From Cradle to grave: development, ageing and disease BBS2002

Academic year: 2018–2019

Manual Practical 3B Preparation: Use of data resources (bioinformatics)

Contact

Coordinator: Lars Eijssen, l.eijssen@maastrichtuniversity.nl

UNS50 room 1.308

Practical related to: Case 10 and Academic Project; B-ILO2002.1-5

Preparation

- Read the information in this manual and perform the preparatory assignments; these should take about 2 hours.
- Bring the answers to the preparatory assignments with you to the practical, either digitally or printed. You need these for the practical training.

Introduction

With these practical sessions, you will learn how to use online resources to investigate a disease. Besides textbooks and papers, databases nowadays are important sources of up-to-date information, from which you can already benefit as study materials in the further courses of the bachelor program. Several resources will be addressed during this practical: first we will revisit the genomic database Ensembl, and then check out a protein database (UniProt).

Aim

Get acquainted with online resources to find out more about diseases, their relationship to molecular pathways and the function of involved proteins.

Procedure

Practical 3B: At the start of the Academic Project, a second computer practical is organized, for which preparatory assignments should be completed before the start of the practical. The first part of the preparatory assignment consists out of individual exercises, to extend your knowledge on the previously covered databases and to add one new database; the second part of the preparatory assignment should be made with your group from the Academic Project, to apply the previously studied databases to the disease you selected for the Academic Project. At the start of the practical, there will be another short quiz to check your preparation. After the quiz, a short lecture with demonstration will guide you through the last database. . All answers will be made available after all separate groups have completed the practical. Afterward, you and your project group will apply your acquired knowledge and skills to find more information on the disease you are studying for the Academic Project. Instructors are present to assist and answer questions on the use of the databases. The information you find during this last practical can be used as an integral part of the academic project. **Be aware that the knowledge you gain from this practical is part of the exam material!**

Practical is signed off with a pass if:

- Practical 3B was attended.
- Student showed active participation.

Literature:

- You can use the “Reference guide to online resources” (see Student Portal > Practicals > Practical 3) for more detailed information on the functionalities of the online databases.

Appendices:

Practical 3B Preparatory assignment

Use of online data resources for molecular biology

1. During this preparation for this practical, you need to look at a) databases related to proteins and their function; b) fill in a workflow for the disease you will study for the Academic project.
    a. In case 10, you studied a form of cancer, which affects the large intestine (colon) and rectum, Familial Adenomatous Polyposis Coli (FAP). FAP is an inherited disease with a classical autosomal inheritance pattern, suggesting a monogenetic cause. The APC gene has been associated to this disease. During the preparation assignments, you will study the function of the protein which is encoded by the APC gene, and the effect of mutation in the APC gene on the function and structure of the protein. When you get stuck trying to answer the preparatory questions, use the "*Reference guide online resources"* that was provided to you.
    b. Fill in the workflow sheet, which you can use during the practical to study your disease in more detail, with means of biological databases.
2. You do not have to hand in your answers on paper, however you have to bring your answers with you digitally or on paper, which is needed for the rest of the practical training.
3. We will test how well you prepared with a short quiz at the start of the practical.
4. After the practical, the answers to the preparatory assignments will be uploaded on the Student Portal (as well as the answers to the quiz).

Assignment 1: genetic variants in the APC gene in Ensembl *(40 minutes)*

▶ *Sections 3.1 till 3.4 of the Reference guide.*

Look up SNP rs137854573 in Ensembl.

a) Which allelic change and position in transcript ENST00000257430 and protein ENSP00000257430 are given by dbSNP?

C to T change at position 1660 in the transcript; arginin (Arg) to stop codon (Ter) at position 554 in the protein.

b) Why are there so many names given for the SNP at the Ensembl page (sequence identifiers and positions)?

Because the position (and possibly change) depend on the splice variant or sequence given.

c) The APC gene is encoded by the forward strand. What would happen to the chromosomal alleles given for the SNP (the NC_ code) when the gene would have been on the reverse strand?

They would be complimentary to the SNPs as given by the mRNA. So for a C -> T change in the mRNA, the change on genomic level would be given as G -> A (as this is always based on the reference genome, which gives the sequence of the forward strand).

Open the phenotype data section.

d) To which phenotypes has the SNP been associated according to Ensembl?

Susceptibility to colorectal cancer; Familial Adenomatous Polyposis 1; Hereditary cancer-predisposing syndrome.

Open the Genes and regulation section.

e) To which gene(s) is the SNP associated and what is/are the Ensembl identifier(s)?

Only to the APC gene; its Ensembl identifier is ENSG00000134982.

Now open the page of the APC gene in Ensembl. From there it is possible to open a table with all SNPs linked to the gene, by opening the Variant table.

f) Which types of SNPs are given by Ensembl and which colour code is used?

- upstream gene variant (grey)
- 5 prime UTR variant (pale green-blue)
- intron variant (blue)
- splice region variant (orange)
- splice donor variant (orange)
- coding sequence variant (green)
- missense variant (yellow)
- synonymous variant (bright green)
- stop gained (red)
- frameshift variant (purple)
- inframe insertion (pink)
- inframe deletion (pink)
- 3 prima UTR variant (pale green-blue)
- downstream gene variant (grey)
- (and some others)

(Note: you can see the classes more easily, by clicking the ‘Consequences’ button at the top of the table)

Using the ‘Consequences’ button at the top of the table, select only truncating (PTV) and missense variants to filter the table. Then sort the table by rs number, by clicking the column header. Find the C/G SNP with identifier rs577466163.

g) Why is the SNP present in the table more than once?

Because it is given once for each transcript (splice variant).

h) What are the SIFT and PolyPhen scores of this SNP? What do these scores mean? How are they computed and why are they only given for missense SNPs?

They range between 0.01 and 0.05 for SIFT, and between 0.987 and 0.999 for PolyPhen. Both indicate a ‘deleterious’ or ‘probably damaging’ consequence on the protein. SIFT and PolyPhen are tools to predict the effect of amino acid substitutions on the functionality of a protein. SIFT does this by looking at conservation of the residue only, Polyphen adds annotation such as domains, and if available 3D structural information to this. They are only given for missense SNPs, as for others no predictions can be made by these tools.

Assignment 2: the APC protein in UniProt *(50 minutes)*

▶ *Section 6.1 of the Reference guide.*

Now we turn to the UniProtdatabase, to find more information about the function and structure of the protein encoded by the APC gene.

First, search the human APC protein in the UniProt database.

a) What is the UniProt identifier of the human APC gene?

P25054

The page starts off with a nice concise description of the function of the protein, supported by references (verify this yourself). Then for APC, some generic information on the types and positions of the mutations is given.

b) Does the information on mutations given by UniProt correspond to what you have seen in NCBI/Ensembl?

Yes, it does (many truncating, clustered in mutations cluster region or MCR).

Further on, information on the GO annotations of the protein is given.

c) What is the evidence type associated with the annotation to cell cycle arrest? (Hint: click the Source to find out)

It is 'Inferred from direct assay' (click some others to get an idea which types we can have).

d) Which general subcellular locations are given by UniProt annotation and which by GO annotation?

UniProt: plasma membrane, cytoskeleton.
GO: Cytoskeleton, Cytosol, Nucleus, Plasma membrane.

Then we find an overview of involvement of mutations in the protein in human disease.

e) Where in the sequence do we find the most disease-causing mutations?

In the part from about one third to one half of the sequence.

After the disease-associated mutations, a section is presented on processing of the protein (*e.g.* cleaving, amino acid modifications).

f) Which is the most common modification present for this protein? What is the biological function of this modification?

   Phosphorylation; this is commonly used for activation of the protein; in this case we can also read (below the table) that the GSK3B protein is known to phosphorylate ACP.

Also, scrolling further down, information on interaction partners of the protein is given.

g) With which protein does APC most strongly interact? Which other interacting protein do this protein and APC have in common?

   Strong interaction with CTNB1; AMER1 as common interactor.

Finally, we will have a look at the structural information given in UniProt.

h) Which secondary protein structural elements do exist (in general)?

   The most common ones are: alpha helices, beta strands/sheets (a sheet consists of multiple parallel or anti-parallel strands), and turns.

i) Which secondary structural element is most common for the APC protein, and where in the protein does is mostly occur?

   The helix is by far the most common; helices are present in the first one third of the protein.

j) Which methods have been used to experimentally determine 3D structure of (part of the) protein? What is the common range of the resolutions? Hint: check the list of structures given.

   Either X-ray crystallography or NMR spectroscopy; resolution commonly ranges from 1.5 to 3.1 Angstrom (estimate based on the structures available for APC).

Assignment 3: Database workflow for your own disease *(30 minutes)*

When you want to study a disease in more detail, there are several steps that you can follow. You could start with books, articles, Wikipedia etc. Hopefully, you are now also thinking about using biological databases as a starting point, or as a method to go in more detail on a gene or protein you found to be linked to your disease. Below, a workflow has been provided with general steps that you could follow to navigate your way through the available information online. This workflow will serve as your guideline for Practical 2B, as well as the Academic Project.

For the preparatory assignment, you have to perform the following steps:

1. Fill in the name of your disease and the table sections of this workflow (for part 1-4). Use the knowledge you gained from the preparatory assignments and practical 3A+3B, to fill in which databases and type of information you would like to address for your disease.
2. Bring a digital/printed version with you to the practical.

We advise you to start filling in the first step (Disease) and second step (Molecular Mechanism) with the whole group of the Academic project (under the part titled “Relevant information for my disease”) before the practical.

Practical training:

During the second part of the practical (after the quiz + demo), you could split up the third step (genes) and fourth step (proteins) between the team members, and investigate one gene/protein in-depth per student, using the databases discussed during practical 3A and 3B.

1. Disease: ______________________

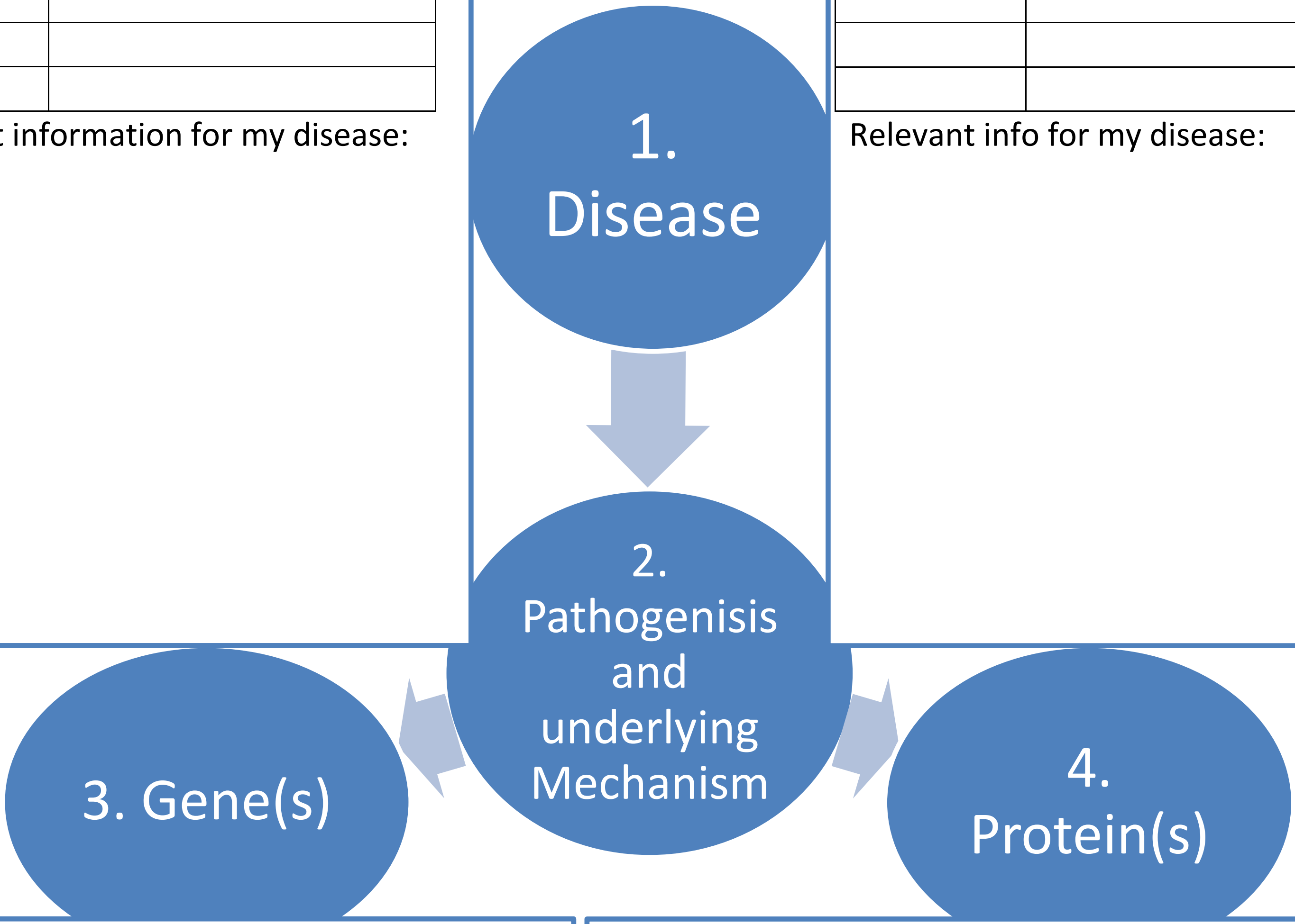
Databases	Type of info

Relevant information for my disease:

2. Molecular Mechanism:

Databases	Type of info

Relevant info for my disease:



1. Disease

2. Pathogenisis and underlying Mechanism

3. Gene(s)

4. Protein(s)

3. Gene(s): ___________________________(from 2)

Databases	Type of info

Relevant information for my disease:

4. Protein(s):___________________________(from 2)

Databases	Type of info

Relevant information for my disease:

Quiz questions

Practical 3B Use of online data resources for molecular biology

* Questions 6 and 8 are directly related to the preparatory assignment. Students should get some time to either go through their notes, or look up the information in the given database.

LEVEL 1 questions (related to preparation):

1. What is SIFT?

    a. SIFT is a tool to visualise the location of SNPs in a gene.
    b. SIFT is a tools to predict the effect of mutations on the functionality of a protein.
    c. SIFT is a tool that allows you to compare phenotypes between diseases.
    d. SIFT is a tool that estimates the genotype frequencies in different populations.

2. Which one of the following could be a Uniprot identifier?

    a. P25054.
    b. NM_000193.3.
    c. GO:0043237.
    d. SHH.

3. Which secondary structural element is most common for the APC protein (Uniprot ID: P25054)?

    a. Parallel beta sheet.
    b. Alpha helix.
    c. Antiparallel beta sheet.
    d. Hydrogen bonds.

4. Which of the following databases is a resource for biological pathways?

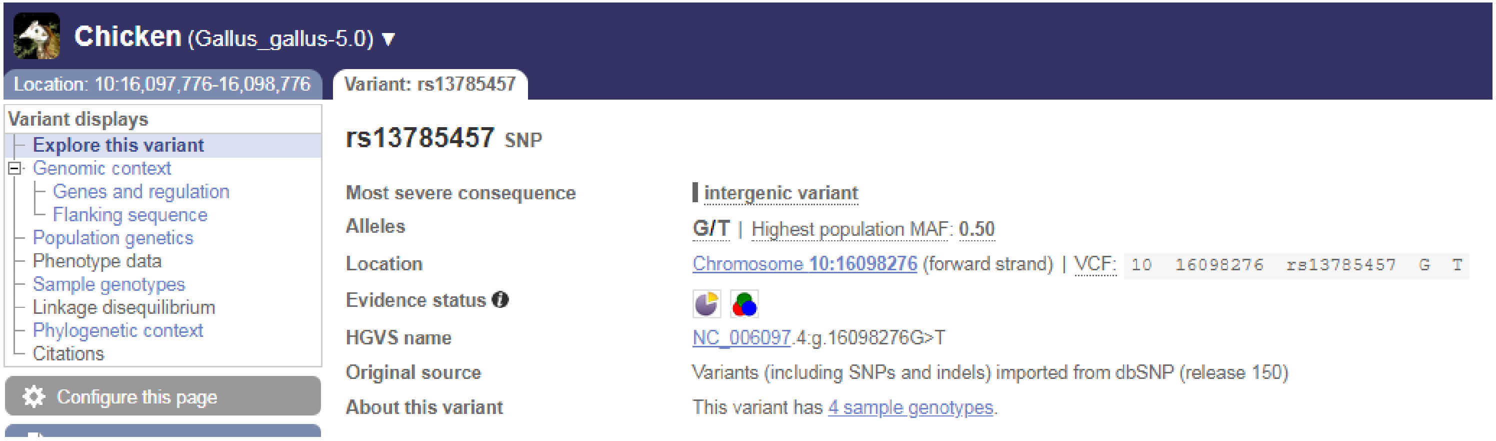
    a. ChEBI.
    b. Reactome.
    c. HMDB.
    d. Ensembl.

5. What does the NR_ stand for in a RefSeq identifier? That the identifier is related to:

    a. An incomplete genomic region.
    b. A predicted mRNA model.
    c. A non-coding RNA region.
    d. A messenger RNA region.

LEVEL 2 questions (related to biological interpretation):

6. What information can you derive from the following Figure?



rs13785457 SNP

Most severe consequence	▌intergenic variant
Alleles	G/T | Highest population MAF: 0.50
Location	Chromosome 10:16098276 (forward strand) | VCF: `10 16098276 rs13785457 G T`
Evidence status	
HGVS name	NC_006097.4:g.16098276G>T
Original source	Variants (including SNPs and indels) imported from dbSNP (release 150)
About this variant	This variant has 4 sample genotypes.

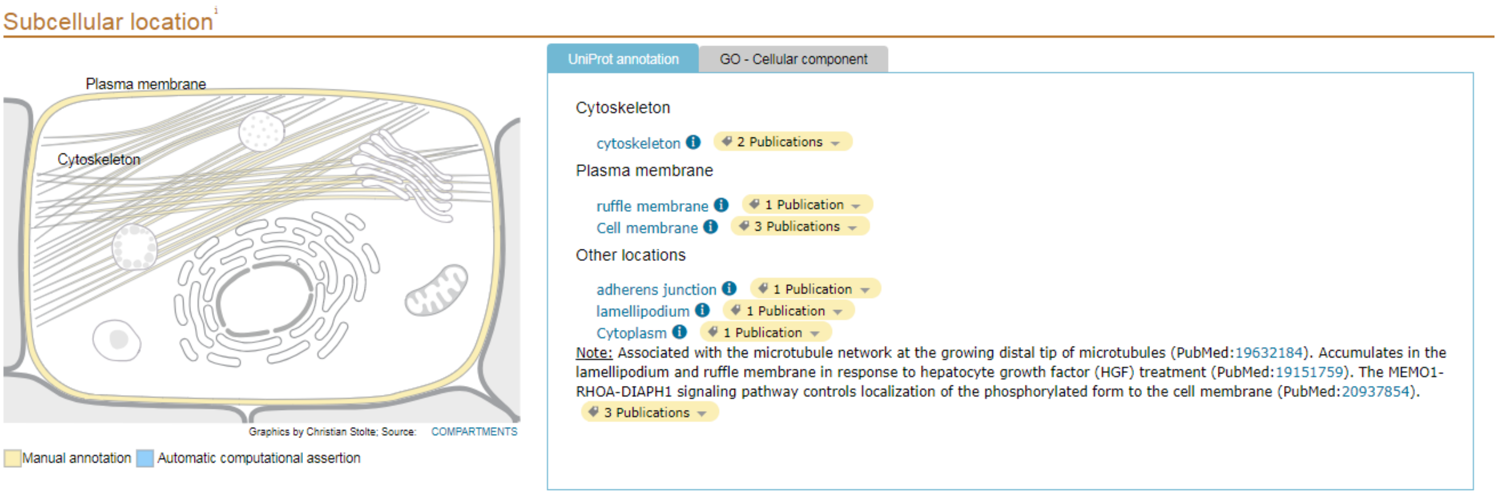
a. This SNP is related to a mutation in humans.

b. <u>That the mutation changed a G to a T.</u>

c. That the mutation is located on the backward strand.

d. That the mutation results in a STOP codon.

7. What information can you derive from the following Figure?



Cytoskeleton

- cytoskeleton 🛈 2 Publications

Plasma membrane

- ruffle membrane 🛈 1 Publication
- Cell membrane 🛈 3 Publications

Other locations

- adherens junction 🛈 1 Publication
- lamellipodium 🛈 1 Publication
- Cytoplasm 🛈 1 Publication

Note: Associated with the microtubule network at the growing distal tip of microtubules (PubMed:19632184). Accumulates in the lamellipodium and ruffle membrane in response to hepatocyte growth factor (HGF) treatment (PubMed:19151759). The MEMO1-RHOA-DIAPH1 signaling pathway controls localization of the phosphorylated form to the cell membrane (PubMed:20937854). 3 Publications

a. That there are 3 publications supporting the presence of the protein under investigation in the plasma membrane.

b. That the protein under investigation is located in the mitochondria.

c. That the protein under investigation is located only in the cytoskeleton, plasma membrane and mitochondria.

d. <u>That treatment with HGF increases the protein to accumulate.</u>