Exercise 8: Variant detection

Please read all the instructions before you start the exercise.
Create a folder called “exercise8” in your WEXAC-mounted classNN folder (testing\classNN\exercise8).
Create a file with your answers called “Exercise_8_2019_answers.docx”.

In this assignment, you will explore and analyze genetic variation data. The data are located at:
course_2019/NGScourse_ex8

1. Manual exploration of genetic variation

Let’s use the IGV browser to examine and judge the quality of a variant calling dataset that was performed with GATK, as was explained in the class.
Please open the IGV browser, select the Human hg38 genome, and upload the bam files of the sequenced individuals (isrlp1.bam and isrlq1.bam) and the vcf file with the variants (Ex8_variants.vcf).

An example: Go to position chr14:20033961. On top, you can see the vcf data for this region, a C->T call, along with the genotypes of the two sequenced individuals, isrlp1 and isrlq1. Here the variant is present only in isrlq1 and the call is heterozygote. Use the cursor to see the vcf content for this call.

a) Which sequencing strategy was used here? (targeted gene sequencing? Whole genome sequencing?). Explain your answer.

b) Examine the below genomic locations. What is the genetic variation in each location, and what is the genotype of each individual (homozygote**/**heterozygote/reference homozygote)?

   chr14:20179531
   chr14:25431612
   chr14:74484297
   chr14:88874949

c) chr14:19679077. The genetic variation of this position has a low quality, and was filtered out from the final dataset. Why?

2. Identifying the disease causing mutation of Spastic Paraparesis 49 disease

Spastic paraparesis 49 (SPG49) is an **autosomal recessive** complicated form of spastic paraplegia, a neurodegenerative disorder of the corticospinal tracts. Patients suffer from delayed psychomotor development, mental retardation, and onset of spastic paraplegia in the first decade. They also have dysmorphic features, thin corpus callosum on brain imaging, and episodes of central apnea, which may be fatal. The disease was identified in several families coming from **Jewish Bukharian** origin, suggesting a single founder mutation.

To underlie the genetic cause of SPG49 **exome sequencing** was applied to 3 patients, following by a functional annotation of the genetic variants using the wAnnovar tool (http://wannovar.wglab.org/).
The excel file ex8_SPG49_chr14.xls contains a set of 35,761 variants from chr14 that were called in the experiment along with a wAnnovar annotation. Note, that some of the information is sometimes missing. For example, polyphen and sift scores, which assess the probability of a mutation to damage the protein

function, can be calculated only for non-synonymous variations (the information is available for most of them).

Before you start, read the detailed explanation about the content of each column in the excel file (worksheet info in the excel).

Questions:

a) What type of genetic variations that might harm a proper function of a protein you identify in the file? See in the **info** worksheet which column to use. (SNV = single nucleotide variation)

b) Assuming that the disease-causing mutation is located on chr14, try to identify the candidate mutation. How? Filter the data to contain only rare variants (below 1% in the human population), exonic, that affect the protein (i.e. without synonymous variations). **Keep this list for question d.** You should be left with 72 variations.
Assuming the disease is recessive (all affected individuals are homozygote to the mutation, all genotypes=1/1), which are the final candidate mutations? In which genes? (you should be left with 2 candidate mutations).

c) Use the Online Mendelian Inheritance in Man database (OMIM, https://omim.org/) to suggest which of the genes is involved in the disease etiology. (Hint: search OMIM which the gene symbol and read the information on the gene).

d) Suppose the disease was dominant (i.e. all the 3 individuals are heterozygotes to the mutations, all have the genotype 0/1), what would be the outcome? Filter the list that you generated in b, to get all SNVs that are heterozygote in all the individuals. You should be left with 5 SNV. Further, use polyphen and sift predictions to select the two most deleterious mutations.
Answer in which genes are the 5 SNVs that you identified? And which are the two most deleterious SNVs that are suggested by SIFT and PolyPhen?

e) Examine the information provided by the ClinVar database. Can you identify additional pathogenic variations in the data (i.e. not causing SPG49)? In which individual?

f) As the DNA was collected under Helsinki permission to search for the HSP49 mutation, should the patient be notify about the other pathogenic mutation that you discovered in question **e**? (this is an ethical question and will not be graded)