Exome sequencing analysis of osteoarthritis patient

Ingrid Meulenbelt and Yolande Ramos
Dept. Molecular Epidemiology

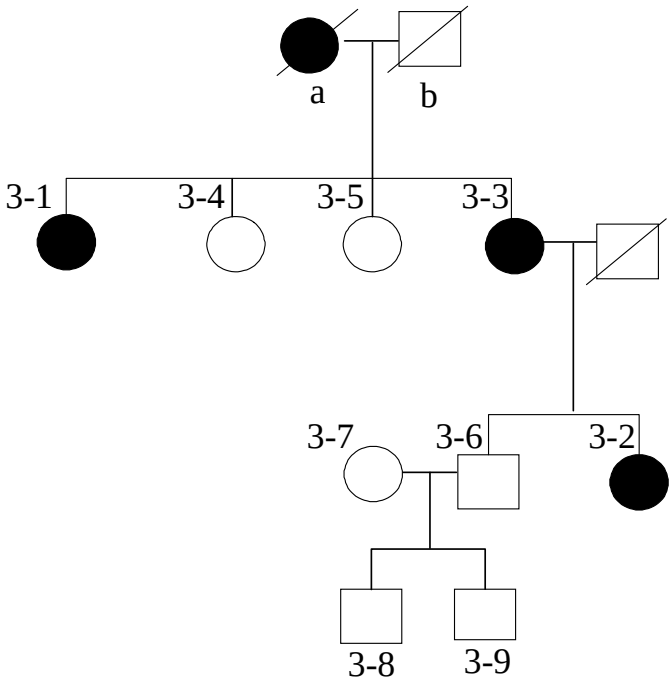
Before you start
1. For this practical you need 2 files called ExoomSeq_Patient.sav and KraakbeenExpressie.xls that can be found on the Github.

2. For the analysis you need SPSS

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
1. Below, you can see the pedigree of the family.



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

2. Family member 3-1 was selected for exome sequencing.
a. Explain in your own words what is exome sequencing.

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

II Explore exome sequencing data

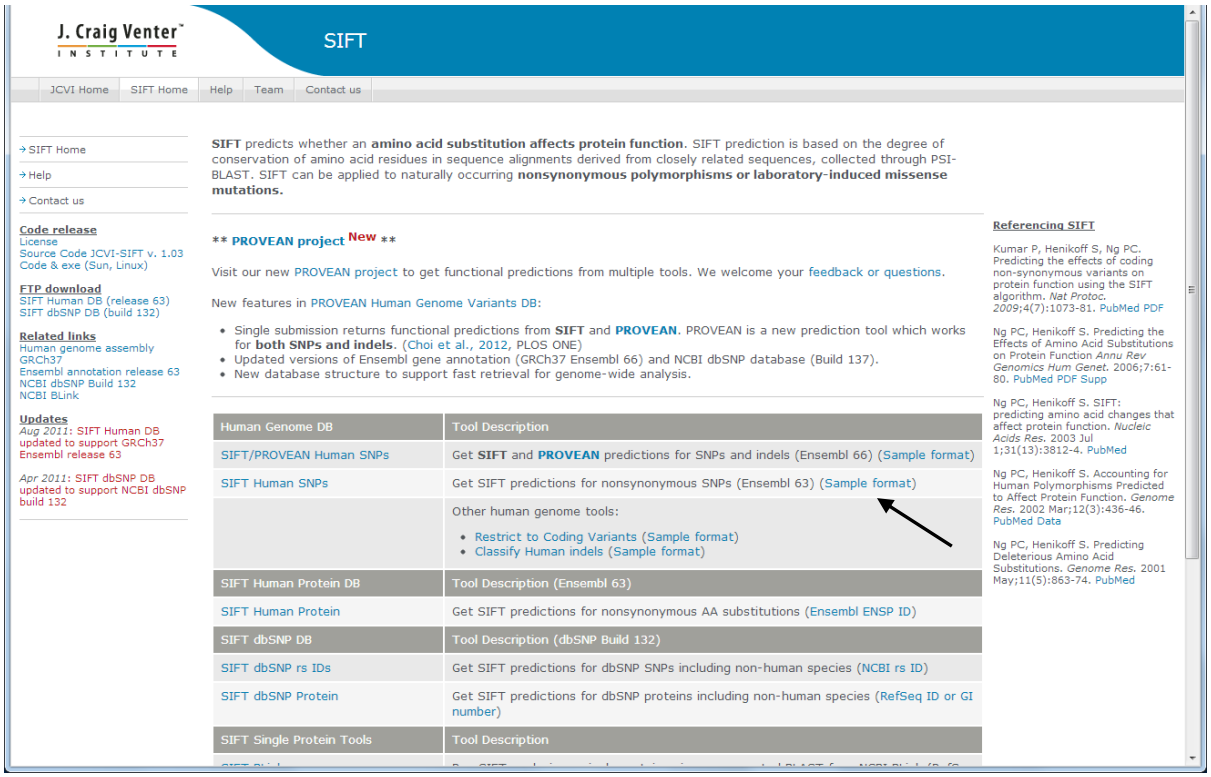
3. In the SPSS file ExoomSeq_Patient.sav you find all genetic variants identified in the OA patient by means of exome sequencing. Open the file.
a. How many variants were found in the patient?
b. Describe proof of principle how is determined whether a base pair is a variant or not (different according to what?).

4. 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’.
a. 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.
b. Are all categories part of the exome?
c. 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?

d. Describe the information found in the other columns. What is in the column ‘state’ the putative implication of ‘known/novel’? The column ‘read_support’ 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 column?

III. Predict harmfulness

5. Not all potentially damaging mutations are in fact damaging. Using SIFT (http://sift.jcvi.org/), it is possible to predict harmfulness.
a. Give an example of an amino acid substitution that has no effect on the function of a gene.
b. Go to the website and check how you have to present your data for analysis with SIFT (see arrow in the screenshot below).



b. Prepare a file for analysis in SIFT based on ExoomSeq_Patient.sav by making the following selections:

- a first selection is based on the identification of ‘new’ variants (Data\Select → Select if condition: state=”novel”)
- subsequently, select for the potential damaging mutations of question 4c (Data\Select → Select if condition: function="frame error" | function="missense" etc.; **NB: first try with ‘filter’, if this is ok repeat with ‘delete’**)
- now, you only keep the following columns:
  chrom, bp, strand, base_change, read_support
- safe the file using ‘save as’ comma-delimited text-file (.csv), and give an obvious name.

- Open in wordpad and delete the first row (titles), all spaces and check that text is separated by commas as answered in 5a (with find/replace).

- Safe file again: now it is ready for use in SIFT.

c. Go to the SIFT website and click 'SIFT human SNPs', choose the file you have just prepared and check 'Gene Name' in the output (see below). Click 'send' at the bottom of the page, wait until you can click on 'SIFT results status page', and wait a little more until the analysis is ready (try 'refresh page' from time to time).

d. Open the results of the SIFT analysis in Excel (download file from SIFT website before opening; it is a tab-separated file). Check whether all columns are loaded well (if not, then maybe you have separated columns on other symbols –space, comma,…- apart from tab). Most important columns are 'dbSNP ID' and 'Prediction'.
How many exome variants have been found in the patient according to SIFT, including 'novel' as well as 'damaging' (**NB: exclude 'damaging with low confidence'**)?

e. As you see, it is still impossible to identify the causal mutation. Logically, mutations causing OA are expressed in the cartilage. In the file KraakbeenExpressie.xls you can find results of a micro-array expression analysis of cartilage material obtained after total joint replacement at the LUMC. From the SIFT-file: delete all variants that are not both 'novel' and 'damaging', sort the file based on gene-name, and paste all columns of KraakbeenExpressie.xls next to the other columns. Check the rows of the table (same number of rows and same gene-names).
- How many of the genes with a new, predicted damaging, mutation are expressed in cartilage (column CartilageExpr=1)?

IV Identification of the putative causal mutation

6. 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. Go to the website login

(b.t.heijmans@lumc.nl / HGMD972695; in case we cannot login all together you may have to make your personal account)

a. 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?

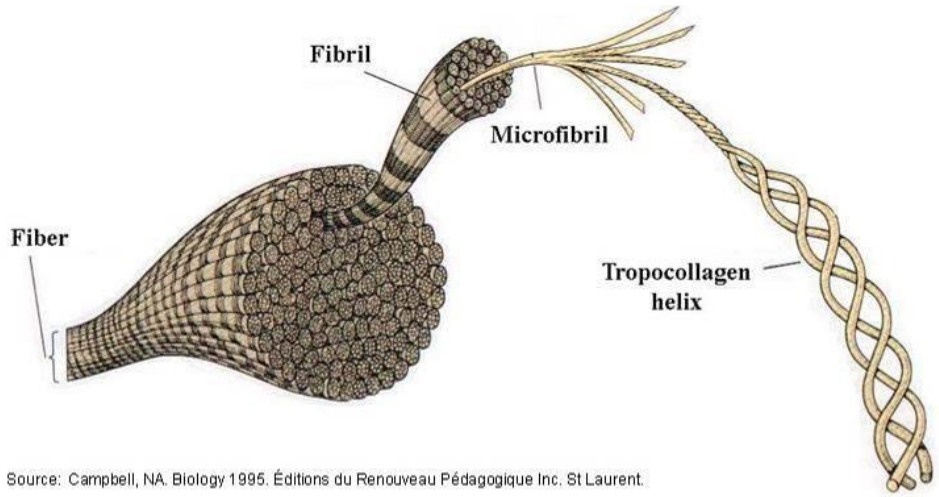
b. Click on the gene in the database. How many missense/nonsense mutations have been found? And has the mutation you identified here also been found before (search in the Excel-file, column 'Substitution', which is the codon it concerns here)? With which diseases are the mutations found close to the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

c. What is the 'support' (in column 'User Comment') for this mutation (how often has the mutation been detected with sequencing)? Given support <4 is unreliable. What to do?

d. The mutation was measured within the family by applying mass spectrometry (Sequenome). In the Table below 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

e. Below 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.



Source: Campbell, NA. Biology 1995. Éditions du Renouveau Pédagogique Inc. St Laurent.