Workshop Part II Exome variant interpretation

Christian Gilissen (Christian.Gilissen@radboudumc.nl)
Galuh Astuti (Galuh.Astuti@radboudumc.nl)
Department of Human Genetics, Radboud UMC

READ ME FIRST

- Actions you need to perform are indicated in **bold**.
- Questions are designated by "**Q**" and are in *italics*.
- Screenshots are just examples, your specific output may look slightly different.
- At the end of this manual there are quick reference guides attached that can help you during this workshop.
- For this workshop you require:
    - This manual
    - Files in following folders:

```
2.Exome_variant_interpretation
                        |--- data
                             |--- vcf_file
                             |--- variant_interpretation
                                                  |--- Case1
                                                  |--- Case2
                                                  |--- Case3
                                                  |--- Case4
                                                  |--- Case5
```

    - Software:
        - Microsoft Excel
        - Internet
- You can ask questions during the workshop!

Contents

Introduction

From the first part of the workshop, we have inspected raw sequence data in fastq format and assessed alignment data in bam format. Thereafter, we perform variant calling to detect genetic aberrations by comparing sequence reads and its reference genome. Variant annotation is the process by which variants and mutations in the DNA are assigned functional information. This information is crucial to identify disease-causing mutations among all variants detected using NGS. In this second part of the workshop, we will learn how to perform variant annotation, interpret SNVs and CNVs.

Variant Effect Predictor (VEP) is free variant annotation tools for both non-commercial and commercial use and is available as a standalone version under the Apache 2.0 license via an intuitive web server (www.ensembl.org/Tools/VEP) or through an API. VEP produces reports in several standard formats as well as a customizable output and annotates SNVs and indels with population allele frequencies, gene/transcript effects, site conservation scores and predicted functional impact scores and classifications based on dbSNP.

Part I: SNVs prioritization in exome data

Case 1

A 40-year-old man with congenital sensorineural hearing loss was admitted to the clinic for deterioration of visual acuity. Ophthalmological assessment is suggestive for retinitis pigmentosa. Based on these features, the patient is suspected of having Usher syndrome. Genetic test on the visual impairment panel is therefore requested. There are no other affected family members.

Go to folder "`2.Exome variant interpretation/data/variant interpretation/Case1/`" and open the file "`Case1_hcdiffs.txt`" by dragging it to Excel. *(Note: sometimes dragging will not work and you will need to open the file directly from Excel)*

You see a list of variants that were extracted from the next generation sequencing data. Figure 2 shows an example of how to filter on columns in Excel by following the numbers.

(Hint: Read the description of each column in the quick guide document. Understanding the information in each column will help you to answer the next questions)

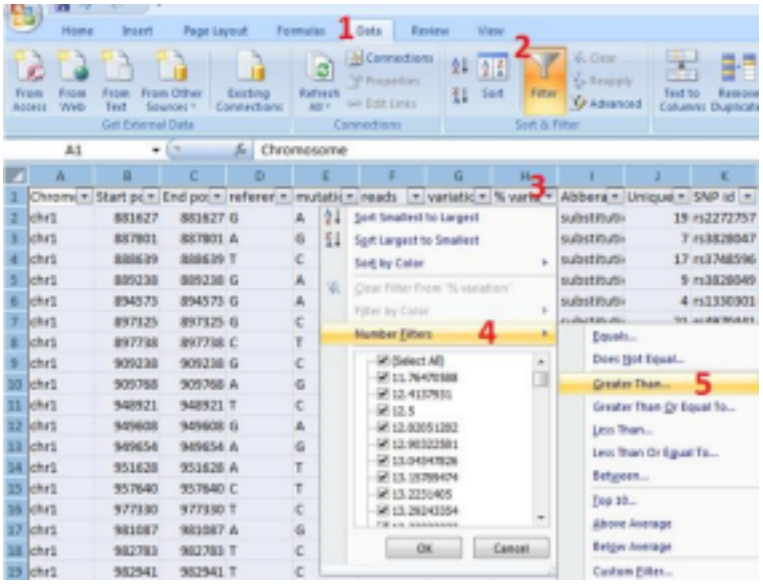

Figure 2. How to filter in Excel example.

Q4. *How many variants were found in this patient?*

Q5. *Of all these, how many are insertions or deletions (indels), and how many are substitutions?*

Q6. *How many variants are located in the exome targets (i.e. exons/ canonical splice sites)?*
(Hint: Look in the column "Gene component")

Q7. *What kind of amino acid consequences are most likely disease-causing?*
(Hint: Look in the column "Mutation AminoAcid". An asterisk "*" represents stop codon and "X" for a frameshift.)

Note: Slightly more detailed this information is also available in the column "Protein effect" based on the Variant Effect Predictor (VEP). See this page for all possible consequences: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html.

Q8. *For a recessive disease that occurs in 1:100,000 individuals, what would be a reasonable allele frequency cut-off?*

Q9. *In addition to gnomAD, we use an in-house allele frequency database. This dataset consists of all variants which have been detected in our center. This information is in the column "**In house Frequency**". What is the advantage of having an in-house allele frequencies database?*

Q10. *PhyloP score represents the degree of evolutionary conservation of the corresponding base (higher score in more conserved base). How would you use this information in variant prioritization?*

(Note**:** Variant pathogenicity prediction programs use evolutionary conservation as one of the criteria to estimate the pathogenicity of a variant. You can think of tools like SIFT, PolyPhen2, CADD, MutationTaster etc. However, the sensitivity and specificity of the predictions is still relatively low. Therefore we use these predictions not for filtering but only on the final set of individual variants for additional proof.)

Q11. *An initial analysis step of an exome focuses on the known disease **genes** for a particular disease. For this reason, we have made virtual gene panels for several genetic disorders, and included this information in the "**Gene Panel**" column. How would you use this information? How many variants do you find for example in the "BLIND" gene panel?*

Q12. *Whereas the panels depict genes previously associated with a disease, we also keep track of individual variants previously associated with disease (column "**Mutation database**"). Common resources for this are HGMD (http://www.hgmd.cf.ac.uk) and Clinvar (https://www.ncbi.nlm.nih.gov/clinvar/). How would you use the column?*

Q13. *How would you further prioritize your candidate genes/variants?*
(Hint: Include variants within coding regions, remove common variants (population allele frequency ≥1%), include nonsense and frameshift variants **OR** missense variants with PhyloP>2.5, and use the Gene Panel column (contains BLIND). With the suspected inheritance pattern for the patient in mind, **use any other filters you think are appropriate**)

Q14. *What is your gene of choice to be the most likely cause for this disorder upon variant prioritization?*

Q15. *Which variant is most likely to be causative in this patient?*

Close file "`Case1_hcdiffs.txt`"

Case 2

A 25 year-old woman in her first pregnancy was admitted to the clinic following a fetal scan at 21 weeks with abnormal position of limbs, bell-shaped thorax, cystic kidneys, aortic stenosis, bowing femur, large irregular pancreas. The phenotype is ciliopathies disorder. There is no family history of congenital anomalies. Prenatal WES was requested.

Go to folder "`2.Exome_variant_interpretation/data/variant_interpretation/Case2/`" and open the file "`Case2_hcdiffs.txt`" by dragging it to Excel.

Perform variant prioritization following the same steps and hints as in Q13.

(Hint: Fetal phenotype is indicative for ciliopathy disorder, filter **Gene Panel** column with "CILIO".)

Q16. *Assuming that the mode of inheritance is recessive, what would be your candidate genes upon variant prioritization?*
(Hint: Check the "Mutation database" column in Excel)

Q17. *Which one of these genes is more likely to be causative?*
(Hint: Check these candidate variants in parent's files, i.e. Case2_father_hcdiffs.txt and Case_mother_hcdiffs.txt)

Part III: CNVs detection in exome data

Case 3

A 3 year-old boy presented to the clinic with developmental delay, macroglossia, single palmar crease and clinodactyly. He is the third child of healthy parents and siblings. Whole exome sequencing was performed to identify the genetic cause underlying the disorder. Identify the potential CNV which may explain his phenotype.

Open your browser and go to: https://igv.org/app/. Load genome assembly by clicking 'Genome' and select Human (GRCh37/hg19).
Open bedgraph file by clicking 'Tracks' from IGV dropdown menu, select 'Local File' and load "2.Exome variant interpretation/data/variant interpretation/Case3/case3.bedgraph" into IGV. Adjust the height of the graph by right-click on the sample name, for a single sample a value of 450 is appropriate (see Figure 3):

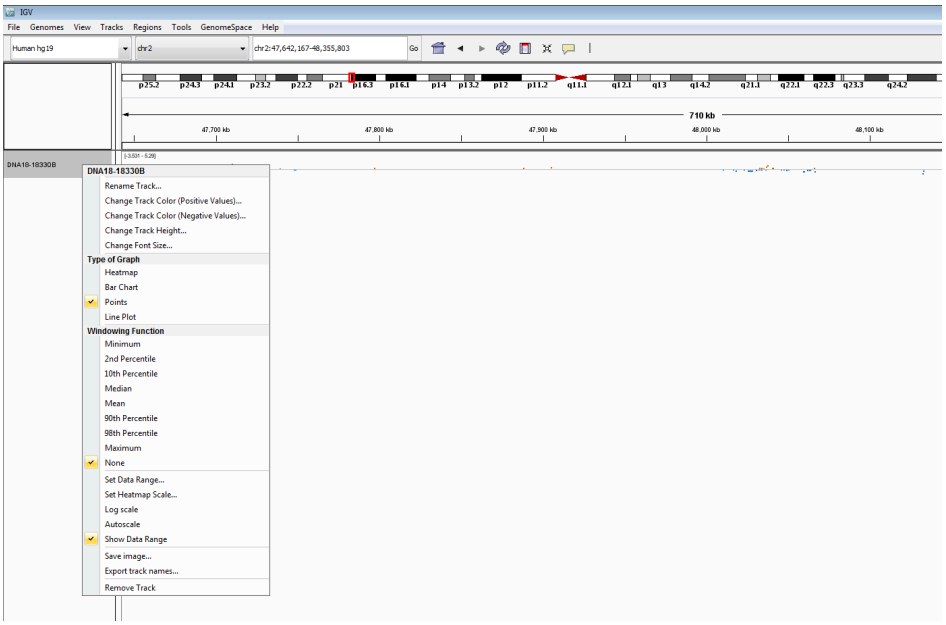

Figure 3. Adjust track height of bedgraph in IGV (red square), and select which chromosomes to look at (purple square).

First look at all chromosomes (select "All" in the dropdown menu, see Figure 3) so you get a good impression of the whole data file, which allows you to notice regions with extreme values.

Q18. *What do you think each dot represents? Do you see changes in the landscape of dots in a particular region?*

Also open a session of Excel, and drag the file "Case3.segments.txt" (from the same folder) into Excel.

As you can see, this file contains 29 rare CNVs that potentially cause the phenotype of this patient. One of the potential CNVs is located on chr11, and you can see this outlier dot also in your bedgraph.

Copy that position ("chr11:84996263") in IGV and zoom out until you see dots again.

Q19. *Do you think this is a real CNV? Do you think this explains the phenotype of the patient?*

(Hint: Check the "Disease gene" column in the segments file in Excel and search online.)

Now go to "chr21".

Q20. *In the segments-file in Excel you see many different variants for chromosome 21. Do you think this is possible? What do you think (considering the bedgraph in IGV ánd phenotype of the respective patient) is the genetic diagnosis for this patient?*

Case 4

6-year-old girl with developmental delay and multiple striking physical features highly suspected of a genetic syndrome diagnosis. Please analyse the WES data for intellectual disability.

Go to the partIII data-folder and drag the file "`/case4.bedgraph`" into IGV, adjust the height of the graph and inspect all chromosomes together first.

Q21. *Can you notice a region with extreme values, and if so, where?*

Now drag the file "`case4.segments.txt`" (from the same folder) into Excel.

Q22. *Do you see a CNV call in this file that corresponds with the region of extreme values from the previous question?*

Case 5

A boy with psychomotor retardation, large ears, and kidney stones. Consanguineous parents, but no positive family anamnesis. In an attempt to establish a genetic diagnosis, the patients DNA is sequenced and your goal is to analyse and identify the potential CNV that causes his phenotype.

Go to the partIII data-folder and drag the file "`/case5.bedgraph`" into IGV, adjust the height of the graph and inspect all chromosomes together first.

Q23. *Can you notice a region with extreme values, and if so, where?*

Now drag the file "`case5.segments.txt`" (from the same folder) into Excel.

Q24. *Do you see a CNV call in this file that corresponds with the region of extreme values from the previous question?*

The Online Mendelian Inheritance in Man (OMIM; https://www.omim.org/) is an online catalog of human genes and genetic disorders. OMIM provides extensive evidence (i.e. full-text referenced overviews) on all known Mendelian disorders and for over 16,000 genes, whilst focusing on the relationship between genotype and phenotype.

Q25. *Look up your candidate CNV in OMIM, what do you see? Does this match the phenotype of your patient?*

Q26. *Do the genes affected in the "Gene overlap" column match with the description in OMIM? And what can you say about the inheritance pattern described in OMIM, does this match with the patients' family history? And why?*

Q27. *How would you check whether this is a homozygous or a heterozygous deletion?*