Exome sequencing analysis of osteoarthritis patient.

Ingrid Meulenbelt and Yolande Ramos - Dept. Molecular Epidemiology

30/10/2020

Before you start

For this practical you need 2 files called ExoomSeq_Patient_2020_final.dat and Expression-cartilage.dat that can be found on the Github.

For the analysis you need R.

Introduction

About 15 years ago a family was included in our research of which several members were suffering from relatively severe, generalized early onset osteoarthritis (OA). At that time linkage-analysis was performed, but we could not definitely identify the gene involved.

Last year, next generation exome sequencing was performed using DNA of one of the affected family-members. In this exercise, using the data that were generated you will be searching for the mutation most likely responsible for development of OA in the family.

» Write your answers in a Word document «

I. Insight into the family

Figure 1 shows the pedigree file of the family.

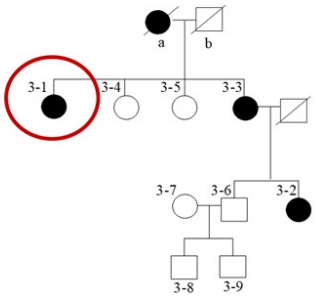

Figure-1

1a. Give an argument whether the disease is inherited in a dominant or recessive way in this family, whether it is X-linked, and whether the penetrance is 100% or smaller. Which factors can play a role in the level of penetrance?

1b. Are patients homozygous or heterozygous for this mutation?

In this family member 3-1 was selected for exome sequencing (see Figure-1).

2a. Explain in your own words what is exome sequencing.

2b. In Figure-2 you see a picture of knee and hand radiography of the OA patient involved. Write down which characteristics of OA you can distinguish.

Figure-2

II. Explore exome sequencing data

In the file ExoomSeq_Patient_2020_final.dat you find all genetic variants identified in the OA patient by means of exome sequencing. Open the file.

```
exome <- read.delim("ExoomSeq_Patient_2020_final.dat", header = T)
```

3a. How many variants were found in the patient? *R tip*

```
length(unique(exome$snp_id))
```

3b. Describe proof of principle how is determined whether a base pair is a variant or not (different according to what?).

In the column 'function' it is indicated where the variant is found and what the effect for the gene is. Calculate the frequencies of the different 'functions' (see box 1).

4a. Make a list of the different categories ranking them by potential damage of gene function. If you think some categories are equally damaging put them together in the list.

4b. Are all categories part of the exome?

4c. What is the total number of variants with a potential damaging effect on the gene-function in these patients? Can you determine which of the variants is causal for OA in this family?

4d. Describe the information found in the other variables. What is in the variable 'state' the putative implication of 'known/novel'? The variables 'read_support' and, related to that, "qc" refers to the so-called 'coverage': how often has the base been sequenced. Why does it sometimes show 2 numbers? What is the implication of low values in this variable?

BOX 1: Frequencies; To obtain frequency table use:

```
library(dplyr)
library(janitor)
Tabfreq <- exome %>% tabyl(function.)
```

III. Predict harmfulness

Not all potentially damaging mutations are in fact damaging this actually depend on the amino acid change. The program (Protein Variation Effect Analyzer (PROVEAN), previously called SIFT is designed to predict the effect (damaging or tolerated) of a base -amino acid change on protein function.

5a. Give an example of an amino acid substitution (missense mutation) that has no effect on protein function.

Go to the website (http://sift.jcvi.org/) and check which information you need to upload to allow PROVEAN / SIFT analysis by selecting PROVEAN Genome Variants (see arrow in **Figure-3**). At the PROVEAN Genome Variants page (Figure4) the format of the file is given. Moreover, at step 1 it is shown that genomic coordinates and variants can be uploaded in 2 ways.

5b. What information does PROVEAN need to predict the effect of the base / amino acid changes?

5c. What are the 2 ways in which data can be entered?

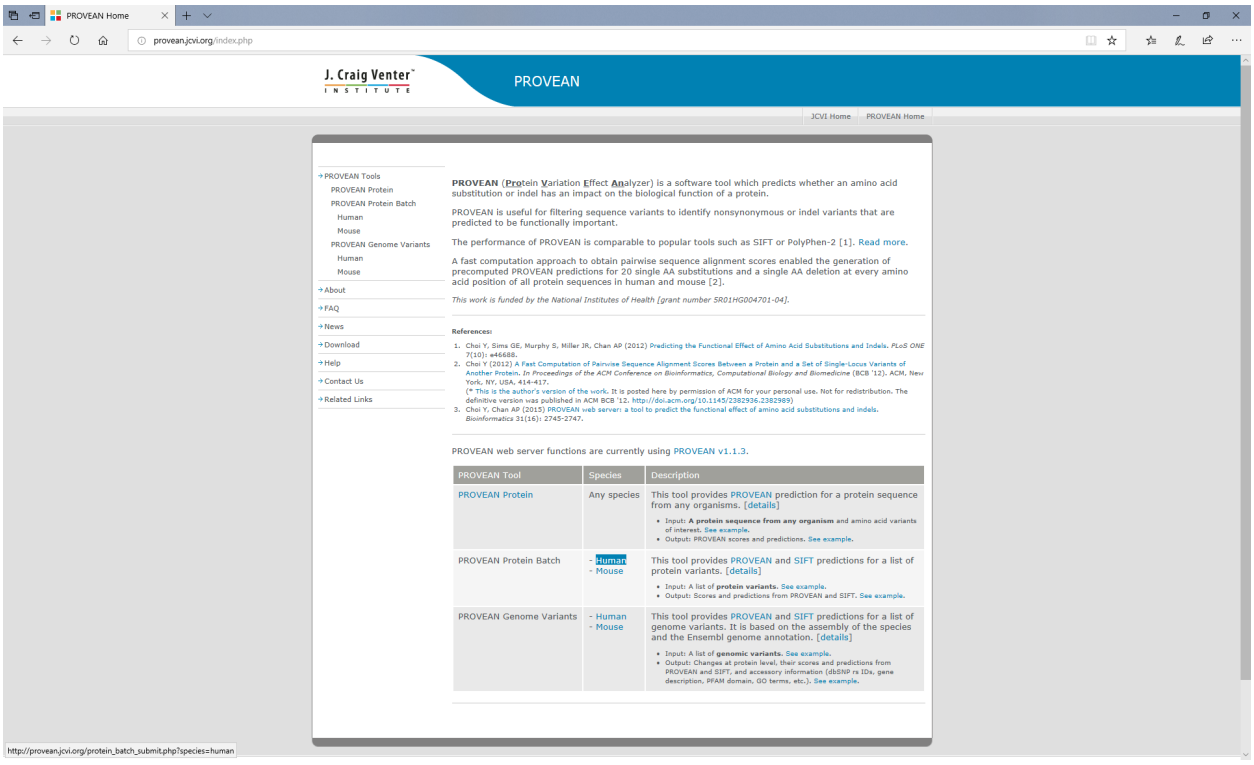

Figure-3

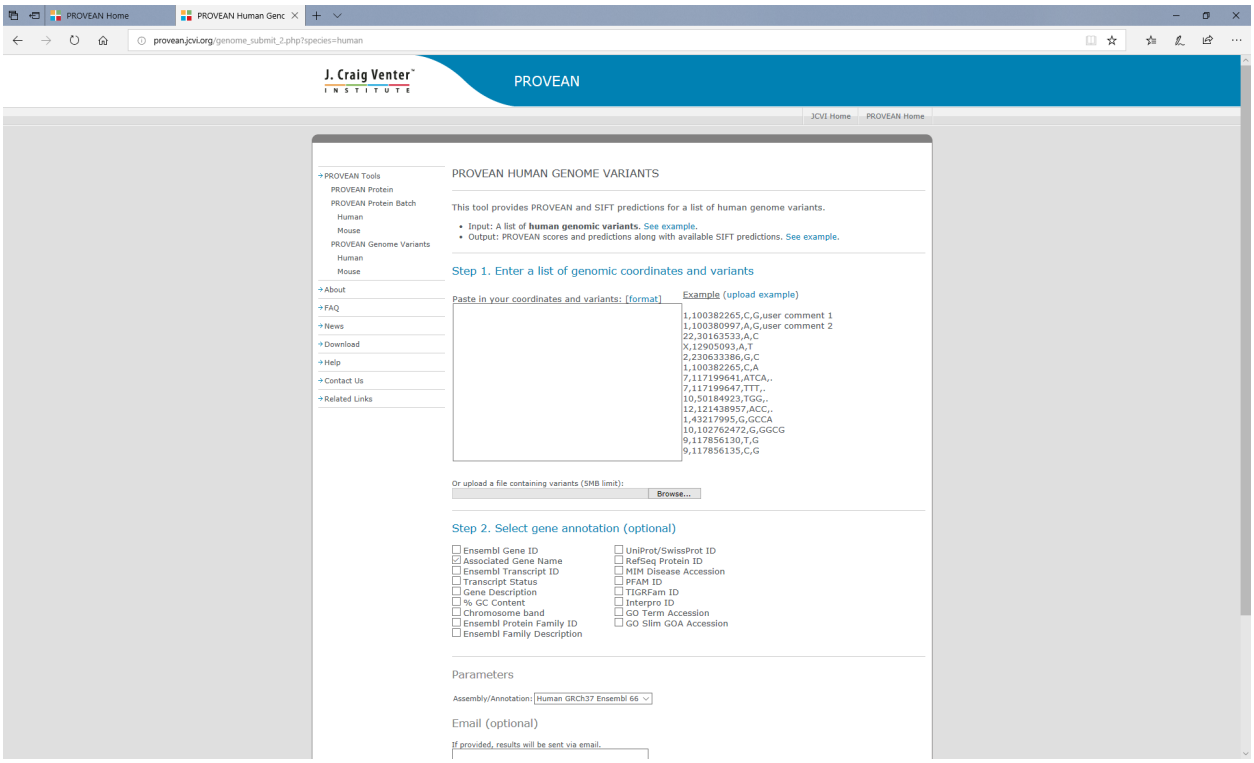

Figure-4

As shown in **Figure-4**, we need a comma delimited file depicting "chromosome", "basepair position on the chromosome", "reference allele", "alternative allele", and as an option you can submit user comments. Alternatively you can enter the base change as C/G, a variable that you find in your dataset. We will now prepare a comma delimited (*.cvs) file from the SPPS file ExoomSeq_Patient.sav to allow the analyses of our exome variants in PROVEAN / SIFT.

- First select for the "novel" variants because it is likely that a high impact mutation causing early onset OA in a family is private.

```
New_exome <- exome[exome$state == "novel",]
```

- Next select for the variants with a qc >= 70 percent to have robust base changes only.

```
New_exome <- New_exome[New_exome$qc >= 70,]
```

- Generate a dataframe in R that only contains only 'novel' variants qc >=70 (see R code). Check if right filter is applied.
- Select variable 'function' and use selection criteria "frameerror" | "missense" | etc.;

Generate a file in SPSS that only contains mutations that may be harmful to the respective protein function and are robustly detected (see Box 2 and your answer to question 4c and 4d). Check if right filter is applied (filter out unselected cases) and delete all unselected variants by selecting output "delete unselected cases".

Remove all variables that you **do not** need in PROVEAN / SIFT analysis and save the file as a comma delimited text-file (.cvs). Alternatively use **save as** and select **only** the variables i.e. chrom, bp, base_change, read_support. Do not save variables names. Give file another name.

Open file in Wordpad or Notepad and check whether the file is ready to upload in PREOVEAN. Since PROVEAN is not able to think the file should exactly look as the example provided (See also Figure-4) i.e. no titles, no spaces and only separations by commas. Use "find -> replace" in Wordpad or Notepad to further modify you file. Safe the final PROVEAN input file that is ready for uploading.

- Upload comma delimited file or use "copy paste" to submit our variants to the PROVEAN / SIFT website (http://sift.jcvi.org/).
- Check at Step 2 – Select gene annotations additional information that you would like for the genes/proteins in which the variants reside e.g. "Associated Gene Name" and/or "Gene Description". Another informative annotation could be 'MIM Disease accession' as it provide you with disease associated to mutation in these genes.
- Click 'send' at the bottom of the page, wait until you are directed to the 'results status page', and wait a little more until the analysis is ready (try 'refresh page' from time to time).

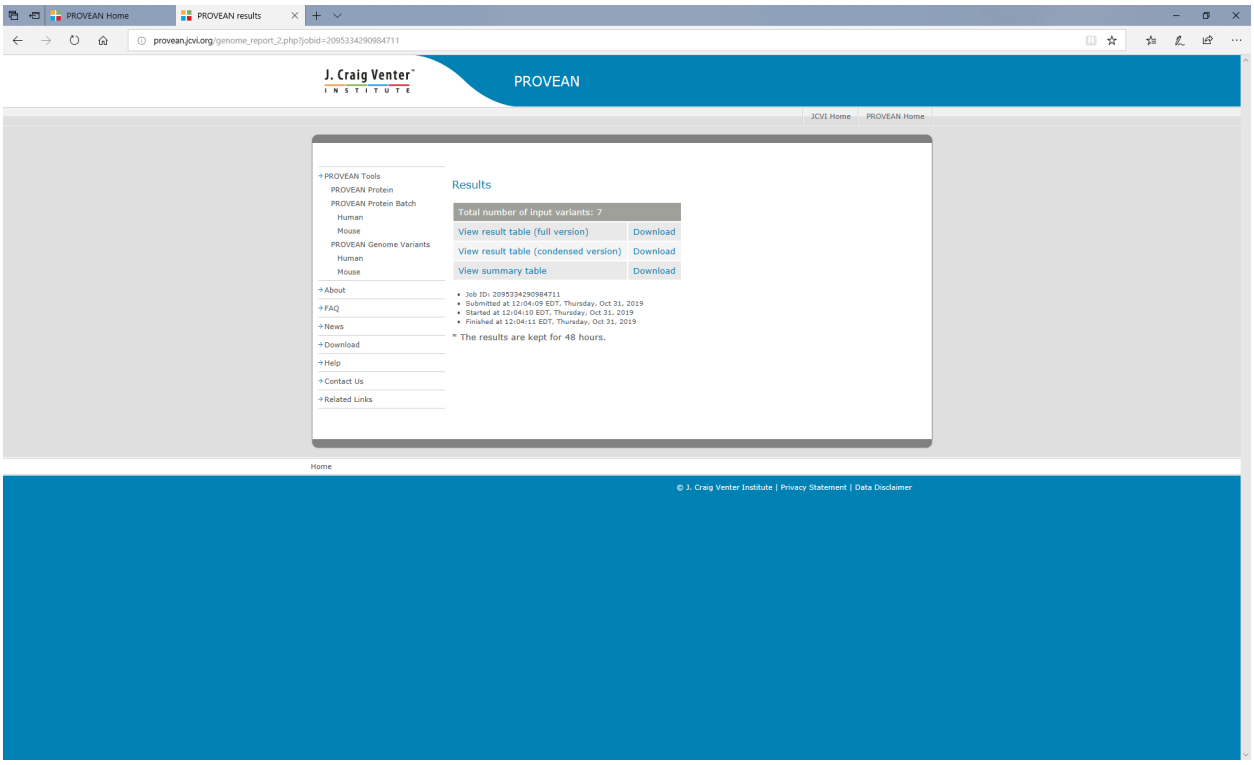

Figure 5

6f. View results of the PROVEAN / SIFT analysis (use the condensed version - middle output, Figure). Which 2 predictions are provided? Are they identical? Why not?

Now download the file and open in R. Check whether all variables are loaded well (if not, then maybe you have separated variables on other symbols –space, comma,. . .- apart from tab). Most important variables are 'dbSNP ID' and 'Prediction'.

6g. How many exome variants have been found in the patient according to PROVEAN / SIFT, including 'Damaging' as well as 'Deleterious'?

What do you think of the damaging variants that have SNP or rs-number? Please select only novel variants.

6i. How many deleterious, damaging and novel variants are present in family member 3-1?

As you see, it is still impossible to identify the causal mutation, the number is still way to high. Logically, mutations causing OA are expressed in the relevant disease tissue i.e. cartilage. In the file Expression-cartilage.dat you can find results of a RNA sequencing expression analysis of preserved and lesioned cartilage material obtained after total joint replacement at the LUMC.

Open the file . Sort the file based on gene-name, and paste all variables of Expression-cartilage.dat file next to the other variables. Check the rows of the table (same number of rows and same gene-names).

6j. How many of the genes with a novel, deleterious, damaging, mutation are highly expressed in cartilage (variable Expression_level_quartiles=4)?

IV. Identification of the putative causal mutation

Using the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk) it is possible to check which mutations have been found for a particular disease (Figure 7). Go to the website login (b.t.heijmans@lumc.nl / HGMD972695). In case you cannot login all together you may have to make your personal account)

Figure 7

7a. Search in which genes mutations have been found for OA (osteoarthritis); see screenshot. In how many genes OA-mutations have been found? Is one of these among the genes selected in question 5e?

7b. Click on the gene in the database and get mutations. How many missense/nonsense mutations have been found? And has the mutation you identified here also been found before (search in the Excel-file, variable 'Substitution', which is the codon it concerns here)? With which diseases are the mutations found close to the mutation of your patient associated?

7d. The mutation was measured within the family by applying mass spectrometry (Sequenome). In Figure 8 you see the genotypes we have found. Does the genotyping confirm the presence of the mutation in the patient that was sequenced? Check with the pedigree shown in 1 whether the mutation is inherited with the disease (A is the mutated allele).

ID	Genotype
3-1	C/A
3-2	C/A
3-3	unknown
3-4	C/C
3-5	C/C
3-6	C/C
3-7	C/C
3-8	C/C
3-9	C/C

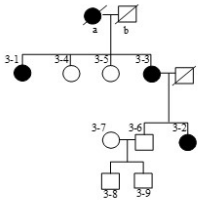

Figure 8

7e. In Figure 9 you see the structure of the protein encoded by the gene. This protein forms a string together with 2 other proteins. Speculate on how the mutation can damage this structure.

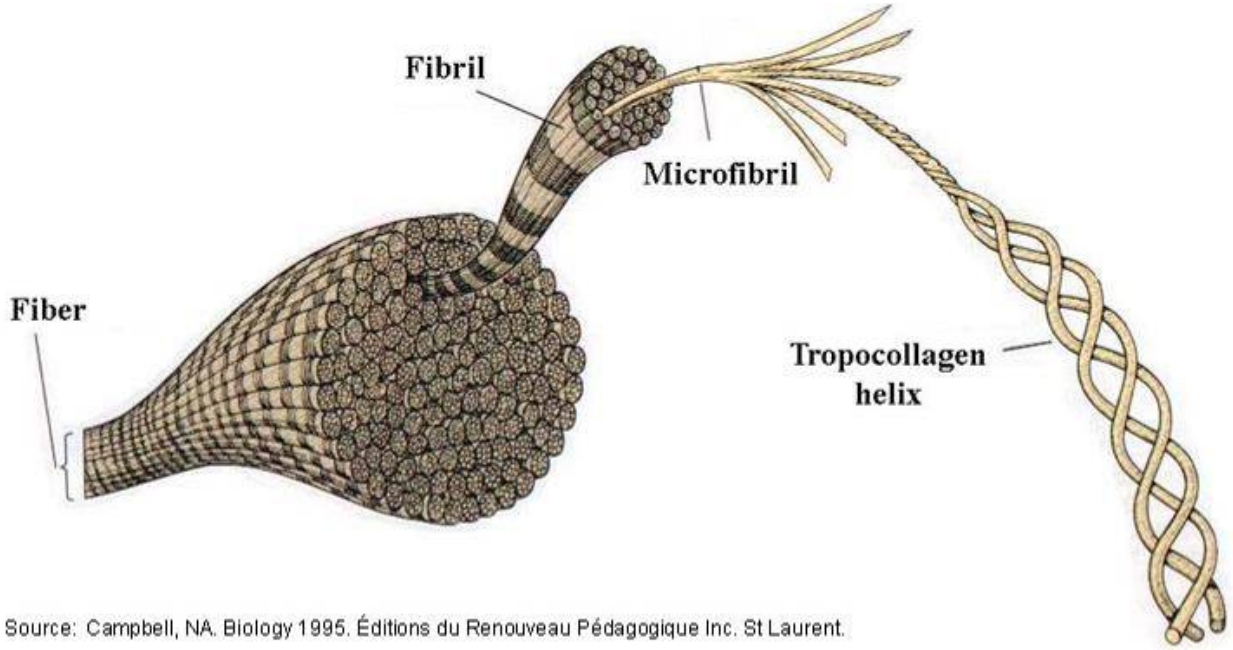

Figure 9