# Bioinformatics module – Zuyd Hogeschool

## Practical Session 1: Assignments Biological Sequence Databases

## Date: May 15th, 2017

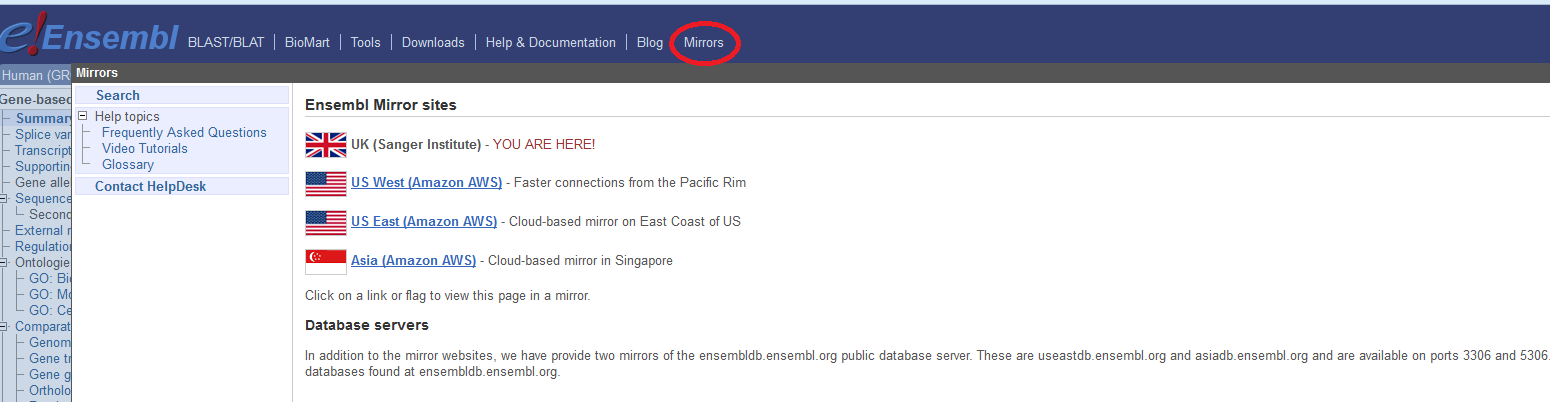
In this practical session you will use two online-available genome-centered data portals to find information on genes, transcripts, proteins and their function. The first one is the Ensembl genome browser/database at the European Bioinformatics Institute (EBI). The second one is the US National Centre for Biotechnology Information (NCBI) data portal, which gives access to a large collection of interconnected databases. The two portals share a lot of underlying information and many of the resources are synchronized on a daily basis, but the interface at two sites differs largely. The choice which one you want to use will be influenced by your personal preference and the task at hand.

Read the practical assignments carefully and try to answer the assignments yourself. If you have any questions you can ask one of the instructors.

***Ensembl genome browser***

Ensembl provides genes and other annotation such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on mRNA evidence in the NCBI RefSeq databases. Website: [*www.ensembl.org*](http://www.ensembl.org)

*Note that if for some reason the Ensembl website is slow you can choose another mirror (see image below)!*



*Assignment 1: Introduction to Ensembl*

First you are going to learn how to use the Ensembl database. Together with one of the instructors you will walk through the Ensembl tutorial available in the file ‘Tutorial\_Ensembl-walk\_through.pdf’.

*Assignment 2: Striated muscle contraction*

Striated muscle contraction is the process where muscle tissue is activated by a signal from the nervous system. In case of voluntary action the nervous signals are initiated from the brain by so called action potentials. In case of reflexes these potentials are coming directly from the spinal cord. Striated muscles are a group of muscles including skeletal and cardiac muscle tissue.

**

*http://wikipathways.org/index.php/Pathway:WP383*

You will use Ensembl (<http://www.ensembl.org>) to find additional information on the **DES** gene.

🖰 Find the human DES gene and answer the following questions:

* 1. As explained in today’s lecture, in Ensembl all the genes have a unique identifier. What is the Ensembl ID of the human DES gene?

ENSG00000175084

<http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000175084;r=2:219418377-219426739>

* 1. What is the full name of the DES gene?

Desmin

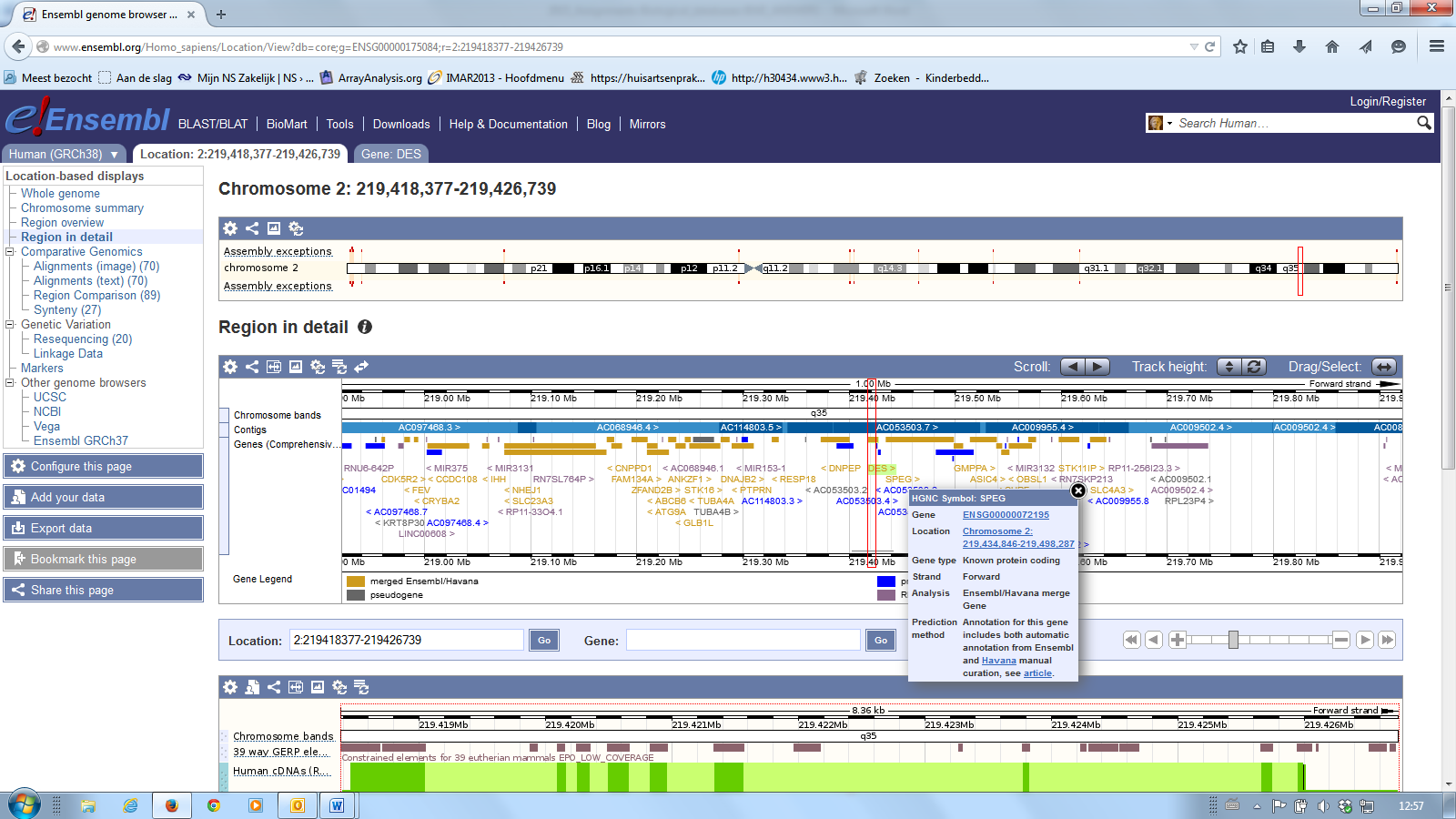
* 1. Find in Ensembl on which chromosome and which strand the gene is located?

Chromosome 2: 219,418,377-219,426,739 forward strand.

* 1. Genes that are located on the same chromosomal region are linked genes. This means that they have the tendency to be inherited together. Therefore, it is of interest to know which protein coding genes are in close proximity. What is the next known protein coding gene on the sequence to the right of the DES gene, on the same strand? (Hint: go to the Location tab).

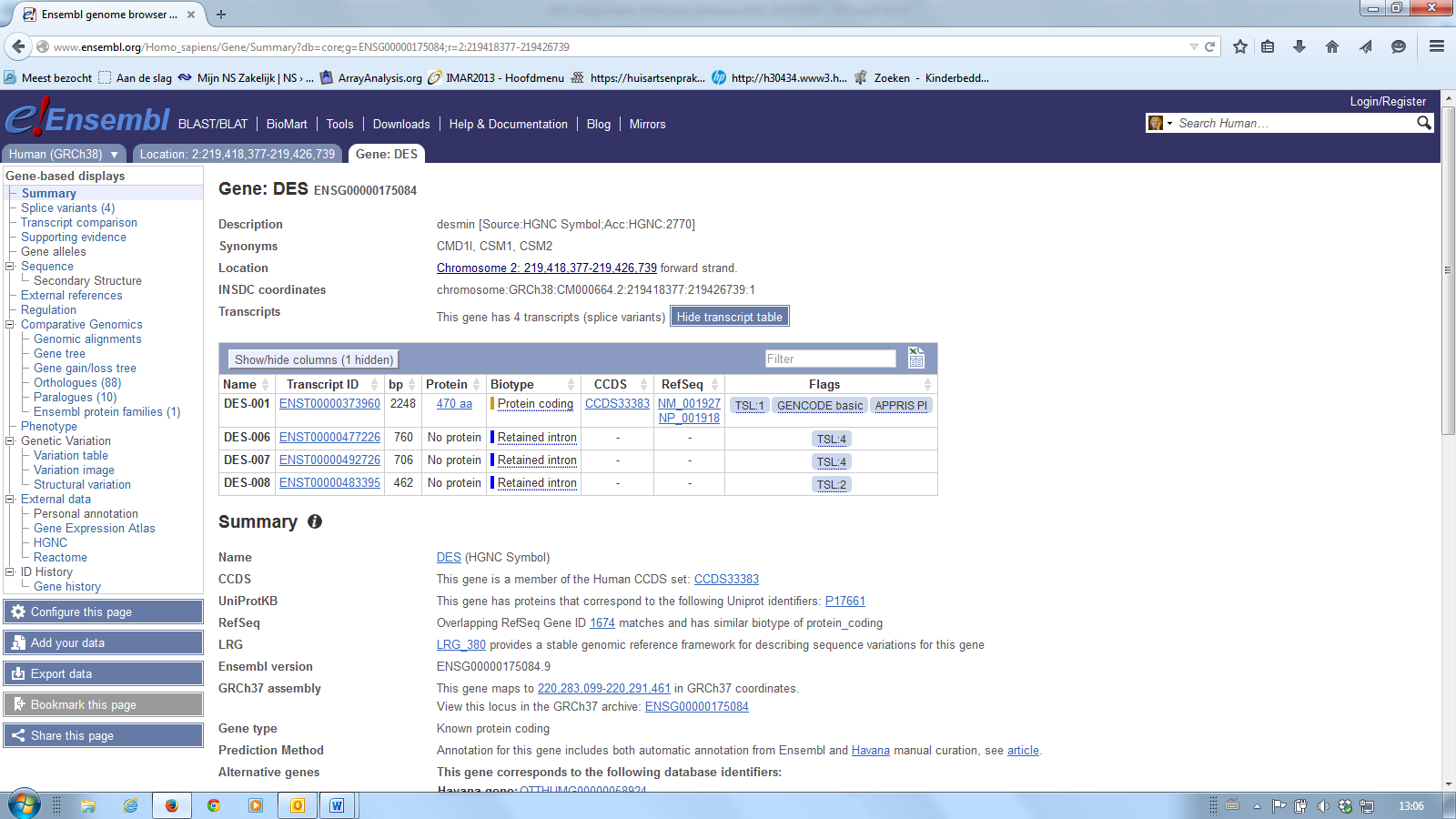
SPEG gene

Make sure you only look at “protein coding genes” indicated in yellow. The strands are indicated with arrows, > means foward stand and < means reverse strand.



* 1. Genes can have more than one transcript due to alternative splicing. These transcripts can give rise to proteins which differ in their sequences and often in their activities or can cause changes in regulatory elements such as translation enhancers or RNA stability domains, which may have a dramatic effect on the level of protein expression. How many transcripts have been predicted for this gene?

Four transcripts



* 1. How many of these transcripts are protein coding, and what are the matching Ensembl transcript ID and RefSeq mRNA and protein IDs?

1 protein coding transcript; ENST00000373960; NM\_001972; NP\_001918

* 1. How many base pairs does the first transcript have and what is the length of the protein (the number of amino acids) it encodes?

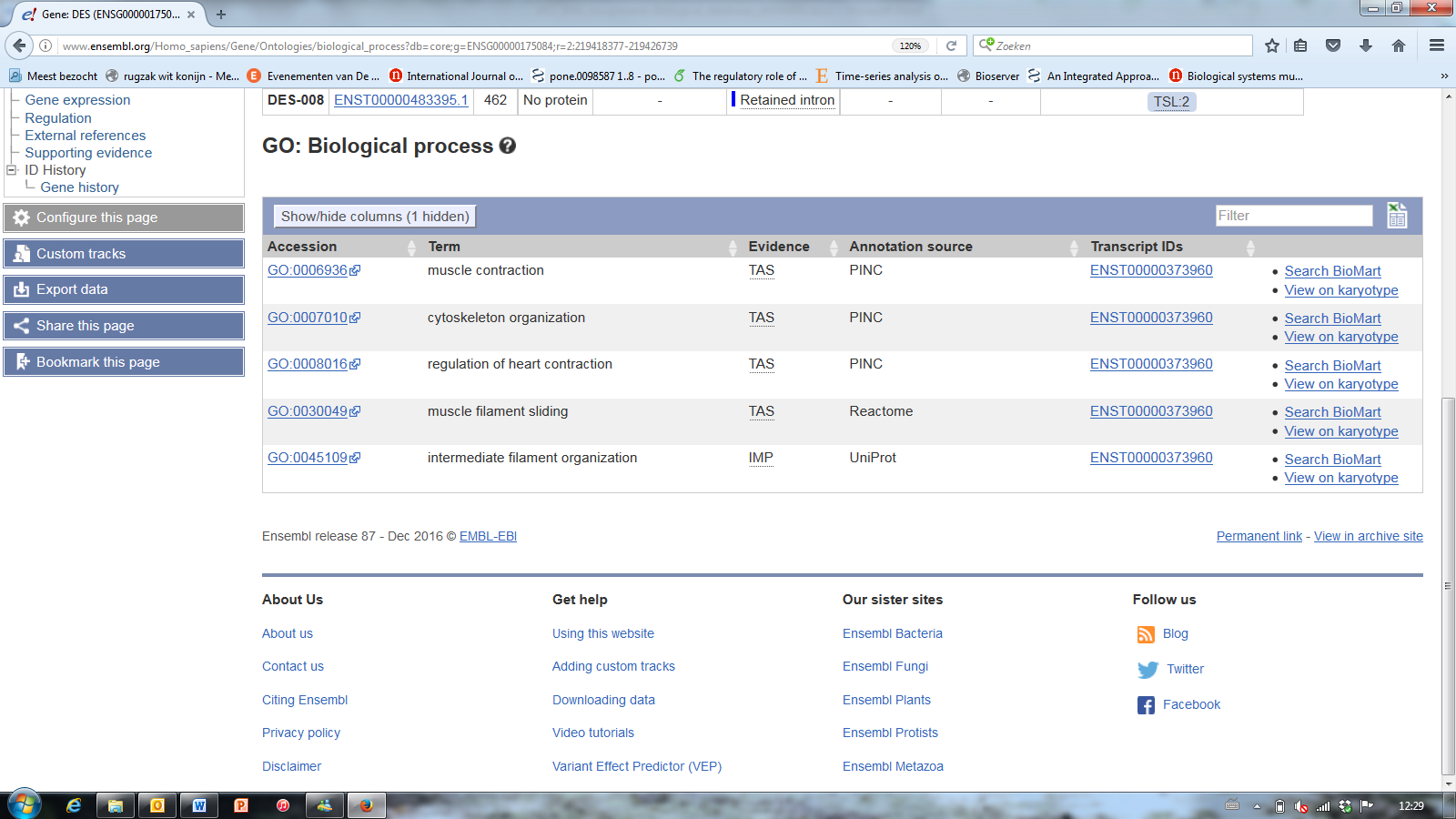
Length transcript: 2248 bp (=base pairs)

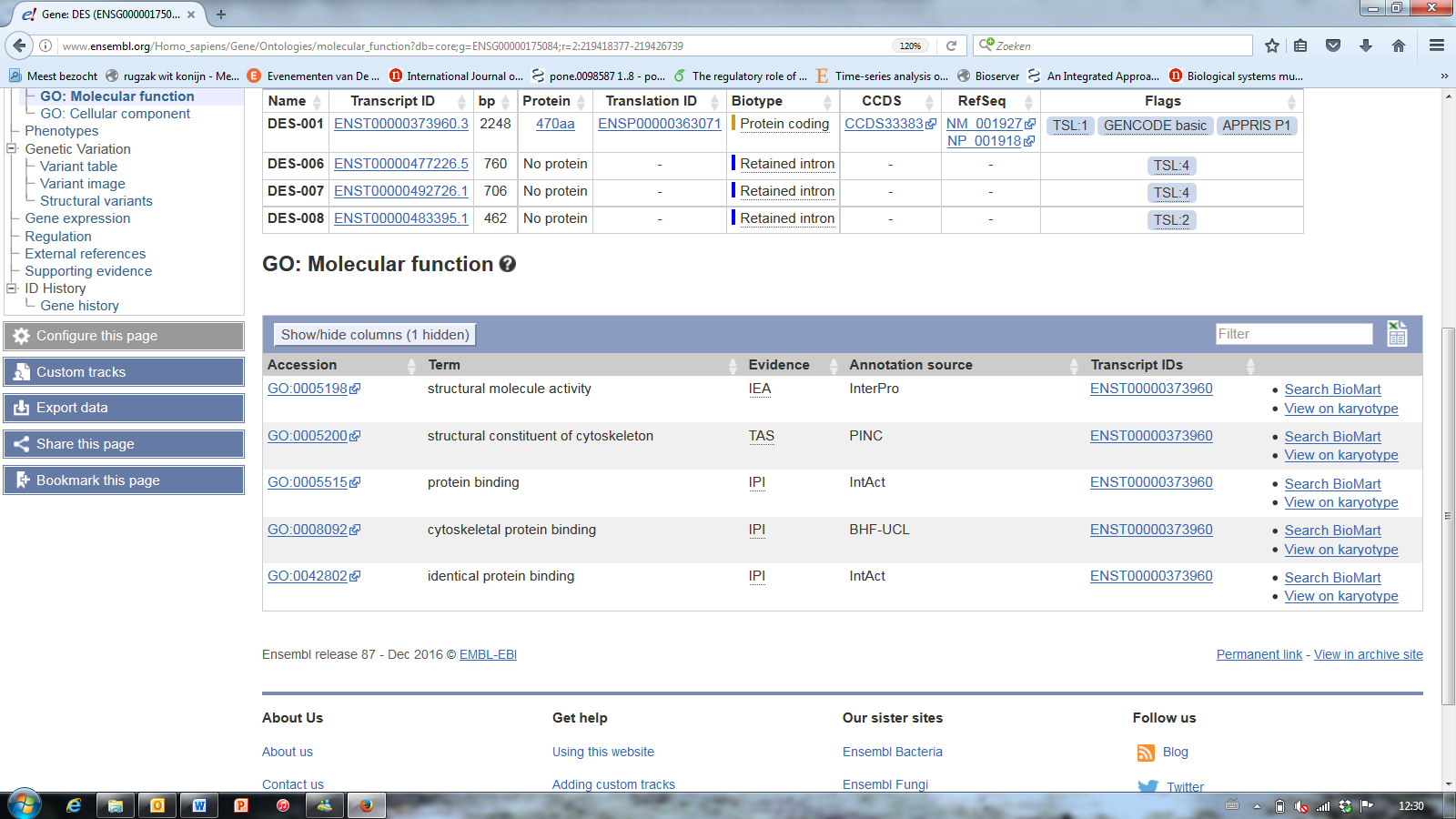
Length protein: 470 aa (=amino acids)

* 1. Gene Ontology (GO) terms give additional information on the protein. Have a look at the GO information of the gene. Give three biological processes and two molecular functions in which DES is involved.

Click on the left hand side under *Ontologies* at the dedicated GO classes.

Biological process

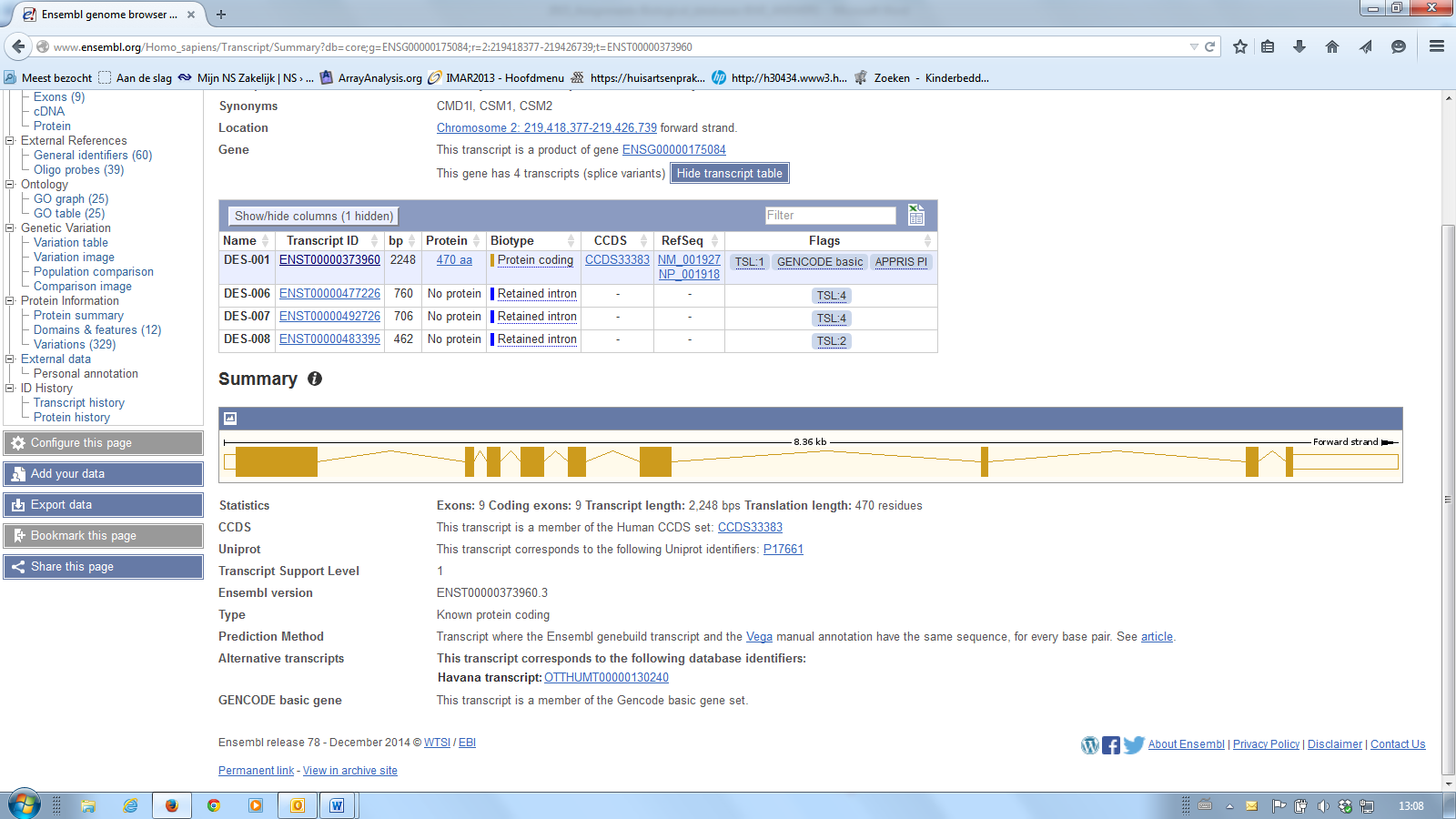


Molecular function

🖰 Click on the transcript ID of the first transcript to navigate to the Transcript tab

* 1. In many genes, the exons contain part of the open reading frame (ORF) that codes for a specific portion of the complete protein. However, the term exon is often misused to refer only to coding sequences for the final protein. This is incorrect, since many non-coding exons are known in human genes. How many exons does this transcript of DES have and how many are coding?

Nine exons, all are coding (note that the first and the last exon contain UTRs or untranslated regions, that are including in the mRNA, but not translated into protein).



***NCBI***

NCBI has a text search and retrieval system that integrates the PubMed database of biomedical literature with 39 other literature and molecular databases, including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression.

*Assignment 3: NCBI*

In assignment 2 you used the Ensembl website to gather information on the human DES gene. In this exercise you will use several NBCI-Entrez databases to find more information about this gene. NCBI offers a training and tutorial page at: <http://www.ncbi.nlm.nih.gov/guide/training-tutorials/>. You can consult the materials offered there if you need further information.

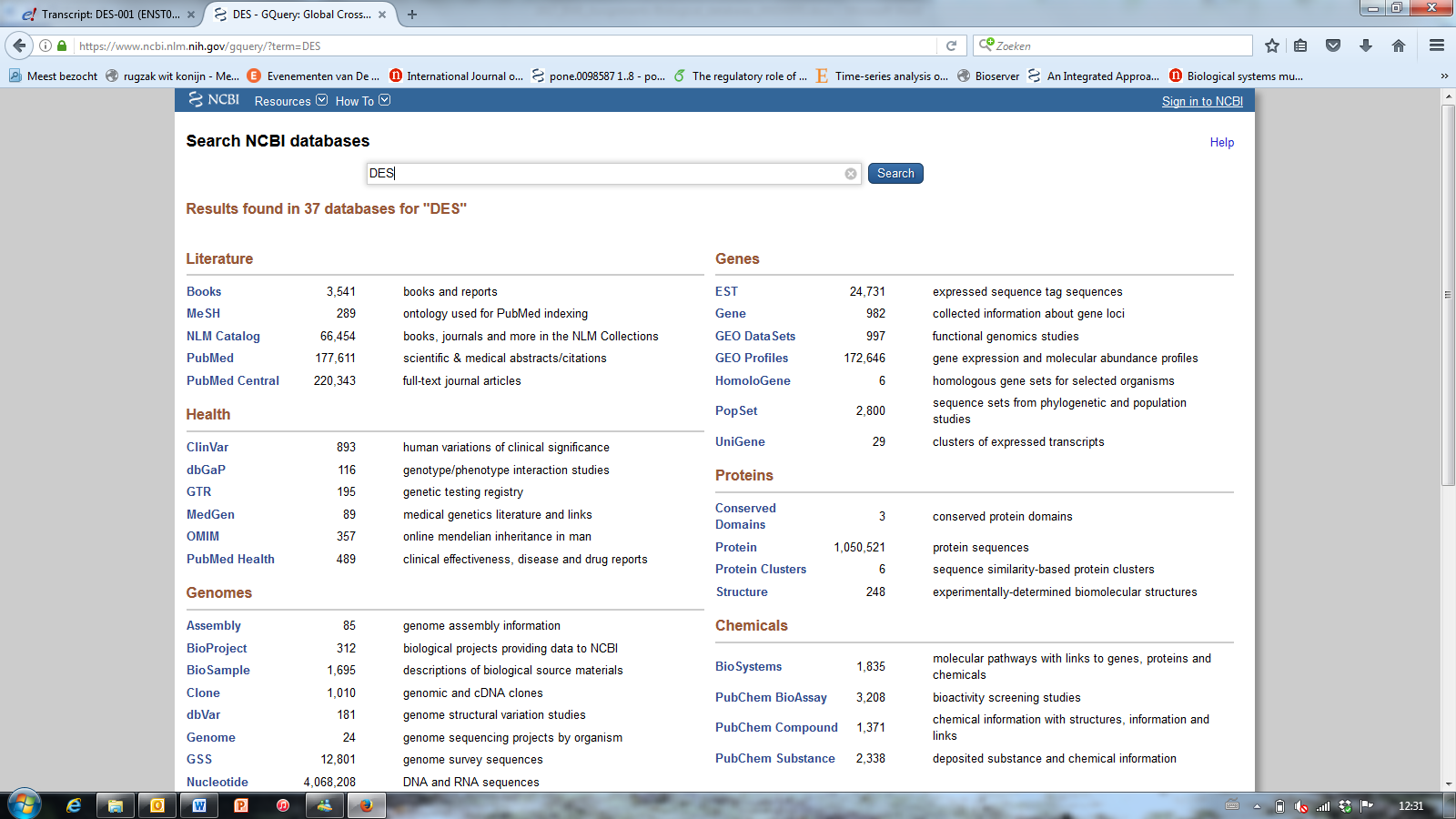
One of the databases offered on the NCBI website (<http://www.ncbi.nlm.nih.gov/>) is the *Gene database*.

! Hint: Background information on this database is available at <http://www.ncbi.nlm.nih.gov/books/NBK21085/> !

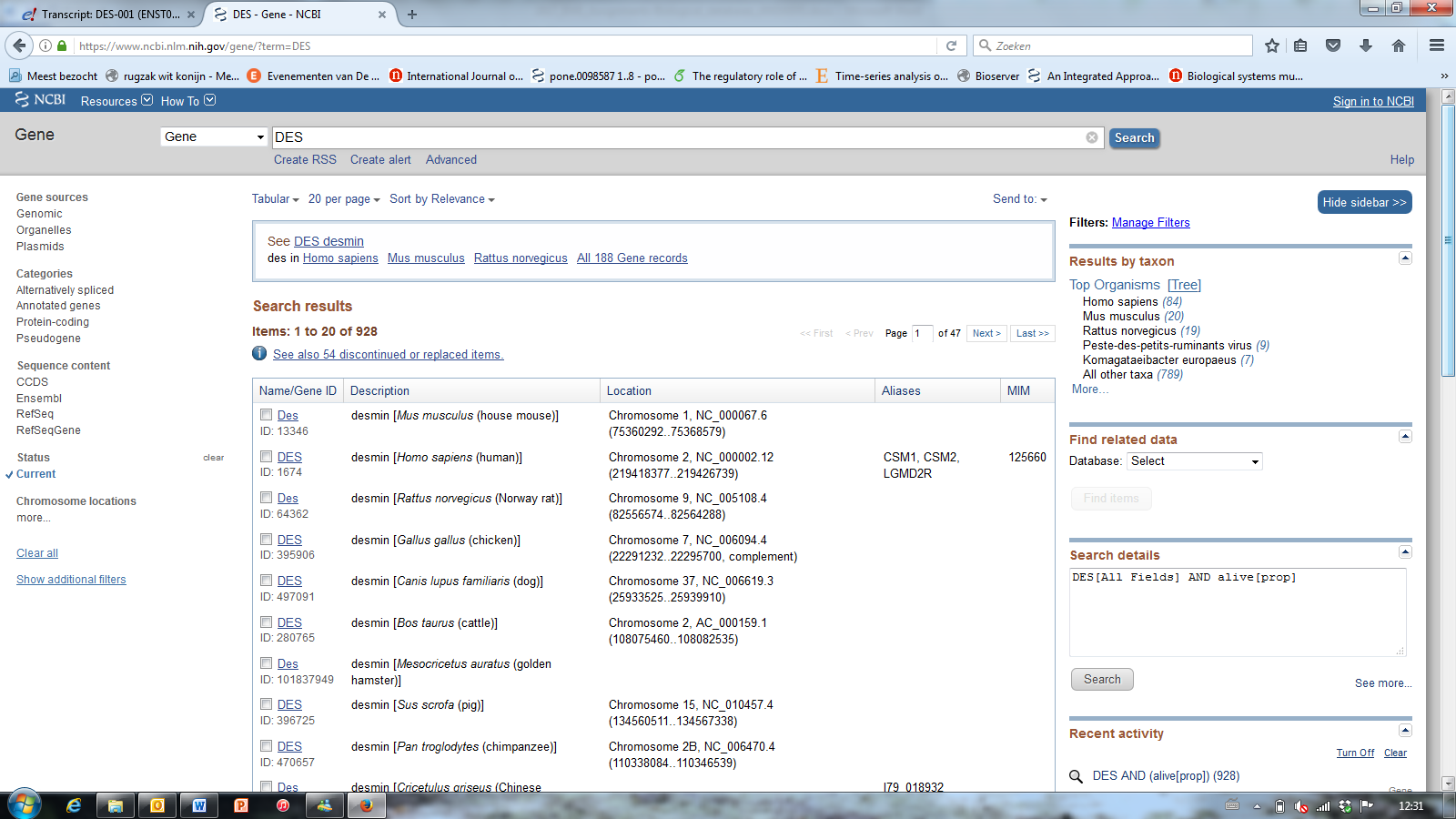
Start with opening the NCBI Gene page for human DES and answer the following questions. *Hint: apart from information given directly on this page, you may also want to use links to other pages (for example at the right side of the page).*

1. NCBI gene also uses specific identifiers. What is the NCBI Gene ID of the human DES gene?

Step 1: Search for human DES in NCBI and click at gene



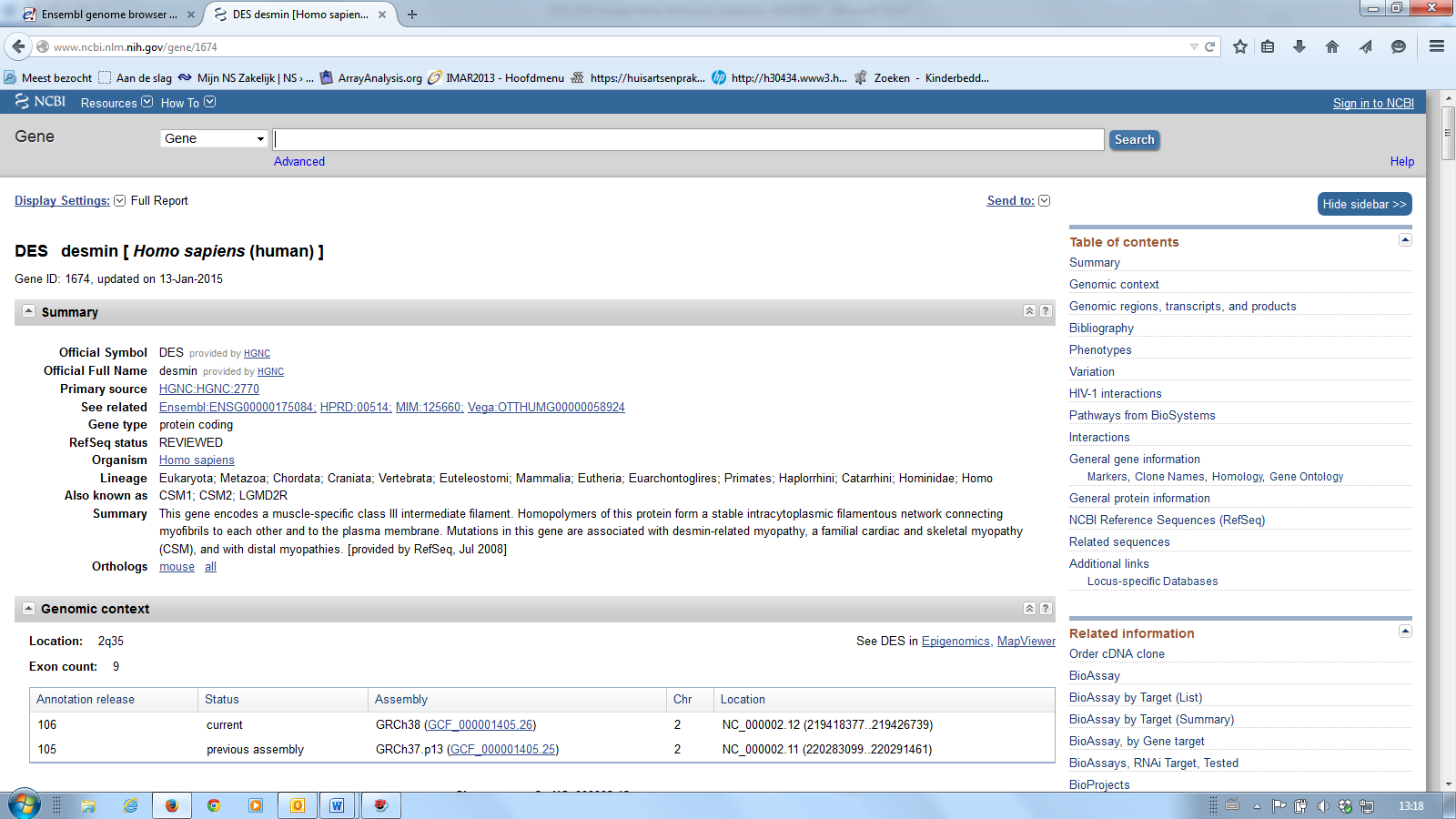
Step 2: Select the human DES gene



Step 3: In the top corner at the left you will find the NCBI Gene ID

Gene ID: 1674

http://www.ncbi.nlm.nih.gov/gene/1674



1. Human genes are given an official name by the HUGO Gene Nomenclature Committee (HGNC). However, one gene can have other names which are not official. Give two alternative names (not IDs) of the human DES gene.

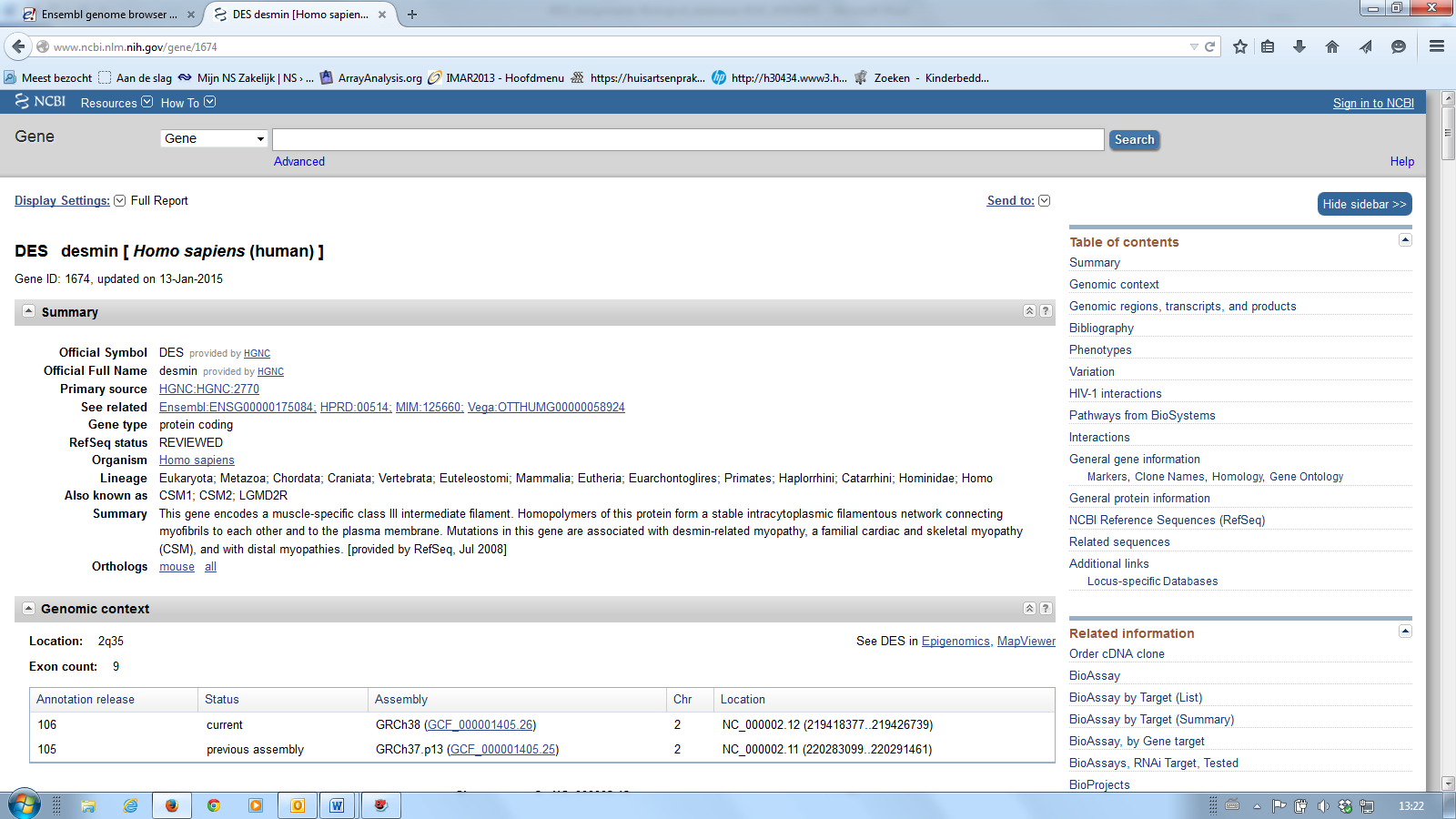
The alternative gene names can be found on the NCBI gene page of the human DES gene.

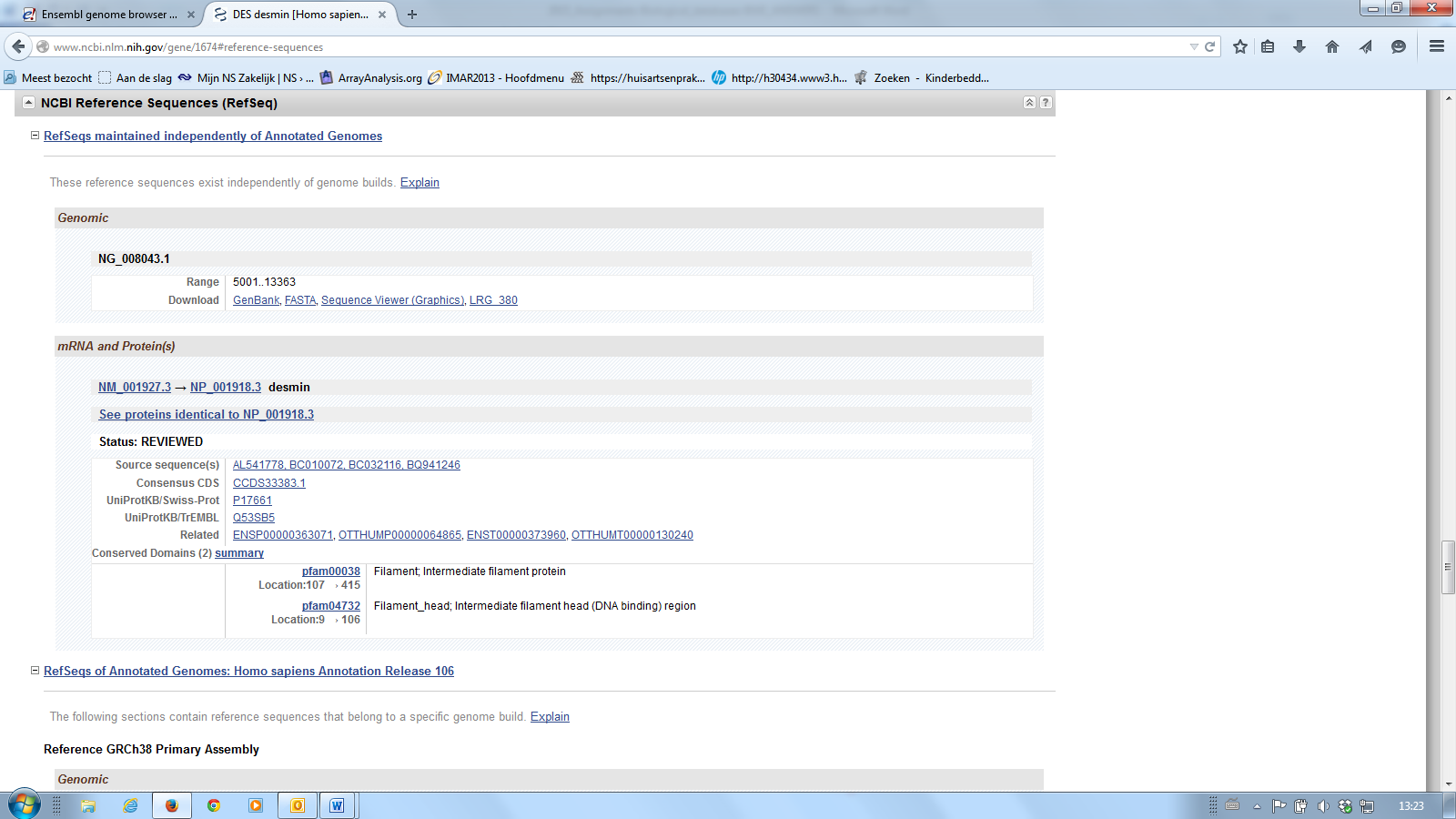
All listed alternative names (Note you only need to give two names):

CSM1; CSM2; LGMD2R

1. In the NCBI gene entry the NCBI reference sequences are given. What is the RefSeq ID of the mRNA of this gene and of the protein?

Click at “NCBI Reference Sequences (RefSeq)” on the right hand side.





mRNA: NM\_001927.3 – protein: NP\_001918.3

1. Do the RefSeq identifiers in the previous question correspond to those given by Ensembl in 2f?

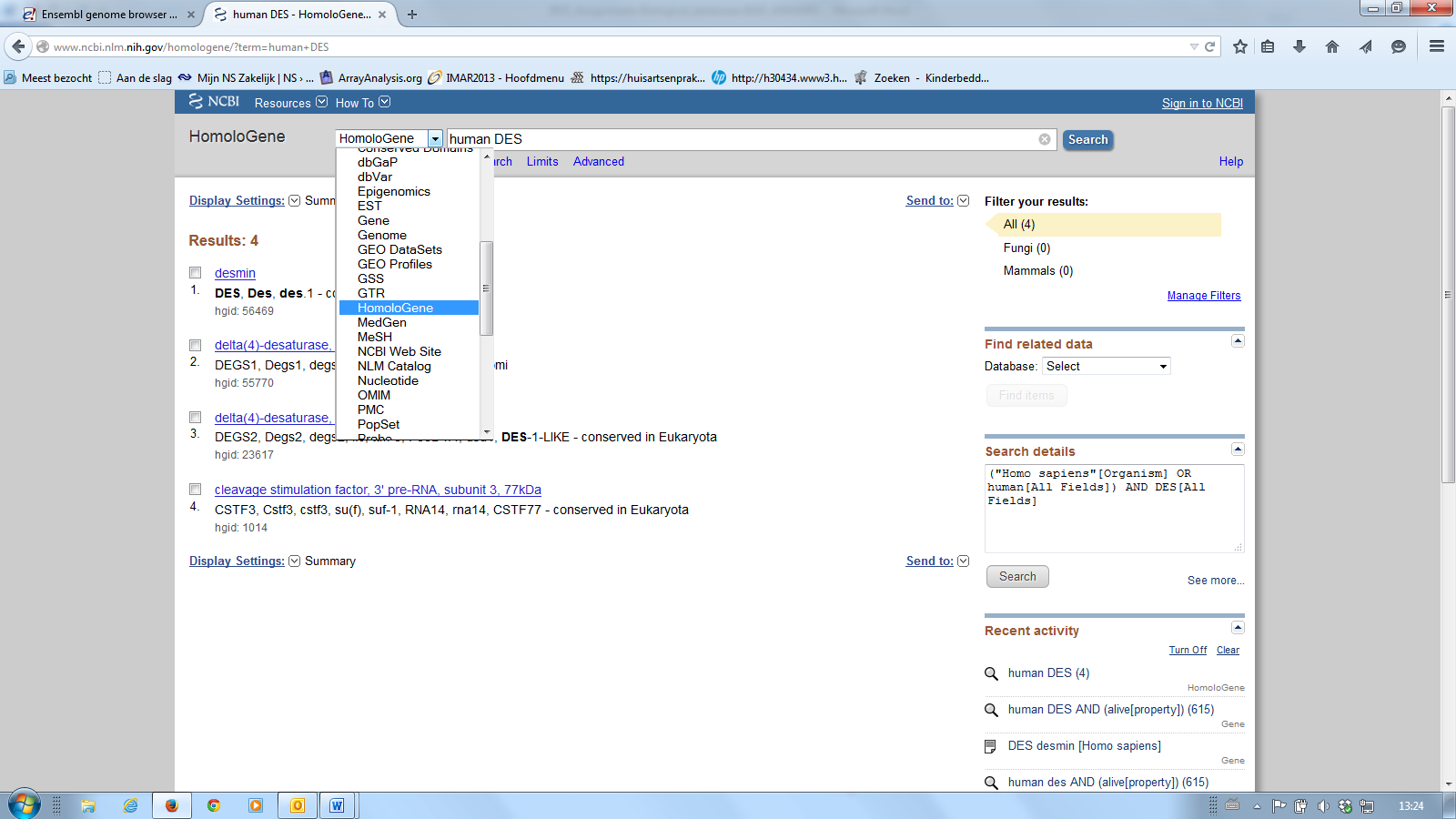
Yes they do (note the number after the dot is a version number which is increased with every change in the record – the core ID may be given without this version number)

1. What are the NCBI Gene IDs of the homologues of human DES in mouse and rat? How/where did you find this?

! Hint: Background information on the *Homologene database* is available at

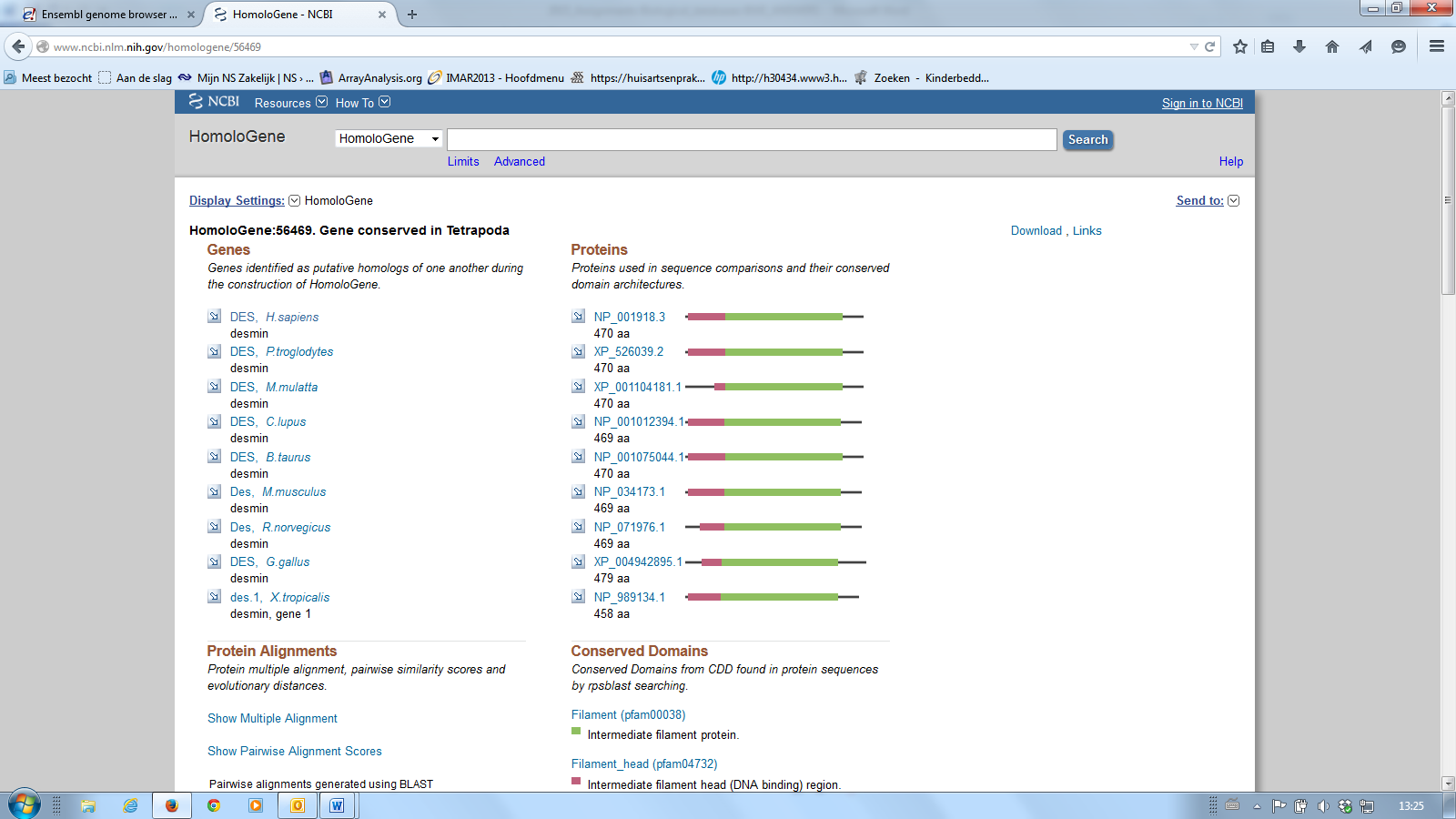
<http://www.ncbi.nlm.nih.gov/books/NBK21083/#A866> !

Use the Homologene database, see image below how to select the databases



Mouse (=*M.musculus*): 13346

Rat (=*R.norvegicus*): 64362



http://www.ncbi.nlm.nih.gov/homologene/56469

*Assignment 4: NCBI – OMIM*

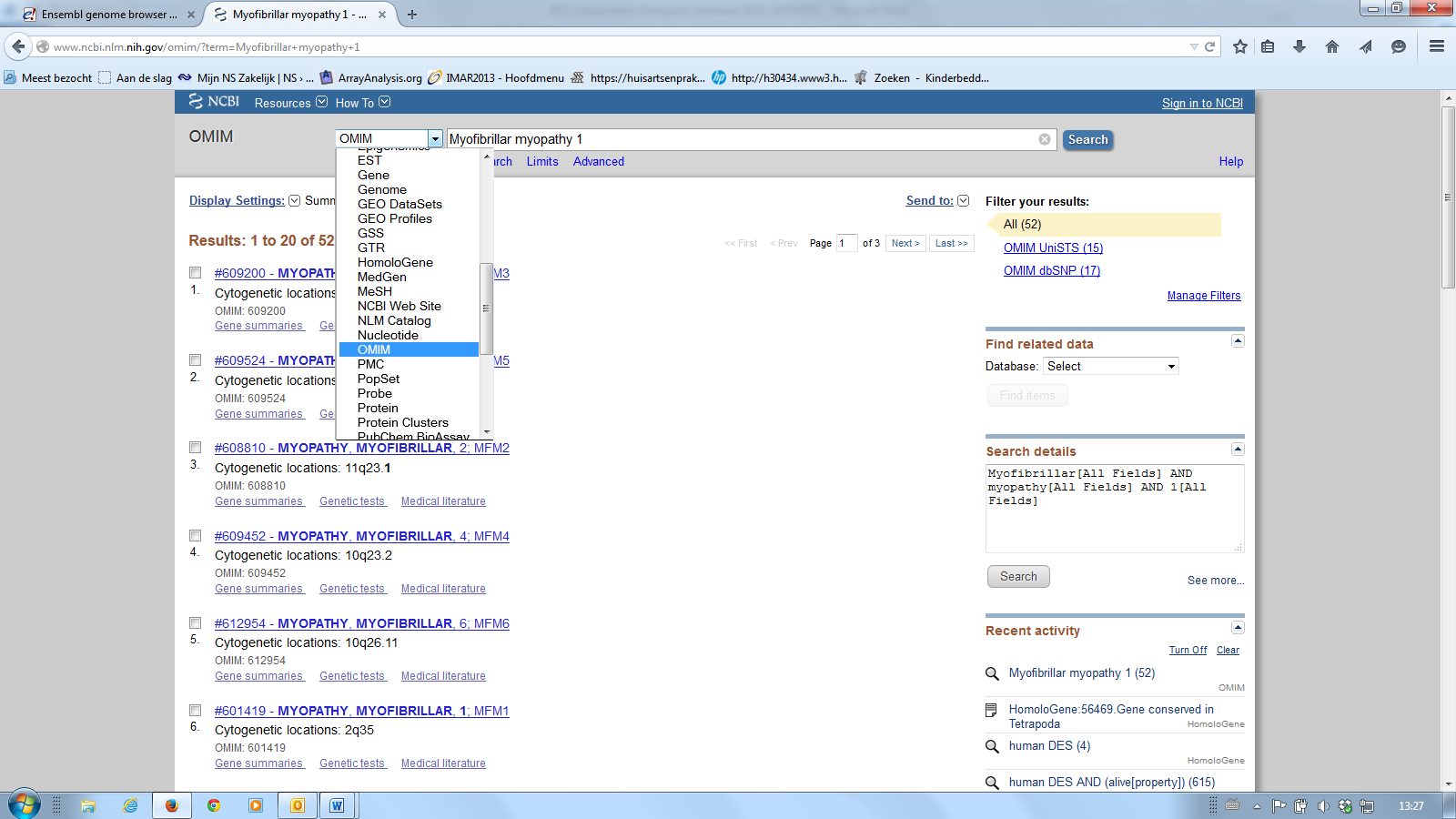
A myopathy is a muscular disease. The muscle fibers do not function properly, which results in muscular weakness. Myofibrillar myopathy (MFM) is a group of diseases of chronic neuromuscular disorders that are morphologically homogenous, but genetically heterogenous. The morphological changes in skeletal muscle in MFM result from disintegration of the sarcomeric Z-disc and the myofibrils, followed by abnormal ectopic accumulation of multiple proteins involved in the structure of Z-discs.

To get more information on Myofibrillar myopathy 1 you will use OMIM which is part of NCBI. Go to the OMIM entry of Myofibrillar myopathy 1.

!Hint: Background information on the OMIM database is available at <http://www.ncbi.nlm.nih.gov/books/NBK21108/> !

1. What is the OMIM identifier of the disease?

Select the OMIM database at NCBI and search for Myofibrillar myopathy 1.

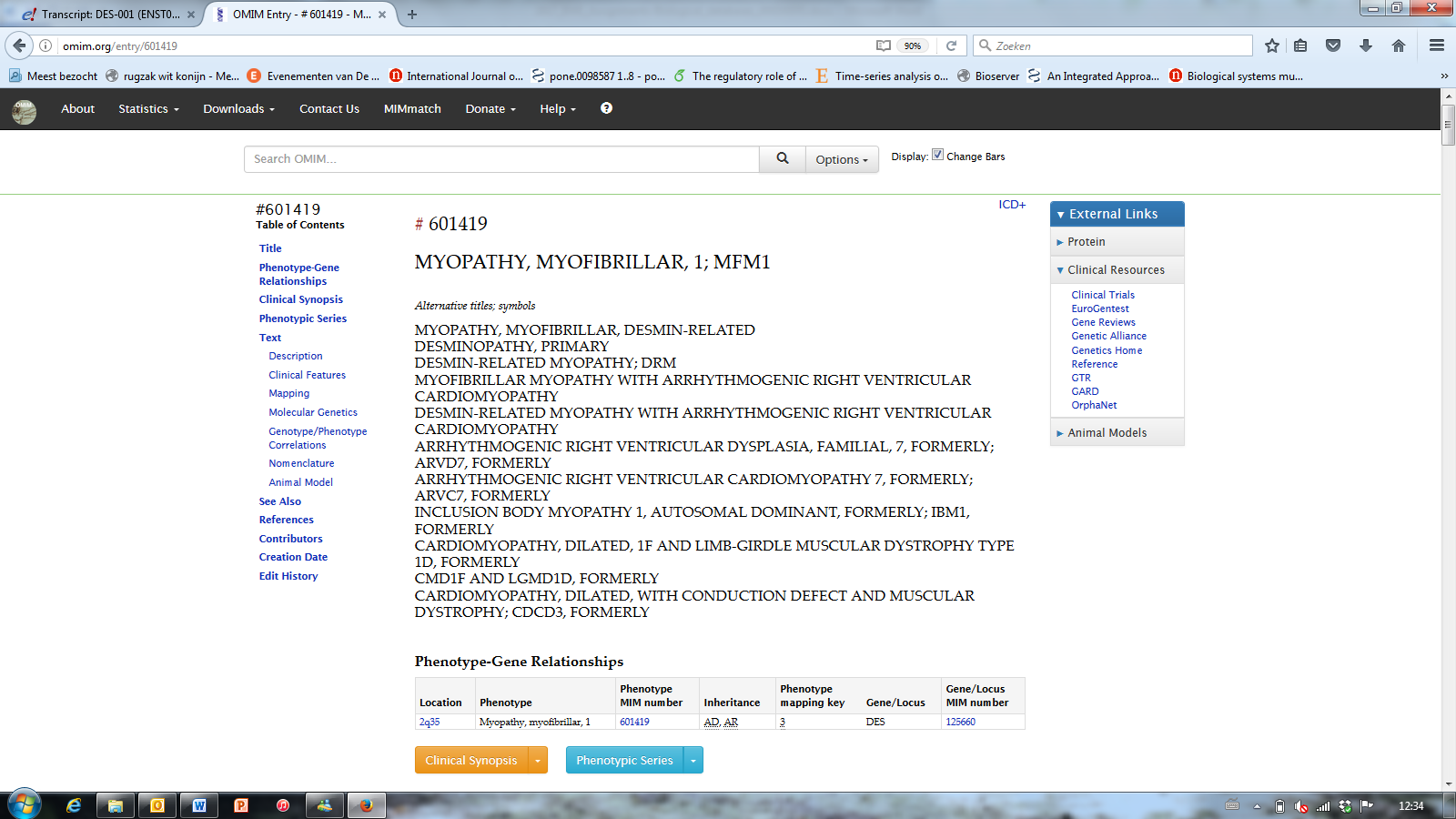


#601419 is the OMIM identifier

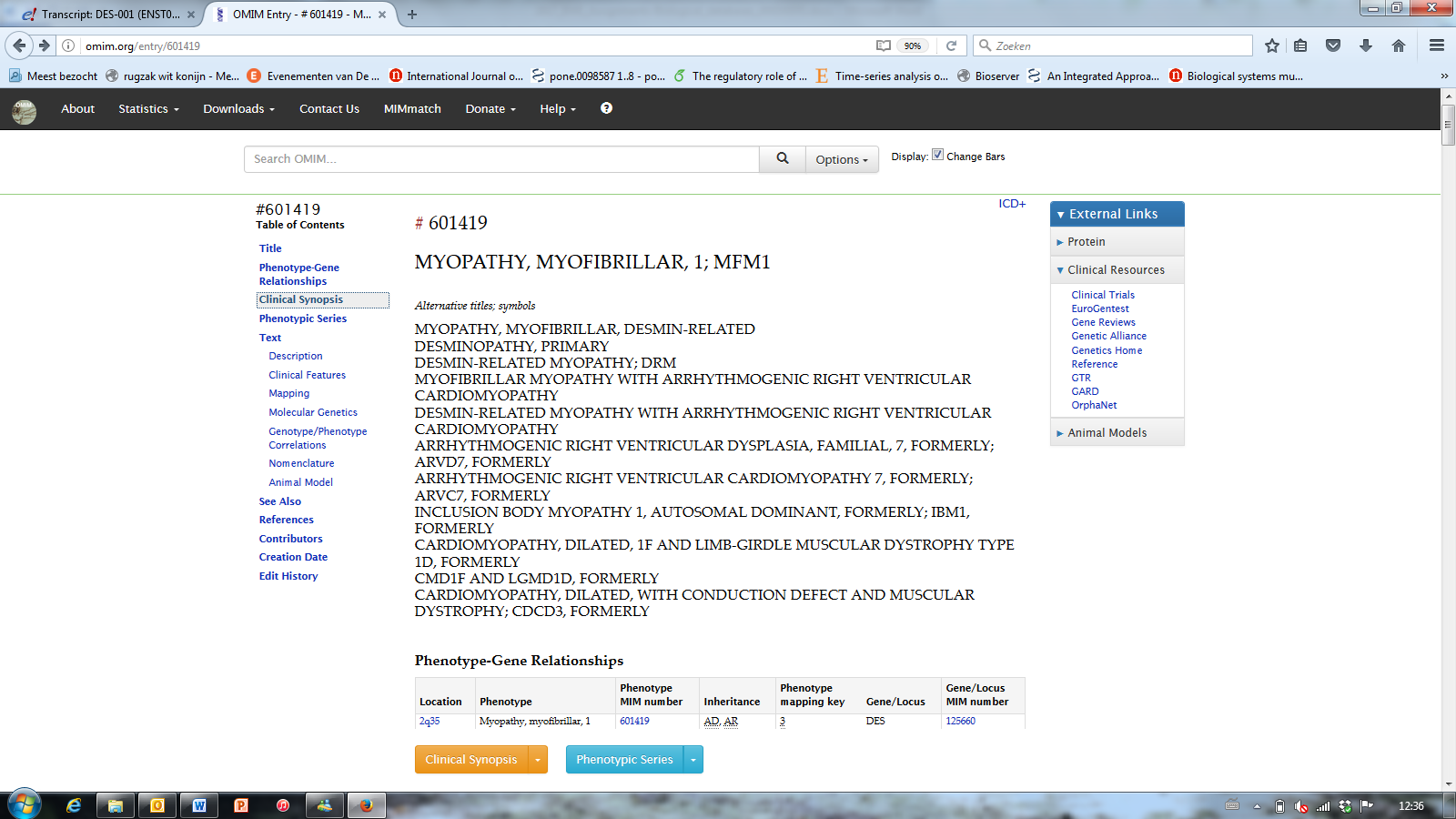
<http://omim.org/entry/601419>

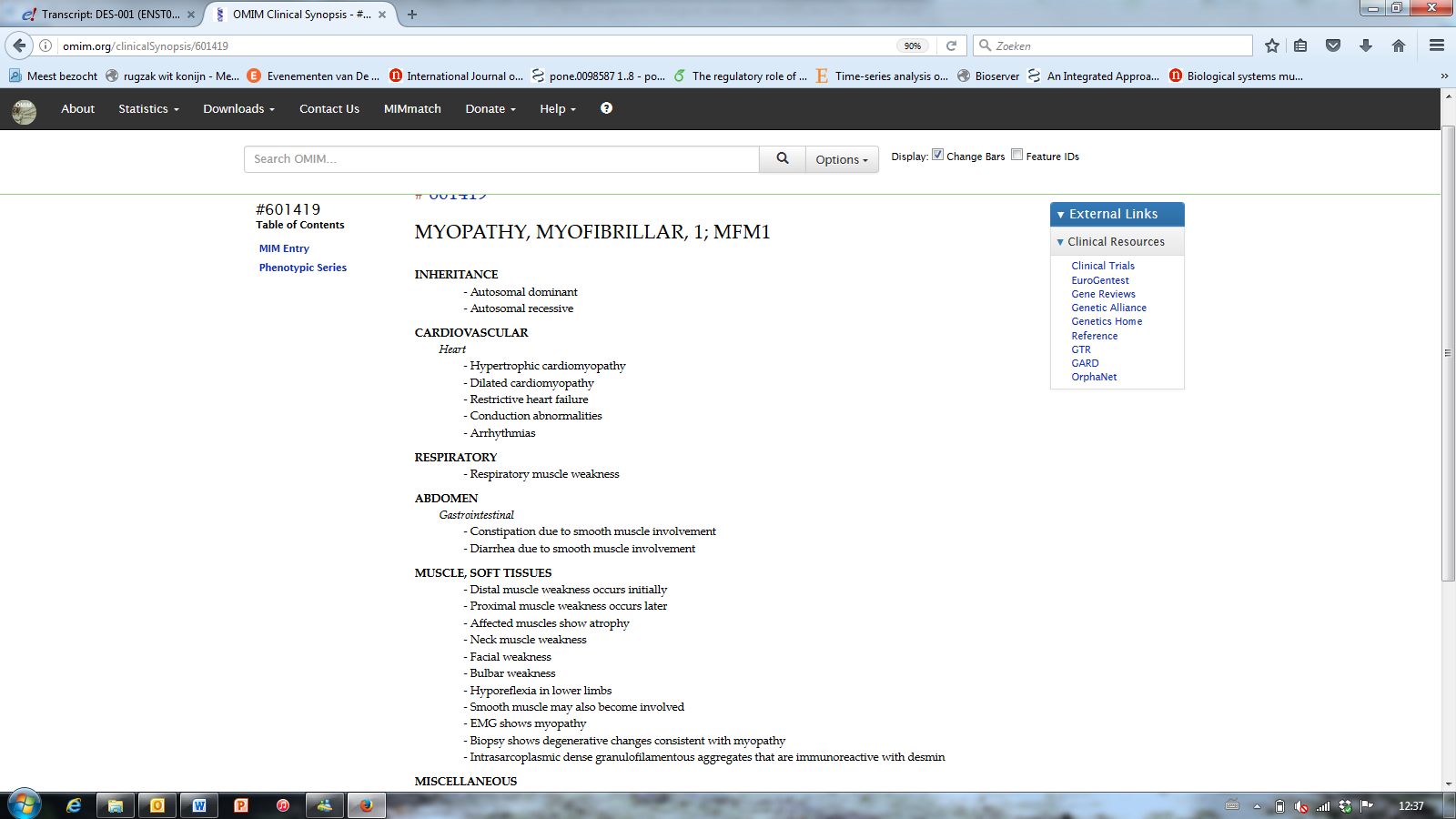
1. Which gene is associated with this disease?

DES gene



1. Go the *Clinical Synopsis* in the menu on the right hand side. A mutation in this gene seems to affect different tissues. Which tissues?



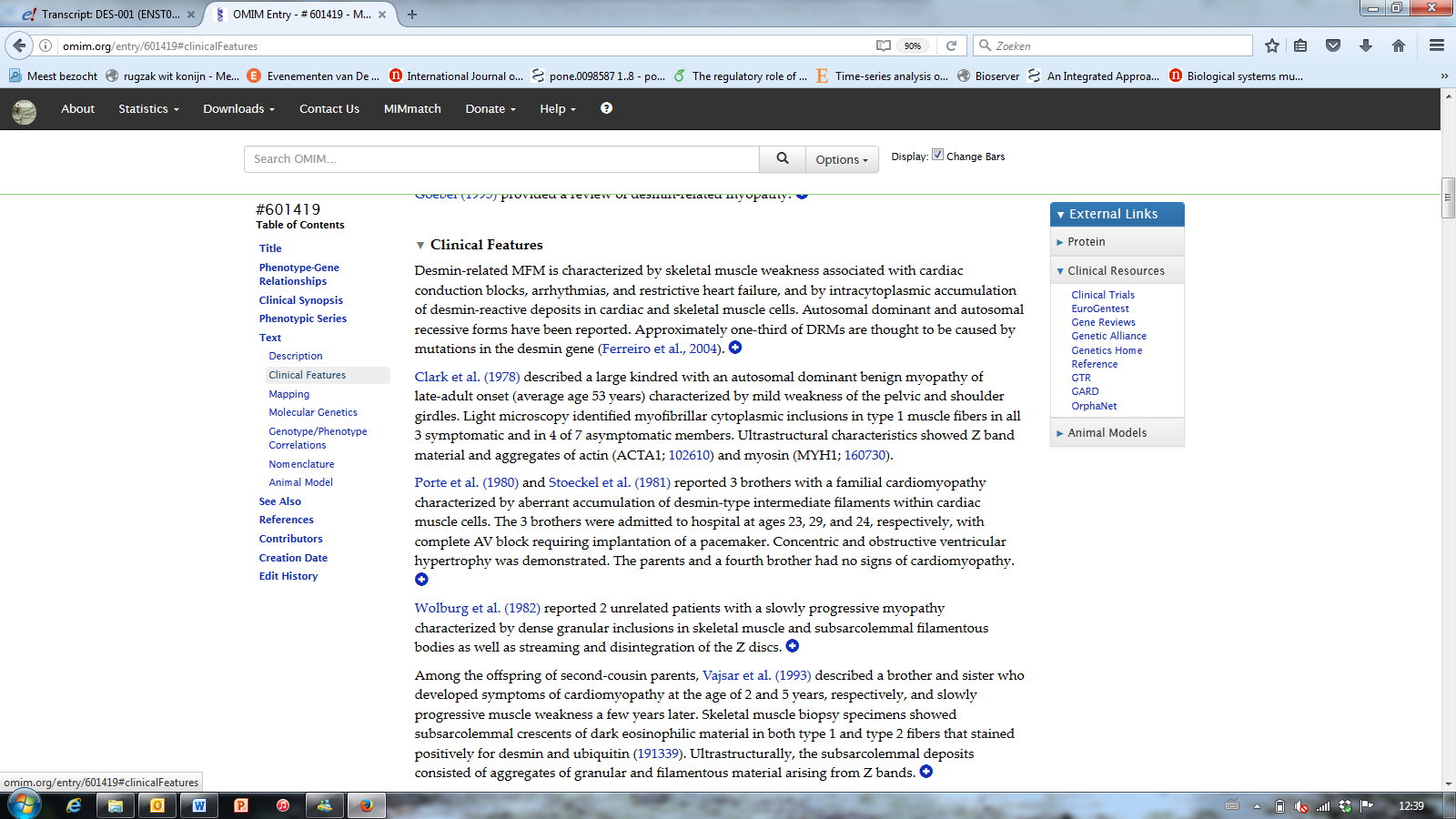


Apart from skeletal muscle, also heart muscle.

1. What are the clinical signs of Myofibrillar myopathy 1?

Hint: go the to the *clinical features*.

Desmin-related MFM is characterized by skeletal muscle weakness associated with cardiac conduction blocks, arrhythmias, and restrictive heart failure, and by intracytoplasmic accumulation of desmin-reactive deposits in cardiac and skeletal muscle cells.



*Assignment 5: Running a marathon*

# Born to run: genetic test can reveal those best able to run marathons

## A new genetic test can reveal whether runners are likely to be able to complete a marathon in a good time.



Scientists have discovered that to run a marathon in a good time requires the right combination of genes. Several studies have shown that the following “marathon” genes are significantly associated with the endurance performance:

1. AMPD1
2. COL5A1
3. PPARGC1A

To gain more insight in the structure and function of the genes you will explore Ensembl and/or NBCI. Additionally, you will investigate whether the genes are involved in inherited diseases.

1. What are the full names, Ensembl IDs and NCBI IDs of the three genes associated with endurance performance?
2. AMPD1 - adenosine monophosphate deaminase 1, ENSG00000116748, 270
3. COL5A1 - collagen, type V, alpha 1, ENSG00000130635, 1289
4. PPARGC1A - peroxisome proliferator-activated receptor gamma, coactivator 1 alpha, ENSG00000109819, 10891
5. In which GO biological processes are the three genes involved? List three GO processes per gene. Which processes are mechanisms that are known to play a role in endurance training?
6. AMPD1

purine nucleobase metabolic process, IMP salvage, purine-containing compound salvage, small molecule metabolic process, nucleobase-containing small molecule metabolic process.

1. COL5A1

blood vessel development, heart morphogenesis, cell adhesion, axon guidance, cell migration, extracellular matrix disassembly, extracellular matrix organization, collagen fibril organization, collagen catabolic process, collagen biosynthetic process, wound healing, spreading of epidermal cells, tendon development, etc.

1. PPARGC1A

temperature homeostasis, response to hypoxia, cellular glucose homeostasis, galactose metabolic process, gluconeogenesis, regulation of transcription, mRNA processing, protein complex assembly.

* AMPD1 -> energy production in skeletal muscle
* COL5A1 -> processes involved in collagen production (muscle–tendon stiffness).
* PPARGC1A -> substrate metabolism

1. Are the three genes involved in inherited diseases? If so, in which disease is each gene involved? Report the database you used to answer this question.

The database used to find the diseases: *OMIM* - Online Mendelian Inheritance in Man

1. AMPD1: Myopathy due to myoadenylate deaminase deficiency
2. COL5A1: Ehlers-Danlos syndrome, classical type.
3. PPARGC1A: no disease

*Additional information*

To select the three “marathon” genes the following two studies were used:

1. Tsianos GI, Evangelou E, Boot A, Zillikens MC, van Meurs JB, Uitterlinden AG, Ioannidis JP. [Associations of polymorphisms of eight muscle- or metabolism-related genes with performance in Mount Olympus marathon runners.](https://www.ncbi.nlm.nih.gov/pubmed/20044476) *J Appl Physiol (1985).* 2010 Mar;108(3):567-74.
2. Brown JC, Miller CJ, Posthumus M, Schwellnus MP, Collins M. [The COL5A1 gene, ultra-marathon running performance, and range of motion.](https://www.ncbi.nlm.nih.gov/pubmed/21934170) *Int J Sports Physiol Perform.* 2011 Dec;6(4):485-96.