**Workshop Part II Exome variant interpretation**

Christian Gilissen (Christian.Gilissen@radboudumc.nl)

Galuh Astuti ([Galuh.Astuti@radboudumc.nl](mailto: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**

**|--- Case1**

**|--- Case2**

**|--- Case3**

**|--- Case4**

**|--- Case5**

**|--- Case6**

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

Contents

[Introduction 3](#_Toc147328920)

[Part I: SNVs prioritization in exome data 4](#_Toc147328921)

[Part II: CNVs detection in exome data 7](#_Toc147328922)

[Description of Annotations Fields 9](#_Toc147328923)

# 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](http://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 the vision disorder “retinitis pigmentosa”. Based on these features, the patient is suspected of having Usher syndrome. Genetic testing of the visual impairment gene panel (virtual panel on exome sequencing data) is therefore requested. There are no other affected family members.*

**Go to folder “2.Exome\_variant\_interpretation/data/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). Note that the variants are called against the* ***GRCh37*** *reference genome!*

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)

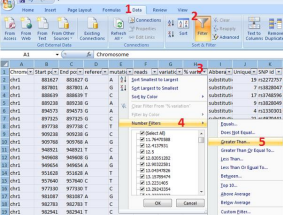


Figure 2. How to filter in Excel example.

***Q1.*** *How many variants were found in this patient?*

***Q2.*** *Of all these, how many are insertions or deletions (indels), and how many are substitutions?*

***Q3.*** *How many variants are located in the exome targets (i.e. exons/ canonical splice sites)?*

(Hint: Look in the column “Gene component”)

***Q4.*** *What kind of amino acid consequences are most likely disease-causing and would you prioritize while trying to diagnose a patient?*

(Hint: Look in the columns “Protein effect” and “Mutation AminoAcid”. Note that an asterisk “\*” represents stop codon and “X” for a frameshift.)

**Note:** the column “Protein effect” is 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.

For rare diseases, we can expect that mutations that cause the disease must appear in the population at a smaller frequency than the frequency of the disease itself. To exclude variants that are too common to cause a disease we can use external databases that contain information about genetic variation in populations. gnomAD (<https://gnomad.broadinstitute.org/> ) is a such a database that collected genetic variation from thousands of exomes of healthy individuals.

***Q5****. For a recessive disease that occurs in 1:100,000 individuals, what would be a reasonable allele frequency cut-off?*

***Q6.*** *In addition to gnomAD, we use an in-house allele frequency database. This dataset consists of all variants which have been detected in our medical center. This information is in the column “****In house Frequency****”. What is the advantage of having an in-house allele frequencies database?*

***Q7.*** *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.

***Q8****. 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?*

*Note: “BLIND” stands for “Blindness” which is the name for our gene panel for genes involved in visual impairment*

***Q9****. 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*](http://www.hgmd.cf.ac.uk)*) and Clinvar (*[*https://www.ncbi.nlm.nih.gov/clinvar/*](https://www.ncbi.nlm.nih.gov/clinvar/)*). How would you use the column?*

***Q10****. How would you further prioritize your candidate genes/variants?*

(Hint: Include variants within coding regions, remove common variants (population allele frequency ≥0.01 ), 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**)

***Q11.*** *What is your gene of choice to be the most likely cause for this disorder upon variant prioritization?*

***Q12.*** *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 ultrasound scan at 21 weeks of pregnancy which showed abnormal position of limbs, bell-shaped thorax, cystic kidneys, aortic stenosis, bowing femur, large irregular pancreas. The phenotype fits in the group of disorders that is called ‘ciliopathies’. There is no family history of congenital anomalies. Prenatal WES was requested.*

**Go to folder “2.Exome\_variant\_interpretation/data/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”.)

***Q13.*** *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)

***Q14.*** *Which one of these genes is more likely to be relevant for the disease?*

**Case 3**

*A 4 year old girl presented to the clinic with autism spectrum disorder, learning disability, delayed motor development and slow potty training. Please analyse the WES data for intellectual disability/neurodevelopmental disorders. The corresponding gene panel is called ”MR” which stands for “Mental Retardation”.*

**Go to folder “2.Exome\_variant\_interpretation/data/Case3/” and open the file “Case3\_hcdiffs.txt” by dragging it to Excel.**

***Q15.*** *Besides recessive, dominant and x-linked inheritance what other “inheritance” pattern can cause disease?*

Up to 40% of intellectual disability cases can be explained by de novo mutations in protein-coding genes.Prioritize this exome data for de novo variants. (Hint: Check the “De novo assessment” column in Excel).

***Q16.*** How many “potential de novo mutations” are there? Why do you think are there so many?

(hint: have a look at the quality of the variants? i.e. number of reads etc.)

***Q17.*** What is the likely genetic cause of disease in this patient?

# Part II: CNVs detection in exome data

**Case 4**

*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 IGV on your computer (alternatively you can do this in 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/Case4/case4.bedgraph”  into IGV.**

**Adjust the height of the graph by right-click on the sample name, for a single sample, choose “Change Track Height” and set the value to 450 (see Figure 3):**

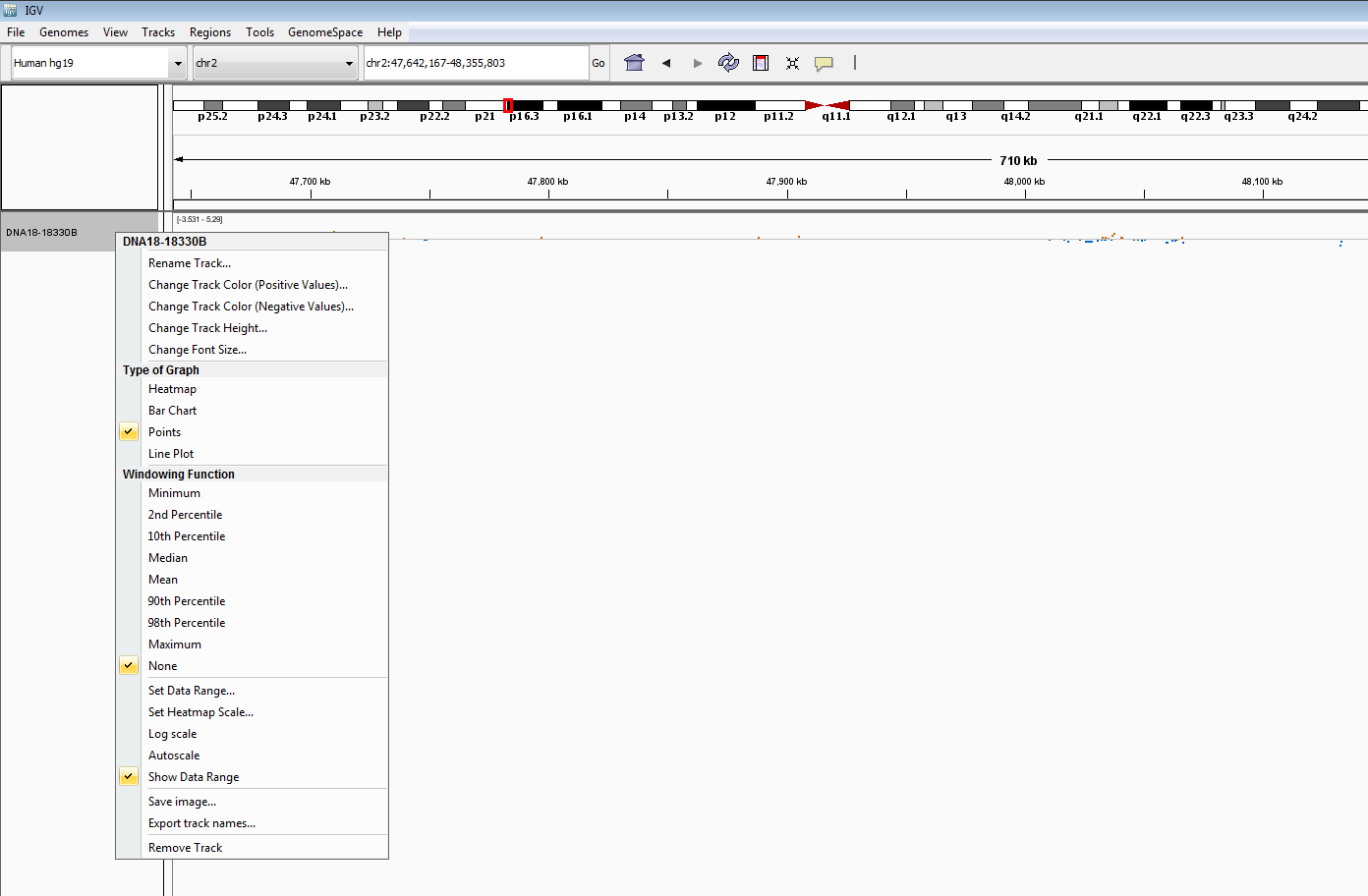


Figure 3. Adjust track height of bedgraph in IGV (“Change Track Height”), and select which chromosomes to look at (purple square).

**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 “Case4.segments.txt“ (from the same folder) into Excel.**

As you can see, this file contains 1081 CNVs calls. Select only those that are larger than 10kb (column “Length”) and that overlap with a known disease gene (column “Gene Panel” should not be empty).

Q17. How many potential CNVs are left? One of the potential CNVs is located on chr17, and you can see this outlier dot also in your bedgraph.

that potentially cause the phenotype of this patient.

**Copy that position (“chr17:44239697”) 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 5**

*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 data-folder and drag the file “/case5.bedgraph” into IGV, adjust the height of the graph (track height) and inspect all chromosomes together first.**

***Q21.*** *Can you notice a region with extreme values, and if so, where? (Look at the column “Value” this represents how much lower or higher the coverage is than what is expected. More extreme values are more likely to be genuine).*

**Now drag the file ”case5.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 6**

*A boy with psychomotor retardation, neonatal seizures, feeding problems, large ears, and kidney stones. Consanguineous parents, and a negative family history for similar problems. In an attempt to establish a genetic diagnosis, the patient’s DNA is sequenced and your goal is to analyse and identify the potential CNV that causes his phenotype.*

**Go to the data-folder and drag the file “/case6.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 ”case6.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/**](https://www.omim.org/)**) is an online catalogue 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 by searching on the cytogenic coordinates (e.g. 15q31)? 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?*

# Description of Annotations Fields

|  |  |
| --- | --- |
| **Field** | **Description** |
| **Chromosome** | Chromosome where the identified variant is located on, i.e. chr1 – chr22, chrX, chrY and chrM (mitochondrial chromosome). |
| **Start position** | Start position of the identified variant (genomic position on the respective chromosome). |
| **End position** | End position of the identified variant (genomic position on the respective chromosome). In case of a substitution or insertion the start and end position are identical, in case of a deletion the end position is the start position + the length of the deletion. |
| **Reference** | Reference nucleotide (wild type) at the given genomic position (based on HG19/NCBI build 37). |
| **Variant** | Variant nucleotide detected at the respective position. |
| **Total reads** | Total number of reads at the given genomic position. |
| **ALT reads** | Total number of reads with DNA variant at the given genomic position |
| **VAF** | Percentage of sequence reads with DNA variant divided by the overall coverage at that position. |
| **Quality score** | Variant call confidence normalized by depth of sample reads supporting a variant. |
| **Variant type** | Type of detected aberration (substitution, deletion, insertion, complex). |
| **SNP id** | rs‐number as given in dbSNP (dbSNP version is given in the meta info of the variant file). |
| **SNP state** | **Same SNP**: the exact identified variant is reported in dbSNP.  **Overlapping location**: there is a variant reported in dbSNP at the same position, but the nucleotide change is different.  **Empty field**: the identified variant is not reported in dbSNP |
| **SNP reference** | Reference (wild type) allele of the reported SNP. |
| **SNP variant** | Variant allele of the reported SNP. |
| **SNP Frequency** | SNP frequency as given in dbSNP. If no frequency information is available you will see an artificial value of ‐1. |
| **Mutation database** | When the identified variant has already been reported to be causative (e.g. by in our in‐house database, Alg018 or HGMD) there will be an entry listed in this column. This entry tells us the project for which the variant has been identified to be causative. |
| **Gene name** | Name of the gene in which the variant was identified. |
| **NC Gene name** | Name of the non-coding gene in which the variant was identified. |
| **Gene component** | The gene component tells whether a variant is located within a gene, which part of the gene or outside the gene. |
| **Local position string** | Text that describes the location of the variant with respect to the nearest start / end of the genomic component (exon/intron) it is located in. |
| **Protein Effect** | Text that describes the location of the variant with respect to the nearest start / end of the genomic component (exon/intron) it is located in. |
| **Reference Amino Acid** | Reference amino acid (wild type) at the respective position. |
| **Mutation Amino Acid** | Variant amino acid based on the mutation sequence at the respective position.  ‘\*’ indicates a stop codon, whereas ‘X’ indicates a frameshift. |
| **Synonymous** | TRUE: the identified variant is synonymous (= does not lead to an amino acid change) FALSE: the identified variant is non synonymous (= lead to an amino acid change) |
| **Hgvsg** | HGVS genomic nomenclature (e.g. 1:g.69511A>G) |
| **Hgvsc** | HGVS coding nomenclature (e.g. ENST00000335137.4\_2.1:c.421A>G) |
| **Hgvsp** | HGVS protein nomenclature (e.g. ENSP00000334393.3:p.Thr141Ala) |
| **phyloP** | PhyloP score for evolutionary conservation on nucleotide level. Score ranges from ‐7 till +7. The higher the score, the better the conservation. |
| **CADD\_PHRED** | Combined Annotation Dependent Depletion (CADD) Phred score. *Kircher M, et al, Nat Genet 2014* suggests to put a cutoff somewhere between 10 and 20. |
| **Grantham Score** | Grantham scores categorize codon replacements into classes of increasing chemical dissimilarity (Granthan R., Amino acid difference formula to help explain protein evolution. Science 1974 185:862‐864).  *\*Note that this is an unreliable score and is best not used for variant prioritization.* |
| **De novo assessment** | Only relevant for trio (de novo) analyses.  **MV** = variant was called in the mother  **PV** = variant was called in the father  **Maternal** = variant was not called in the mother, but inspection of the alignment showed inheritance from the mother  **Paternal** = variant was not called in the mother, but inspection of the alignment showed inheritance from the father  **Shared** = variant was not called in the mother or father, but inspection of the alignment showed inheritance from both  **Possible de novo** = variant was not called in the mother or father and no additional evidence for inheritance was found in the parental alignment files |
| **Index calls** | Only relevant for trio (de novo) analyses. In that case it represents the individual base calls of the reads at this position. |
| **Paternal calls** |
| **Maternal calls** |
| **Gene Panel** | Information on whether the gene in which a variant was found is part of one (or more) of our gene packages. In addition, the known inheritance pattern is given (e.g. blindness (AR,AD)). |
| **EXAC AF** | Allele frequency from 60,706 unrelated individuals. |
| **gnomAD-E AF** | Allele frequency of the matching variant from the gnomAD exome database. |
| **gnomAD-G AF** | Allele frequency of the matching variant from the gnomAD genome database. |
| **In house frequency** | Allele frequency of the matching variant from our in house Exome sequencing data of 59,210 individuals |