From Cradle to grave: development, ageing and disease BBS2002

Academic year: 2018–2019

Manual Practical training 3: Use of data resources (bioinformatics)

Contact

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

UNS50 room 1.308

Practical related to: Case 2; 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 one disease database (OMIM), second a pathway database (WikiPathways), and last one genomic databases (Ensembl.

Aim

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

Procedure

Practical 3A: The full practical will take place in a computer lab. In order to effectively participate in this practical, you should first complete the preparatory assignments which will cover three databases: OMIM, WikiPathways and Ensembl. You do not have to hand in the answers, however we will check your preparation by means of a short quiz (results will be stored in the Grade Center; this is also a way for yourself to check if your preparation was sufficient). After the quiz, an interactive lecture with demonstration of various databases will take place. Finally, individual exercises have to be made during the second part of the practical for hands-on work to put the preparatory assignment and demo in context; instructors are available for guidance. All answers will be made available after all separate groups have completed the practical. During practical 3A, we will explore information in molecular biological databases, taking examples from the myelination pathway. The aim of this practical is to get you more familiar with these databases and the information you can find. The assignments of this practical cannot target detailed biological interpretation, but we strongly advise you to take these resources into account when preparing for the next cases (and furthermore throughout your study). **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 3A 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 3A Preparatory assignment

Use of online data resources for molecular biology

1. In case 2, you studied how myelin sheets form a protective layer for neurons. Building these sheets is a complex process, in which many molecular interactions are involved. Databases provide a reliable source of up-to-date information and can complement knowledge in text books and other literature sources.
2. There are four preparatory assignments, which we expect you to prepare before the practical. When you get stuck in trying to answer the preparatory questions, please use the "*Reference guide online resources"* that was provided to you for tips. 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 questions you will work on during the practical).

Assignment 1: Important biological databases *(30 minutes)*

Watch the video on "Important Biological Databases" posted on YouTube (available at https://www.youtube.com/watch?v=WrUcxVJOwL0). Answer the questions at the end of the video, which are also described below. A suitable book for further reading is "Lacroix, Z., & Critchlow, T. (2003). Bioinformatics : Managing scientific data . San Francisco." Available as e-book from the Maastricht University Library (this reference is not mandatory reading material for the preparatory assignment).

Questions related to video:

a) Explain the relevance of Biological Databases in Bioinformatics.
b) Briefly describe the different types of Sequence Databases.
c) Describe Entrez Search Engine.
d) Explain the Primary Nucleotide sequence database.

Assignment 2: genetic variants in the SHH gene in OMIM *(30 minutes)*

"The SHH gene encodes sonic hedgehog, a secreted protein that is involved in establishing cell fates at several points during development." [OMIM: 600725]. There are several diseases associated with this gene.

Look up the SHH gene in OMIM (https://omim.org/, a database at NCBI, which contains expert-curated information on the links between genes and genetic variants and human disease).

a) To which diseases has the gene been linked, which identifiers are mentioned for these diseases (displayed in the table as "Phenotype MIM number") and with which inheritance pattern (if known)?

   Four diseases are mentioned under the OMIM:600725 website (accessed 2018-06-13).

1. Holoprosencephaly 3	142945	Autosomal dominant (AD).
2. Microphthalmia with coloboma 5	611638	AD
3. Schizencephaly	269160	unknown
4. Single median maxillary central incisor	147250	AD

b) How many literature references have been used to construct this page for OMIM, and when was the page last updated?

111 references are provides, the last update was on 07/07/2016 (accessed on 2018-06-13).

Open the Table View of Allelic Variants for the SHH gene from the left-side menu.

c) How many mutations are described in OMIM and which types of mutations do they include?

Twenty mutations are described here*. Under the column titled "Mutation", one can check which amino acid is changed at which location, or if a deletion of a given amount of base pairs has been removed.

* By clicking on the link "20 selected examples", one will find the following additional information: "How are mutations catalogued in OMIM?

> Mutations are catalogued in OMIM in the Allelic Variants section of gene entries (see 1.2). For most genes, only selected mutations are included. Criteria for inclusion include the first mutation to be discovered, high population frequency, distinctive phenotype, historic significance, unusual mechanism of mutation, unusual pathogenetic mechanism, and distinctive inheritance (e.g., dominant with some mutations, recessive with other mutations in the same gene). Most of the allelic variants represent disease-causing mutations. A few polymorphisms are included, many of which show a positive correlation with particular common disorders."

d) Which phenotype has been associated with the largest amount of the described variants?

HOLOPROSENCEPHALY 3

Take a closer look at variant 600725.0002 from the table.

e) Answer the following questions:

I. Which change(s) in the protein are reported?
II. At which position in the protein is this change reported?
III. What is the rs number of the SNP?
IV. Which base changed in the gene?
V. What is consequence of this change?

I. Glutamine (GLN) to stop codon (TER)
II. at position 100 in the protein;
III. the rs number is rs137854573.
IV. The base was changed from G to A.

V. The consequence is, that only part of the gene can be transcribed, causing the protein to be shorter and potentially malformed.

Click on the link for variant 600725.0004.

VI. How many literature references have described this variant? Describe in your own words, what these references found on the influence of this particular variant on the transcribed protein.

There are 2 literature references mentioned.

- The first paper [Roessler et al. (1996)] describes: “TGG-to-GGG transversion resulting in a trp117-to-gly (W117G) substitution of the SHH gene in a family with HPE3 (142945). The W117 residue is invariant in all hedgehog protein sequences and occurs immediately following the first alpha-helix of the murine Shh N fragment.”
- The second paper [Maity et al. (2005)] describes: “W117G mutation in mouse Shh caused failure of Shh processing, leading to retention of the immature glycosylated protein within the ER of transfected cells. In vitro binding assays using recombinant proteins showed that the mutation caused a temperature-dependent conformational change that allowed Shh to bind Ptc (PTCH1; 601309) at 4 or 32 degrees C, but not at 37 degrees C. The W117G mutation drastically reduced signaling potency in chicken embryo neural plate explant assays.”

Assignment 3: The biological context of SHH from WikiPathways *(30 minutes)*

The SHH protein play a vital role in development and the differentiation of multiple cell types, including neuronal cells. Look up the human “Oligodendrocyte Specification and differentiation (including remyelination), leading to Myelin Components for CNS” pathway in WikiPathways (https://www.wikipathways.org/index.php/Pathway:WP4304). This database (WikiPathways) shows you interactive drawings of pathway diagrams, which help to understand the biological context of genes, proteins and metabolites.

a) In which step(s) does the SHH protein play a role according to this pathway?
Specification of Multi-potent neural stem cells to Oligodendrocyte progenitor cells.

b) Which other protein(s) are needed for this step(s), and which protein(s) inhibit this step?
FGF (Fibroblast growth factor) is also needed for the specification. BMP2 and BMP4 (part of the Bone morphogenetic protein group) inhibit this specification.

Click on the SHH box, to display its identifiers from several other databases.

c) What are the Entrez Gene (NCBI Gene), Ensembl, RefSeq, HGNC symbol and Uniprot identifiers of the human SHH gene or protein? Tip: click on the grey arrow in the pop-up box for each of the database names.
Entrez: 6469
ENSEMBL: ENSG00000164690
RefSeq (gene): NM_001310462, NM_000193, NP_000184, NP_001297391

HGNC: SHH
Uniprot: Q15465, F8WEH4, F8WB84, C9JC48

d) Which of these databases give you information on a gene, and which one on a protein? Why can it be relevant to make a distinction between these two for biological databases?
Genes: Entrez, ENSEMBL, RefSeq (starting with NM), HGNC.
Proteins: Uniprot and RefSeq (starting with NP).
Genes first have to be transcribed and translated for a protein to be formed. This protein can then be used as an enzyme, to catalyse a reaction. In order to understand how a mutation in a gene affects a protein, one needs to know what happens to the transcript of the gene, how this affects the structure of a protein, and what the original function of a protein was (without a mutation). Combining all of this information will lead to a better understanding of for example the underlying biological cause of a genetic disease.

e) There are also other databases linked to a gene, such as OMIM (which you looked at for assignment 2). Which identifiers are linked from WikiPathways to OMIM? Compare these numbers with the ones you found for question 2a. Are these numbers the same?
OMIM: 142945, 269160, 600725, 611638, 147250.
All four disease ID numbers are mentioned here, with the 3rd number (600725) being the identifier for the SHH-gene itself.

Even though there is much more information that can be gained from this database, we will now continue to a database with lots of details on genes and the function of a gene, called Ensembl, which is database containing information on comparative genomics, evolution, sequence variation and transcriptional regulation. You can go to the Ensembl page for the human SHH encoding gene, by clicking on the link provided at WikiPathways (or by this link: http://ensemblgenomes.org/id/ENSG00000164690).

Assignment 4: Genetic and functional information on SHH from ENSEMBL *(30 minutes)*

You will now study the SHH gene in more detail, and part of the function of the protein encoded by the SHH gene. For this we will use Ensembl, which is database containing information on comparative genomics, evolution, sequence variation and transcriptional regulation. *Note: If too many people from the same country/IP-address try to use Ensembl, it can respond slowly. If that happens, click 'Mirrors' at the top of the page and try another mirror; this will guide you to a similar website, related to another country.*

a) What are other names of this gene, besides to official HGNC name SHH?
HHG1, HLP3, HPE3, MCOPCB5, SMMCI, TPT, TPTPS

b) What is the chromosomal position (chromosome, location, and strand) of this gene?
Chromosome 7: 155,799,986-155,812,273, reverse strand. (Accessed on 2018-06-13).

c) How many transcripts (splice variants) are known for this gene? How many are protein coding?
(Click on show transcripts). 5 known, 2 protein coding (accessed on 2018-06-13).

d) Give the name of one Gene Ontology (GO) annotation for each of the three GO domains (cellular component, molecular function, biological process). Pick annotations that originate from Uniprot, which are related to the basal lamina (which influencing cell differentiation, migration, and adhesion)

Cellular component: GO:0005615, extracellular space (similar term is mentioned by Reactome).
Molecular Function: GO:0043237, laminin-1 binding.
Biological Process: GO:0001708, cell fate specification (there are other valid answers).

e) Explain in your own words, how Gene Ontology terms can aid you, if you want to do research on a gene (start with https://en.wikipedia.org/wiki/Gene_Ontology_Term_Enrichment).

"Gene Ontology (GO) term enrichment is a technique for interpreting sets of genes making use of the Gene Ontology system of classification, in which genes are assigned to a set of predefined bins depending on their functional characteristics. For example, the gene FasR is categorized as being a receptor, involved in apoptosis and located on the plasma membrane.

Researchers performing high-throughput experiments that yield sets of genes (for example, genes that are differentially expressed under different conditions) often want to retrieve a functional profile of that gene set, in order to better understand the underlying biological processes. This can be done by comparing the input gene set each of the bins (terms) in the GO – a statistical test can be performed for each bin to see if it is enriched for the input genes. FunRich[1] can also be used for Gene Ontology enrichment analysis."

Quiz questions

Practical 2A Use of online data resources for molecular biology

* Questions 2, 3, and 4 are all 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 ENSEMBL?
    a. A metabolite database.
    b. A gene database.
    c. A protein database.
    d. A disease database.

2. How many diseases have been associated with the SHH gene [OMIM: 600725]?
    a. Two.
    b. Three.
    c. Four.
    d. Eleven.
    e. Twenty.

3. Which inheritance patterns are described on OMIM for all the diseases related to the SHH gene [OMIM: 600725]?
    a. Autosomal recessive (AR) and unknown.
    b. Autosomal dominant (AD) and multifactorial (MF).
    c. Autosomal dominant (AD) and unknown.
    d. Autosomal recessive (AR) and multifactorial (MF).

4. Which cells are formed under the influence of the SHH protein according to the myelination pathway from WikiPathways [WP4304]?
    a. Multi-potent neural stem cells.
    b. Immature oligodendrocytes.
    c. Oligodendrocyte progenitor cells.
    d. Mature oligodendrocytes.

5. Which of the following GO-terms originates from the Molecular Function-tree for the SHH gene (ENSG00000164690)?
    a. GO:0043237, laminin-1 binding.
    b. GO:0005615, extracellular space.
    c. GO:0007507, heart development.
    d. GO:0001525, angiogenesis.

LEVEL 2 questions (related to biological interpretation):

6. What could be the effect of a mutation resulting in a stop codon in gene for the corresponding protein?

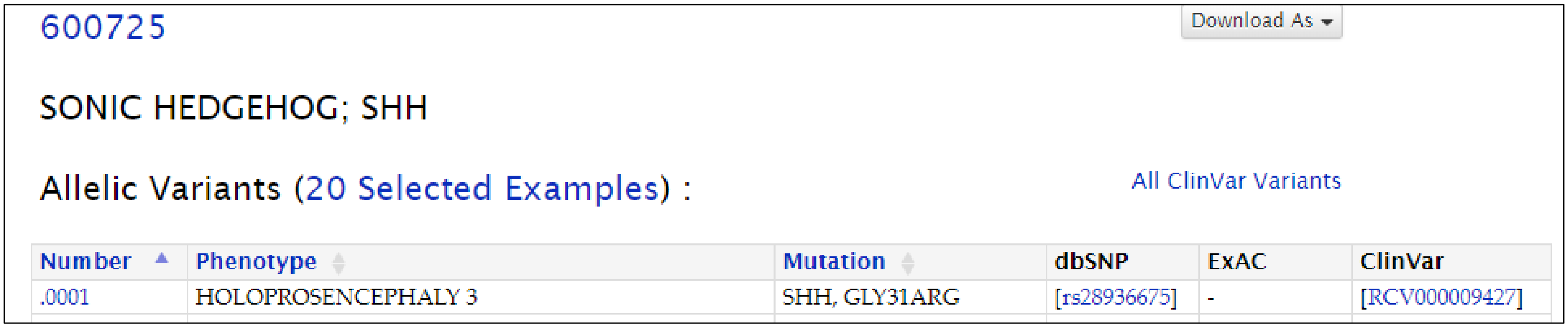
    a. The gene can still be transcribed completely; however the translation will not be complete, leading to a malformed protein.

b. The gene can be transcribed and translated normally; however the protein will not have its normal tertiary structure.
c. The gene will be transcribed in two parts, leading to two different proteins.
d. Only part of the gene can be transcribed, causing the protein to be shorter and potentially malformed.

7. What is the meaning of an autosomal dominant disease?

a. **A pattern of inheritance in which an affected individual has one copy of a mutant gene and one normal gene on a pair of autosomal chromosomes.**
b. A pattern of inheritance in which an affected individual has two copies of the mutant gene.
c. A mode of inheritance in which a mutation in a gene causes the phenotype to be expressed in all males and in females who are homozygous for the gene mutation.
d. Inheritance of which can only be transmitted from father to son.

8. What information can you derive from the following Table?



600725

SONIC HEDGEHOG; SHH

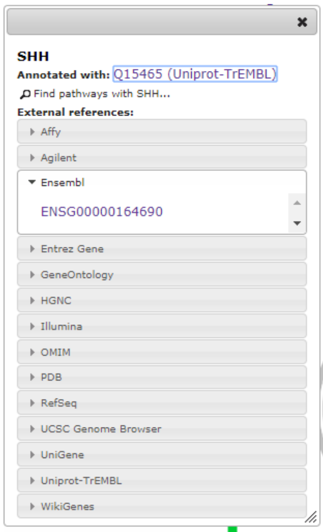
Allelic Variants (20 Selected Examples) :

All ClinVar Variants

Number ▲	Phenotype	Mutation	dbSNP	ExAC	ClinVar
.0001	HOLOPROSENCEPHALY 3	SHH, GLY31ARG	[rs28936675]	-	[RCV000009427]

a. The allelic variant of SHH has number 600725.
b. The mutation is autosomal dominant.
c. The gene affected by the mutation is called GLY31ARG.
d. An ARG amino acid is changed to GLY at position 28936675.
e. There are at least 20 allelic variants known for the gene SHH.
f. Holoprosencephaly 3 only has one variant, with identifier GLY31ARG.

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



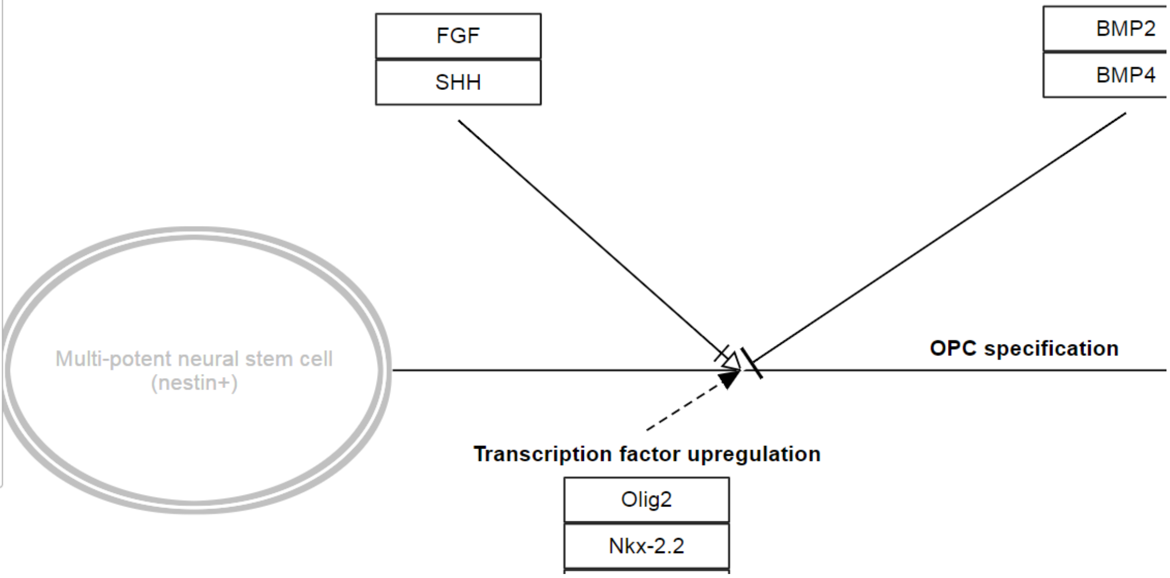
SHH

Annotated with: Q15465 (Uniprot-TrEMBL)

Find pathways with SHH...

External references:

- Affy
- Agilent
- Ensembl
  - ENSG00000164690
- Entrez Gene
- GeneOntology
- HGNC
- Illumina
- OMIM
- PDB
- RefSeq
- UCSC Genome Browser
- UniGene
- Uniprot-TrEMBL
- WikiGenes


FGF
SHH
BMP2
BMP4
Multi-potent neural stem cell
(nestin+)
OPC specification
Transcription factor upregulation
Olig2
Nkx-2.2

a. SHH and Nkx-2.2 have an effect on nestin+.
b. OPC specification is needed for transcription factor upregulation.
c. FGF, SHH, BMP2 and BMP4 all have the same effect on OPC specification.
d. The identifier for FGF is ENSG00000164690.
e. Multi-potent neural stem cells have a positive effect on the SHH gene.
f. BMP2 and BMP4 are transcription factors.

Practical 3A Practical training exercises

Use of online data resources for molecular biology

- With the preparatory assignment, you studied a specific gene (SHH) related to the growth of myelin sheets. However, SHH is related to establishing cell fates at several points during development.
- During this practical, we are going to explore more information in molecular biological databases for SHH, starting with a pathway from WikiPathways on heart development, and then progressing to another gene database (Entrez gene) from NCBI.
- When you get stuck trying to answer the questions, use the "*Reference guide online resources"* that was provided to you; if this doesn't help, ask one of the instructors.
- After the practical, the answers to these assignments will be uploaded on the Student Portal.

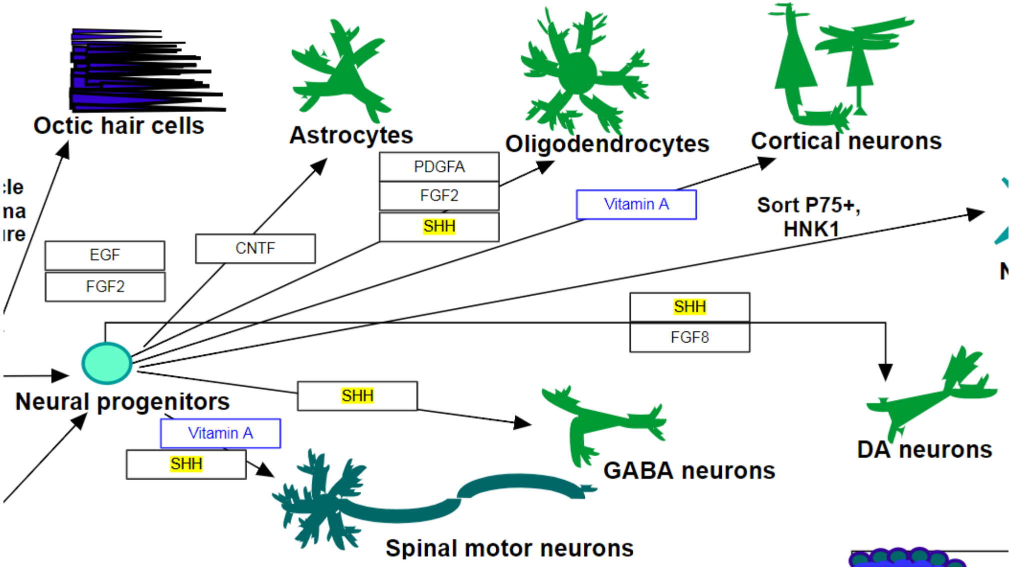
Assignment 1: Another biological context of SHH from WikiPathways *(20 minutes)*

Look up the human "Differentiation Pathway" in WikiPathways (https://www.wikipathways.org/index.php/Pathway:WP2848). This database (WikiPathways) shows you interactive drawings of pathway diagrams, which help to understand the biological context of genes, proteins and metabolites.

a) In which step(s) does the SHH protein play a role according to this pathway? (Tip: search for SHH with the key-shortcut ctrl + f, to see it light up in the pathway).

SHH seems to be involved in four different steps, all originating from neural progenitor cells. These steps are:

1. Development of oligodendrocytes
2. Development of DA neurons
3. Development of GABA-neurons
4. Development of spinal motor neurons.



b) Which other protein(s) are needed for the step to develop oligodendrocytes? Is this finding in line with the information you saw in preparatory question 3b?
- FGF2 (Fibroblast growth factor2) and platelet derived growth factor subunit A (PDGFA) are also needed for the specification.
- This is partially in line with the answer from prep. Question 2b, since FGF was mentioned here as well (however not specified to factor 2). However, PDGFA was not mentioned for the step called "Specification of Multi-potent neural stem cells to Oligodendrocyte progenitor cells", nor are the inhibitory BMPs. Since the Differentiation Pathway" includes several cell-types, it could be more general, where the "Oligodendrocyte Specification and differentiation (including remyelination), leading to Myelin Components for CNS" is very specific for just a few types of cells (and therefore includes more specific information).

c) Which other compound(s) are needed for the step to develop spinal motor neurons?
Vitamin A.

Click on the box of this compound (in blue), to display its identifiers from several other databases.

d) What are the HMDB, CHEBI, Pubchem compound and Wikidata identifiers of this compound? Tip: click on the grey arrow in the pop-up box for each of the database names.
HMDB: HMDB00305, HMDB0000305
CHEBI: CHEBI:17336, 17336
Pubchem Compound: 445354
Wikidata: Q424976

Even though we could go into more detail on chemical compounds with help of the databases mentioned in question 2d, we will now return to the SHH gene. We will continue to a database with more details on genes and the function of a gene, called Entrez Gene from NCBI. You can go to the NCBI page for the human SHH encoding gene, by clicking on the link provided at WikiPathways (or by this link: http://www.ncbi.nlm.nih.gov/gene/6469).

Assignment 2: Genetic and functional information on SHH from NCBI *(40 minutes)*

NCBI gene is a database that integrates information from a wide range of species. A record on a gene may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

a) What is the NCBI Gene identifier of this gene?
6469 (this can be seen directly under the title or in the webpage link).

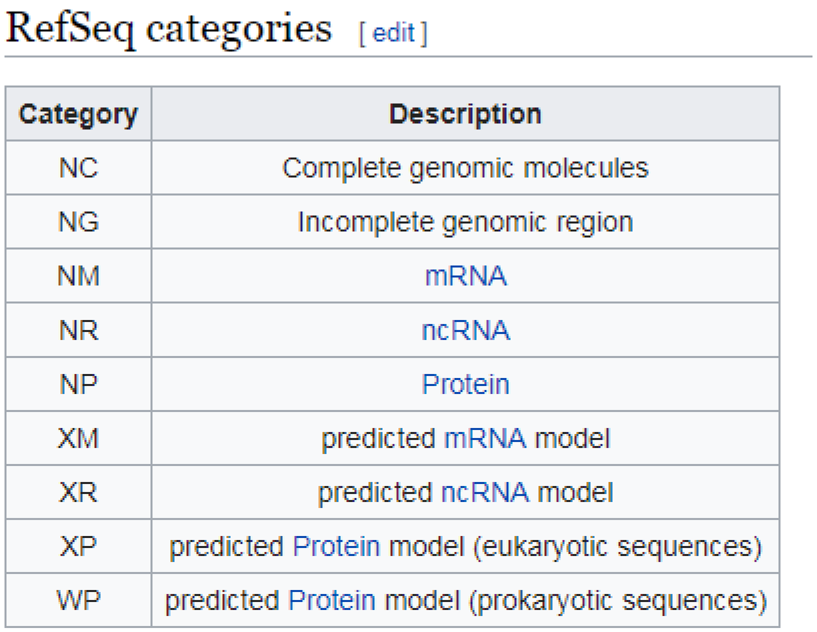
b) What are other names of this gene, besides to official HGNC name SHH?
TPT; HHG1; HLP3; HPE3; SMMCI; ShhNC; TPTPS; MCOPCB5

c) Compare your answer from question 2b to question 4a from the preparatory assignments. Are there similarities or differences? How could differences in names influence your research in other databases? How do databases try to minimalise this influence?
- All names are similar.
- If there would be differences in names, a literature study would be harder to perform (or you might miss information if you only look for 1 abbreviation).

- Therefore, databases can provide a solution (since they mention multiple names, work with identifiers, and link out to other databases).

d) Give the RefSeq identifier of one of the mRNA, with the corresponding RefSeq ID and name for the protein for one of its splice variants.
NM_000193.3 → NP_000184.1, sonic hedgehog protein isoform 1 preproprotein (others are possible).

e) Explain in your own words what Ref-Seq identifiers are, which different categories there are, and how they can aid you, if you want to do research on the relationship between a gene and a protein (start with https://en.wikipedia.org/wiki/RefSeq).
- The Reference Sequence (RefSeq) database is an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein products. This database, provides only a single record for each natural biological molecule (i.e. DNA, RNA or protein) for major organisms ranging from viruses to bacteria to eukaryotes.
- A list of the main categories is given at Wikipedia, with a link to a book giving more details and categories. For this list (see below), we can deduce that NM_ stands for a messenger RNA (mRNA) sequence, and NP_ for a protein sequence.



RefSeq categories [edit]

Category	Description
NC	Complete genomic molecules
NG	Incomplete genomic region
NM	mRNA
NR	ncRNA
NP	Protein
XM	predicted mRNA model
XR	predicted ncRNA model
XP	predicted Protein model (eukaryotic sequences)
WP	predicted Protein model (prokaryotic sequences)

- For each model organism, RefSeq aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. RefSeq is limited to major organisms for which sufficient data are available. Therefore, if one wants to research a gene, the RefSeq identifier NC_ corresponds to a unique nucleotide sequence, which gives the complete sequence of the gene. The messenger RNA from this gene can be described with the NM_ identifiers, and the resulting protein with the NR identifier. The NR identifier sequence belongs to Non-coding regions (ncRNA), therefore not translated to a protein.

f) How many exons does the gene have (total regardless of transcript) (see 'exon count')?
8

g) In which tissue is this gene expressed most (scroll down to ‘Expression’).?

Stomach

h) To which neurological Phenotype is it associated (scroll down to ‘Phenotypes’)? Is this finding in line with the information you saw in preparatory question 2a?

- Holoprosencephaly 3 [OMIM:142945]; Microphtalmia, isolated with coloboma 5 [OMIM:611638]; schizencephaly [OMIM:269160]; single median maxillary incisor [OMIM:147250]. (**Accessed on 2018-07-22)
- These disease are also mentioned on OMIM itself (Q 2a prep.), and the identifiers also correlate.

i) Name all the seven pathways from WikiPathways, in which the gene is involved.

I. Differentiation Pathway, organism-specific biosystem. (also seen at q 1).
II. Dopaminergic Neurogenesis, organism-specific biosystem
III. Ectoderm Differentiation, organism-specific biosystem
IV. Heart Development, organism-specific biosystem
V. Hedgehog Signaling Pathway, organism-specific biosystem
VI. Integrated Pancreatic Cancer Pathway, organism-specific biosystem
VII. Tgif disruption of Shh signaling, organism-specific biosystem
Accessed on 2018-07-22

j) Explain in your own words for two of these pathways, why SHH plays a role in them.

From the information learned with the previous questions, SHH plays a role by establishing cell fates at several points during development. Since several pathways mention “differentiation” or “development”, the presence of SHH in these pathways make sense.