Variant prioritization in NGS studies: Annotation and Filtering (Practical)

Work through the following practical exercises on your own.

The objective of these exercises is to become familiar with the concepts of variant filtering in order for you to develop your own filtering strategy.
• Genome in a Bottle consortium set of high-confidence variant calls from WGS

https://sites.stanford.edu/abms/giab

• Reference genome: GRCh37
• Extracted Agilent V5+UTR target regions to simulate whole exome sequencing study
• Used VCFrandomSample in Galaxy to randomly select 5000 variants (~10%)

• Sample ref: NA12878
• Female
• Utah/European ancestry
DISCLAIMER: I am not a clinician nor a medical geneticist.
The exercises in the following 2 sessions are designed as a conceptual thought process. They should not be interpreted as any kind of clinical guideline or protocol!

You have a female patient who suffers from early onset and recurring deep vein thrombosis. She has a history of unexplained pregnancy loss and a family history of venous thromboembolism. You suspect this follows an autosomal dominant inheritance pattern. You have opted for clinical whole exome sequencing to identify a potential molecular diagnosis.

After processing, the resultant variant calls are in 2_GIAB_GalaxyRandom.vcf.
Go to the ANNOVAR web application: [http://wannovar.usc.edu/](http://wannovar.usc.edu/)

- Have a look at the input options on the submission form.
- Click through to the tutorial tab on the homepage.
- Read through the documentation until you understand the different options you can select during the submission.
Go back to the homepage. We will now submit our sample .vcf for annotation.

Fill in the Basic Information – email address & sample identifier

Click +Input File to navigate to your .vcf

Note: wANNOVAR input must be a .vcf or manual entry of variants in the same format!

We will use default parameters for this run, but make sure the correct reference genome is selected!

Click Submit
You will be taken to a submission confirmation

Click through to the results page and wait... (You will receive an email with the link to the results page once the computation is done)

Tip for Mac users: Check that your user input contains the expected number of lines!

The computation takes ~10 minutes depending on the server load. (If yours is taking too long, ask me for the annotated file!)

While you are waiting, open the ANNOVAR documentation in a new tab/window and read through the filtering descriptions to understand the output file.

The results page will refresh until the computation is done.

Check the results summary. Does this look right?

Our dataset is a simulated WES data set, however it includes UTR variants. If your study design includes ONLY coding variants you can use the “exome summary results”, otherwise use the “genome summary results”.

Right click “TXT file” and save the file

Right click “view”, open in a new tab & go to it
The (~5000) annotated variants will be clearly displayed.

Scroll down to the filtering options below.

Toggle the fields to filter for exonic and UTR variants on chromosome 6 with a MAF less than or equal to 0,05 in 1000g.

Click “Go”

Scroll up to have a look at your filtered variants
• Have a look at the chromosome column – did the filter work?
• How many non-synonymous variants are there?
• How different are the MAF’s across the different populations in 1000g?
• Scroll right to the end – is this individual homozygous for any of these variants?

• Navigate to where you saved the wANNOVAR output .txt file
• Right click and open with Excel

Note: This is a simulated data set consisting of only ~5000 variants and can therefore be feasibly manipulated in Excel. Large WGS datasets consist of ~1-1,5 million variants, and WES datasets consist of ~50 000 variants. This makes the annotated files very difficult to impossible to work with in Excel! You (or your bioinformatician) will need to do the initial filtering steps using the command line or some other VCF manipulation tools.
SIDE NOTE: Working with Filters in Excel

To add filters to your spreadsheet:
• Select the header row of your spreadsheet
• Go to “Data” and click “Filter”

To clear ONE filter
• Click on the filter arrow & click “Clear Filter”

To clear ALL filters
• Go to “Data” and click “Clear Filters”
SIDE NOTE: Working with Filters in Excel

If you try to select and copy a section of the spreadsheet, after you have applied filters – it will still copy all the hidden data!

To copy only your filtered data to a new spreadsheet or file:
- Select the area you want to copy
- Specify visible data only
  - Alt + ; (Windows)
  - Cmd + Shift + Z (Mac)
- Then copy (Ctrl C) and paste (Ctrl V) as usual

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Ref</th>
<th>Alt</th>
<th>Func.ref</th>
<th>Gene.ref</th>
<th>Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38049552</td>
<td>38049552</td>
<td>C</td>
<td>G</td>
<td>intronic</td>
<td>GNL2</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>51820990</td>
<td>51820990</td>
<td>T</td>
<td>G</td>
<td>UTR3</td>
<td>EPS15</td>
<td>NM</td>
</tr>
<tr>
<td>1356</td>
<td>170987636</td>
<td>170987636</td>
<td>A</td>
<td>G</td>
<td>intronic</td>
<td>AADAT</td>
<td></td>
</tr>
<tr>
<td>1370</td>
<td>16081523</td>
<td>16081523</td>
<td>A</td>
<td>-</td>
<td>UTR3</td>
<td>CFA987</td>
<td>NM</td>
</tr>
<tr>
<td>1113</td>
<td>129929673</td>
<td>129929673</td>
<td>C</td>
<td>T</td>
<td>downstream</td>
<td>CPA2</td>
<td></td>
</tr>
<tr>
<td>1173</td>
<td>150742985</td>
<td>150742985</td>
<td>G</td>
<td>A</td>
<td>UTR3</td>
<td>ABC89</td>
<td>NM</td>
</tr>
<tr>
<td>1291</td>
<td>93520799</td>
<td>95206799</td>
<td>G</td>
<td>T</td>
<td>intronic</td>
<td>CD1H7</td>
<td></td>
</tr>
<tr>
<td>2442</td>
<td>93373099</td>
<td>93373099</td>
<td>A</td>
<td>-</td>
<td>UTR3</td>
<td>DIRAS2</td>
<td>NM</td>
</tr>
<tr>
<td>1128</td>
<td>8995756</td>
<td>8995756</td>
<td>A</td>
<td>G</td>
<td>exonic</td>
<td>A2ML1</td>
<td></td>
</tr>
<tr>
<td>1127</td>
<td>9202912</td>
<td>9202912</td>
<td>A</td>
<td>G</td>
<td>exonic</td>
<td>A2ML1</td>
<td></td>
</tr>
<tr>
<td>1130</td>
<td>9203038</td>
<td>9203038</td>
<td>C</td>
<td>T</td>
<td>intronic</td>
<td>A2M</td>
<td></td>
</tr>
</tbody>
</table>
• Apply Filters to your annotated variants file
• Select the columns containing MAF information (K-AC) and replace the “.” with “,” so that excel treats these as numbers

For diseases where you expect one or two variants with large effect sizes, it is a good idea to look for obvious variants first (“low-hanging fruit”!)

Using the filters, check for variants that have previously been implicated in disease or GWAS.
• Using the GWAS_DIS filter, filter for variants that have been implicated in a GWAS by unselecting “.” and “Blanks” (~30 variants)
• Reading through these diseases, do any variants seem interesting in this case?
• Copy the interesting ones to another file (incl. which gene they are in!)
• Clear this filter
• Now repeat, using the CLINVAR_DIS filter and unselecting “.” and “Blanks” (~163 variants)
• That’s still quite a lot, how would you filter further?
Using the CLINVAR_SIG column, filter against variants marked as non-pathogenic, other, untested, unknown & probable-non-pathogenic so only pathogenic variants remain. (~10 variants)

Reading through these diseases, do any variants seem interesting in this case?

Copy the interesting ones to another file

Hang on – there’s 1 variant there that looks worrying but is unrelated to our phenotype of interest. Should we be worried about it? Should we inform our patient?

Clear both CLINVAR filters

This “low-hanging fruit” approach rarely yields results when investigating common or multi-factorial conditions where you are expecting a number of variants with moderate effect sizes.
Filtering on **inheritance pattern** would be a good place to start. The genotypes are in the last column of the spreadsheet

- Which genotype (0/1 or 1/1) would you filter against in this case?

Let’s filter our variants based on their **genomic context**.

- Use “Func.refgene” to filter out intronic & intergenic variants (~3208 vars)
- Now use “ExonicFunc.refgene” to filter out synonymous exonic variants. (~2449 vars - Be careful not to filter against “Blanks” otherwise you will exclude all UTR and splicing variants!)

We’ve filtered out 50% of our variants but 2500 is still a lot to manually inspect! Let’s filter out the **common variants**.

- Looking at our spreadsheet, most variants have 1000g MAF data
- Which 1000g column would you filter on?
- What frequency cut-off would you use - 1%? 5%? 10?
- Try 5% to start with *(Remember to use both ends of the frequency spectrum!)*
You should have ~364 variants left – much more manageable but still quite a lot to manually investigate. Let's see how many of them are predicted to be functional.

- Use SIFT, PolyPhen-HDIV, FATHMM and MutationTaster columns to filter for variants predicted to be functional by at least one of these algorithms
- Remember to clear that algorithms filter in between each one
- At each stage, copy your interesting variants across to another file
- Ordinarily you would use all of the functional prediction algorithms, but this requires advanced filtering or scripting

You should now have a set of variants (hopefully ~100) that are interesting candidates.

One more step before tea-time!
Most of the functional prediction tools included in the wANNOVAR annotation don’t provide scores or predictions for **non-coding variants**.

- Use the “Func.refgene” filter to filter out the exonic variants
- Pick any 10 variants and manually run them through the FATHMM non-coding algorithm to check for predicted functionality
- If any interesting ones come up, add them to your new file
- Ordinarily you would do this for all of the non-coding variants!

Sort your new file of interesting variants and do a quick check to get rid of duplicates (there will definitely be a few!).