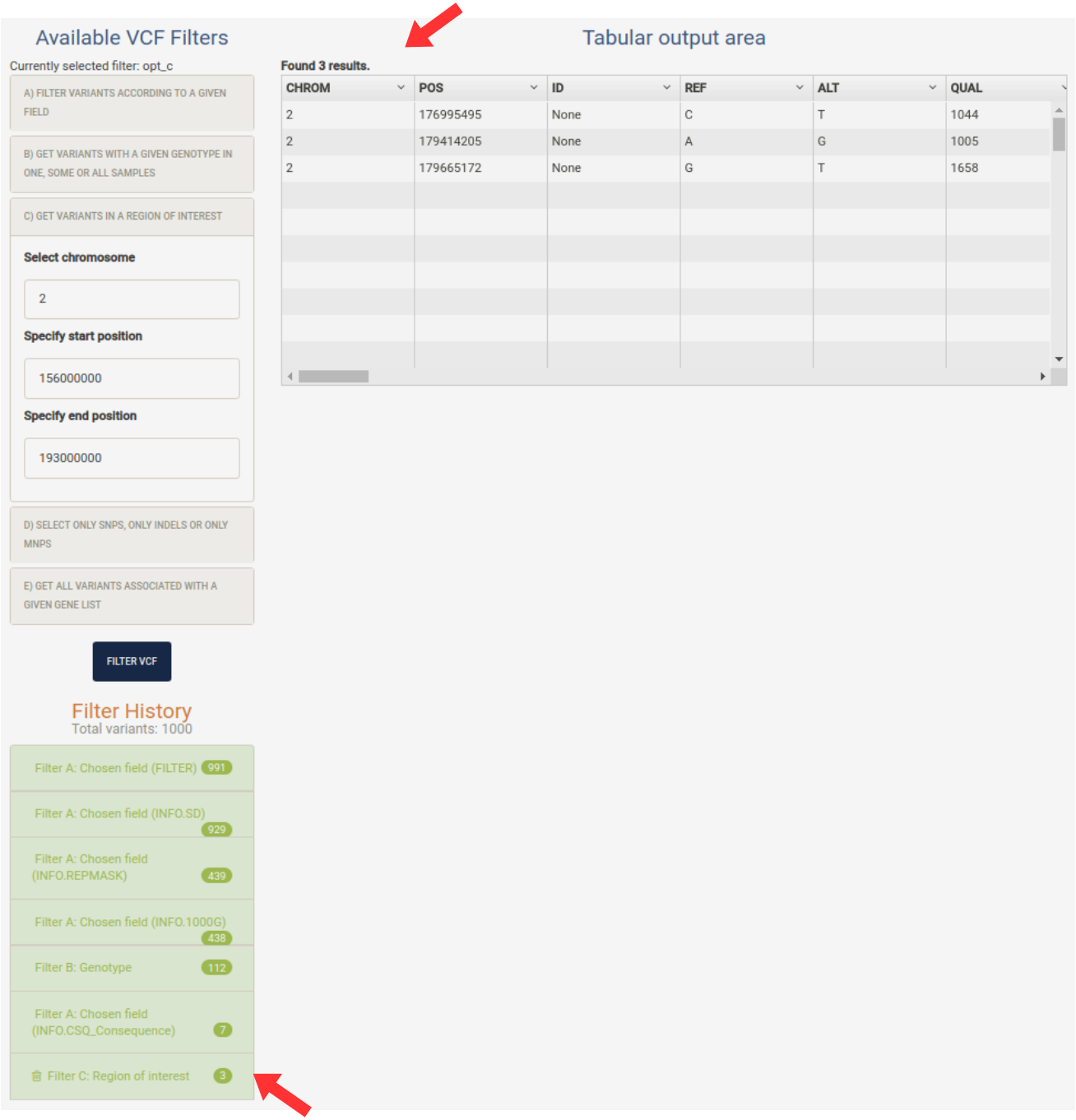


Figure 15

Congratulations!
This is finally a good number of variants to have a closer look on. Now that we're happy with our results, we can proceed to the fourth and last functionality of BrowseVCF, that will allow us to **export the history of our queries and the final set of variants** that passed all the filters (**Figure 16**). Click on the button “4. EXPORT HISTORY & RESULTS” at the top of the page. Click on the green button on the left to export the detailed history of your queries in plain text format. Click on the blue button on the right to save the final results as tab-separated format (compatible with any spreadsheet). A pop-up windows will ask you to choose a name and a destination for the output file (e.g. “My_filtered_variants.xls” on the Desktop).

One last important thing to do: **clean up**!
To free the disk space taken by all the intermediate files generated by BrowseVCF (which, for an annotated whole-genome, could be ~10Gb!), once you have downloaded the final results, simply delete the temporary folder highlighted in green in the figure above and that's it, you're done.

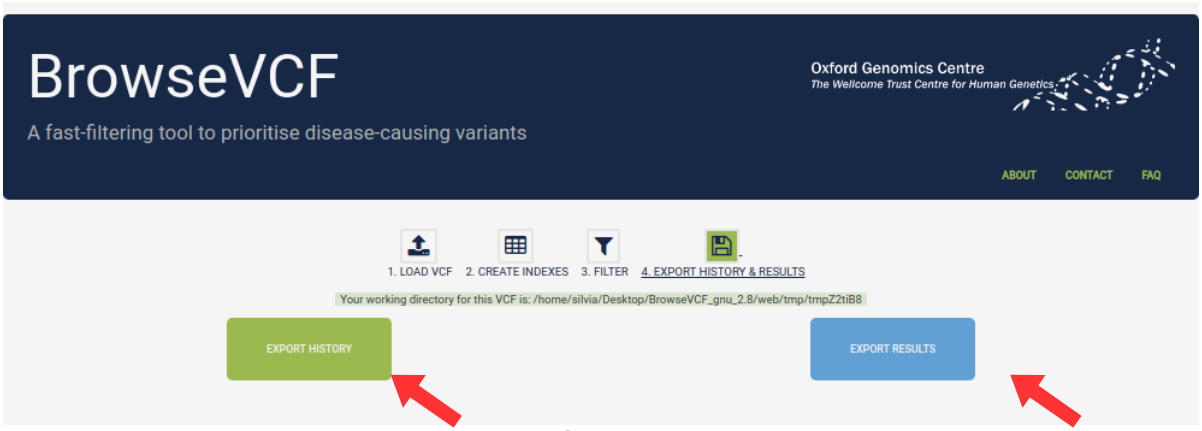


Figure 16

Follow-up

So... What's the end of the story from Hastings' paper? As we can see from Table S1 in the supplementary material, the set of filtering criteria applied was more or less the same we did using BrowseVCF (**Figure 17**).

Table S1: Filtering criteria applied to all variants identified.

1. Within linkage region (LOD > 0), whereby both affected and unaffected individuals were considered in the linkage analysis

2. Exonic or splicing variant,
   i.e. with one of the following consequence terms: 'transcript_ablation', 'splice_donor_variant', 'splice_acceptor_variant', 'stop_gained', 'frameshift_variant', 'stop_lost', 'initiator_codon_variant', 'transcript_amplification', 'inframe_insertion', 'inframe_deletion', 'missense_variant', 'splice_region_variant'

3. Heterozygous and shared by both affected individuals (III-1 and III-4)

4. Called confidently by Platypus (flagged as 'PASS')

5. Overlapping neither segmental duplications nor repeats

6. Allele frequency in 1000 Genomes ≤ 1 % or not reported [1]

7. Observed no more than 7 times in WGS500 [2]

8. Observed no more than 50 times in ESP
   (Exome Variant Server, http://evs.gs.washington.edu/EVS/)

9. Observed no more than 50 times in UK10K
   (UK10K Project, http://www.uk10k.org)

10. Observed no more than 500 times in ExAC Browser
    (Exome Aggregation Consortium, http://exac.broadinstitute.org/)

Steps 1 to 10 are implemented in an automated script, further filtering steps are based on manual inspection:

11. Supporting evidence for the existence of affected transcript(s)

12. Evidence of expression of the gene in the heart both at RNA and protein level using multiple databases (see Expanded Materials)

13. Splice/intronic variants: considered if at crucial position (-2 to +2) or violating consensus rules at position -6 to -3 (for 5' sites) or at position -3 for 3' sites [3]

    Predicted to be tolerated using MaxEntScan
    (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html and
    http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html)

Figure 17

However, after getting to a set of 6 putative candidates, researchers had to manually inspect them using different approaches: (i) one variant was excluded because it was assumed to be an artifact, due to an incorrect transcript being present in Ensembl, (ii) another variant did not segregate with disease in the family, (iii) 2 splice variants were predicted by MaxEntScan to be silent (at positions -5 and -3 of a 3′ splice junction, respectively). Only 2 final candidate variants were considered conceivably linked to the phenotype: a missense change in PDP2 and TTN, respectively. However, PDP2 codes for a protein which has low expression levels in the heart, whereas **TTN** codes for titin, an abundant skeletal muscle and heart-specific protein with crucial functions. Mutations in titin have been associated with cardiomyopathy and skeletal myopathy. The identified missense variant c.533C>A in TTN, which codes for a **p.A178D** change (i.e. from alanin to aspartic acid) at the amino acid level, is absent in ExAC (**Figure 18**).

Table S2: Variants remaining after Platypus filtering (steps 1-10 of Table S1)

											Allele frequency				
CHROMOSOME	POSITION	REFERENCE	ALTERATION	QUALITY	FILTER	GENE	CONSEQUENCE	AA_CHANGE	SIFT	POLYPHEN	1000G	UK10K(AC/AN)	ESP6500(AC/AN)	EXAC	Reason for exclusion
16	66919133	G	A	1484	PASS	*PDP2*	missense_variant	E316K	deleterious(0)	probably_damaging(1)	0	1.3E-04	0	4.9E-05	n/a
2	179665172	G	T	1658	PASS	*TTN*	missense_variant	A178D	deleterious(0)	possibly damaging (0.734)	0	0	0	0	n/a
16	31495991	C	T	1852	PASS	*SLC5A2*	splice_region_variant & intron_variant				0	0	0	8.2E-06	expressed exclusively in kidney and testis; position -3 of a 3' splice junction, predicted to be tolerated
16	67208979	G	T	1242	PASS	*NOL3*	missense_variant	G45V		unknown(0)	0	0	0	2.5E-05	an artefact due to an incorrect, poorly supported transcript (ENST00000564860) present in Ensembl; synonymous R213R change in all other transcripts
2	176995495	C	T	1044	PASS	*HOXD8*	missense_variant	A134V	tolerated(0.08)	benign(0.31)	0	0	0	0	not present in affected individual III-6
2	179414205	A	G	1005	PASS	*TTN*	splice_region_variant & intron_variant				0	0	0	0	position -5 of a 3' splice junction, predicted to be tolerated (for detailed analysis see Table S3)

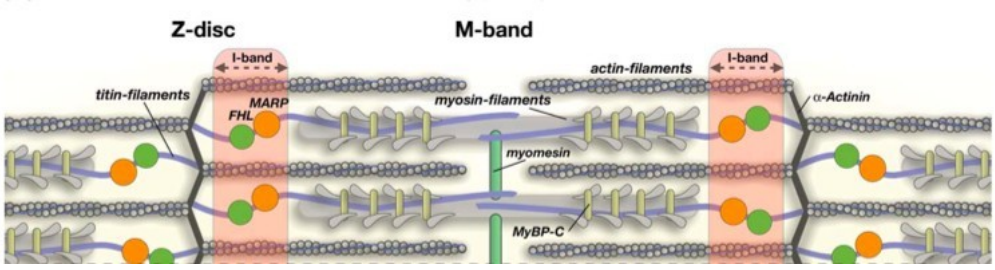
RNA and protein expression in the heart

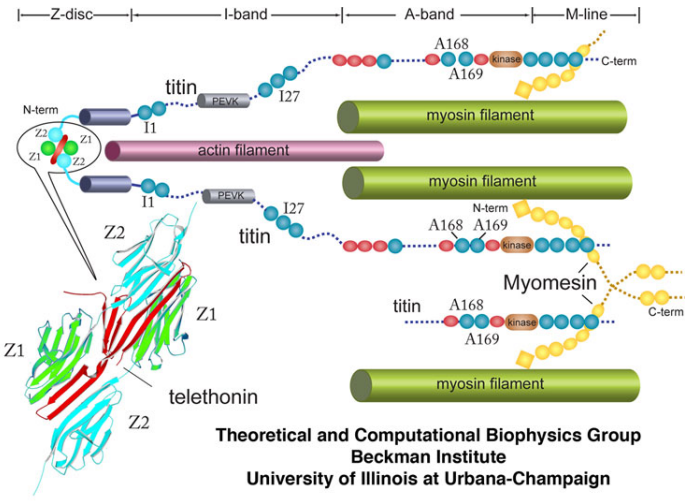
GENE	RNA					Protein					Comment
	GeneCards	Expression Atlas (EMBL)	Protein Atlas	GeneHub	GTex	Protein Atlas	Human Protein Map	Proteomics DB	PaxDB	GeneCards	
TTN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*known cardiomyopathy gene*
PDP2	✓	✓	✓	×	✓ (low)	✓	×	×	×	×	*low expression in the heart*

Figure 18

Titin molecular function

Acting like a molecular bungee cord, ***titin*** protects muscle fibers from damage due to overstretching. It binds to ***telethonin***, which is believed to anchor the ends of two separate titin molecules to the Z-disc. In individuals where this binding is impaired, pathological states arise.





Validation

Sanger sequencing confirmed the cosegregation of the heterozygous mutation with disease in all affected individuals of the family (see “+” and “-” signs in the family pedigree). In addition, researchers performed a number of other validations, including:

- structural biology prediction → the charged aminoacid is likely to impact the protein’s secondary structure (due to steric hindrance) and probably its folding too
- bacterial expression of WT and mutant protein → circular dichroism spectroscopy and x-ray scattering showed that the β-sheet conformation was impaired and that the mutant protein was unfolded
- denaturating gel electrophoresis → reduced stability and a degradation product were observed for the mutated protein, but not for the WT one
- semiquantitative GST pulldown assay → the mutated protein showed impaired binding to two telethonin constructs

...That’s all, folks!

Hope you enjoyed this practical session and found it useful for your future studies.

Thanks for your attention,
Silvia