Practical sessions overview

- **Practical 1: Databases, Alignment, Search** – 31/03
- Practical 2: Motifs, Python & assignment – 28/04
- Practical 3: Phylogenetic trees, protein structures and Gene Ontology – 5/05
- Practical 4: Next Generation Sequencing – 12/05
- Q&A session – 19/05
- EXAM – 26/5

Practicum agreements

- Assignment:
  - **Mandatory**
  - Differs for computer scientists – biologists
  - Will be explained during practicum 2

- Biology students:  ?

Practical session 1

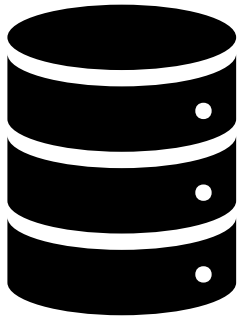
Databases & Alignment

Practical agreements

- There are 2 correction moments:
  - 15u: correction chapter 1
  - 17u: correction chapters 2-3
- All solutions will be provided afterwards
- Homework questions Q1-Q3: solutions will be provided afterwards.

Overview of the practical session

NCBI &
UniProt



gene/protein
info

>Gene_sequence
TATCATCATATGGTAGTAGGTAAAT

>Protein_sequence
MKSGSGGGSPTSLW

Pairwise alignment

```
PDRLRLPLCFLGVFVCYFYYGILQEKITRGK
|..::|.:|.||::..:..:|:|||.||...
PGLIQLAICVLGIYASFLSWGVLQEAITTVN
```

Multiple alignment

```
MSETAPVAQAASTATEKPAAAKKTKKPAK-AAAPF
MSETAPVPQPASVAPEKPAATKKTRKPAK-AAVPF
MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASK
MSEVALPAPAASTSPEKPSAGKKAKKPAKAAAAAK
***..  . .**.:.*** * **:.**** **...
```

Section 1

Databases

Exercise 1.1: NCBI databases (Gene DB)

(1) Look for human insulin in the **Gene database** and answer following questions:

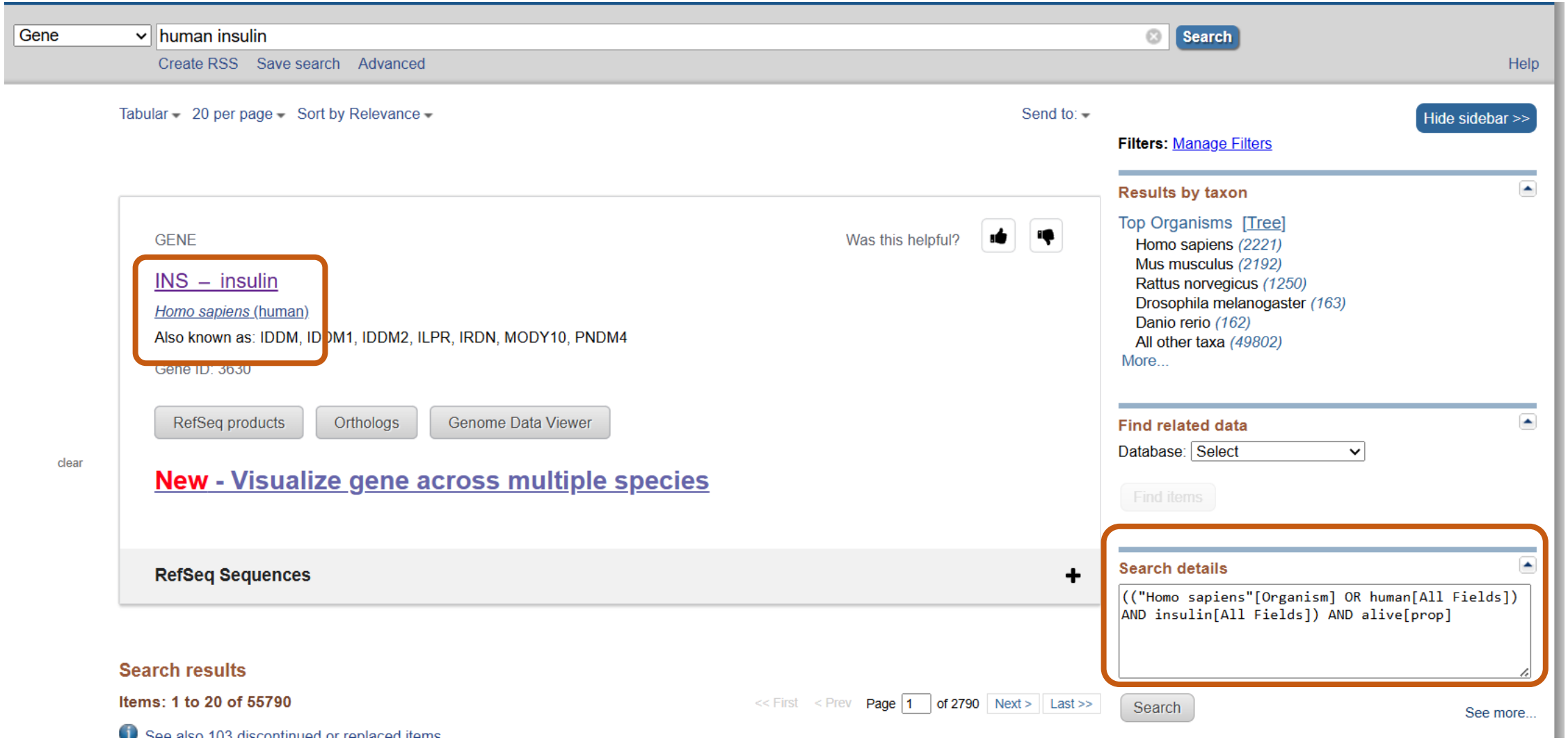
- Which diseases are associated with problems of insulin metabolism?
- Find molecular components that interact with insulin.
- Gene Ontology: Take a look at the 3 gene ontology classes and the associated terms for insulin. What can this tell you about the function of insulin?

Exercise 1.1: NCBI databases (Gene DB)

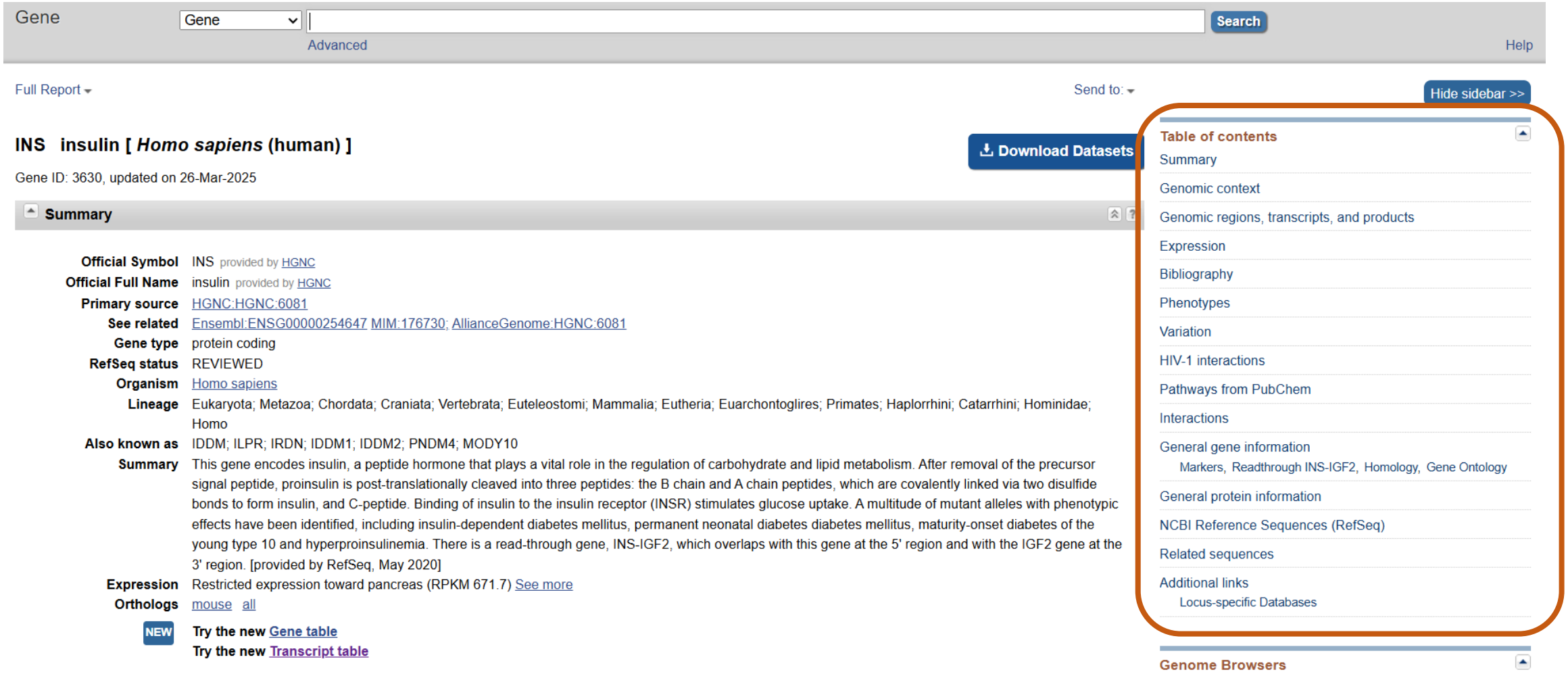
Gene | (homo sapiens[Organism]) AND insulin[Gene/Protein Name] | Search

Name/Gene ID	Description	Location	Aliases	MIM
IGF1 ID: 3479	insulin like growth factor 1 [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (102395860..102481839, complement)	IGF, IGF-I, IGFI, MGF	147440
CTLA4 ID: 1493	cytotoxic T-lymphocyte associated protein 4 [*Homo sapiens* (human)]	Chromosome 2, NC_000002.12 (203867771..203873965)	ALPS5, CD, CD152, CELIAC3, CTLA-4, GRD4, GSE, IDDM12	123890
IGF1R ID: 3480	insulin like growth factor 1 receptor [*Homo sapiens* (human)]	Chromosome 15, NC_000015.10 (98648539..98964530)	CD221, IGFIR, IGFR, JTK13	147370
INS ID: 3630	insulin [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (2159779..2161209, complement)	IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10	176730
MAPK3 ID: 5595	mitogen-activated protein kinase 3 [*Homo sapiens* (human)]	Chromosome 16, NC_000016.10 (30114105..30123309, complement)	ERK-1, ERK1, ERT2, HS44KDAP, HUMKER1A, P44ERK1, P44MAPK, PRKM3, p44-ERK1, p44-MAPK	601795
IGFBP3 ID: 3486	insulin like growth factor binding protein 3 [*Homo sapiens* (human)]	Chromosome 7, NC_000007.14 (45912245..45921272, complement)	BP-53, IBP3	146732
INSR ID: 3643	insulin receptor [*Homo sapiens* (human)]	Chromosome 19, NC_000019.10 (7112257..7294414, complement)	CD220, HHF5	147670

Exercise 1.1: NCBI databases (Gene DB)

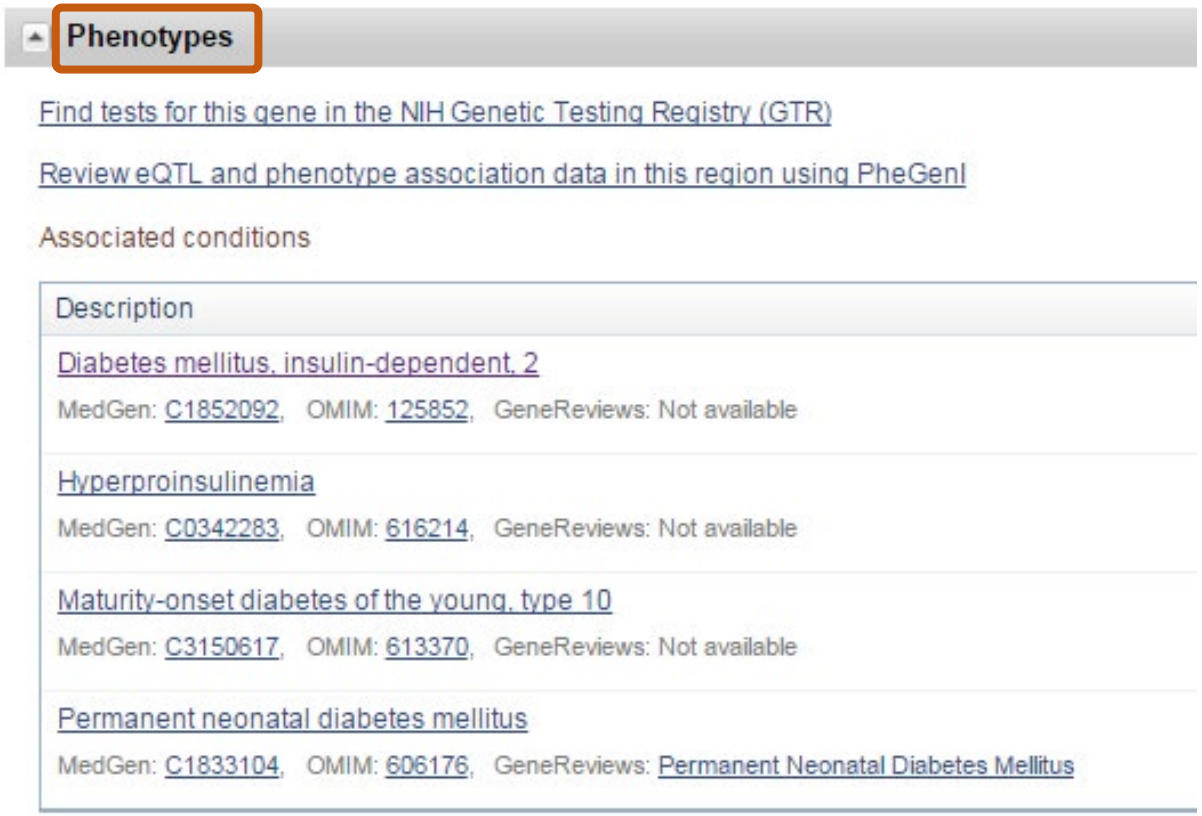


Exercise 1.1: NCBI databases (Gene DB)



Exercise 1.1: NCBI databases (Gene DB)

- Which diseases are associated with problems of insulin metabolism?



Exercise 1.1: NCBI databases (Gene DB)

- Find molecular components that interact with insulin.

Interactions

Items 1 - 25 of 29

Products	Interactant	Other Gene	Complex	Source	Pubs	Description
P01308	P16870	CPE		HPRD	PubMed	
P01308	P07858	CTSB		HPRD	PubMed	
P01308	P07339	CTSD		HPRD	PubMed	
P01308	P14091	CTSE		HPRD	PubMed	
P01308	P35557	GCK		HPRD	PubMed	
P01308	P01906	HLA-DQA2		HPRD	PubMed	

Exercise 1.1: NCBI databases (Gene DB)

- Gene Ontology: Take a look at the 3 gene ontology classes and the associated terms for insulin. What can this tell you about the function of insulin?

General gene information

Function	Evidence Code	Pubs
hormone activity	IC	PubMed
hormone activity	IMP	PubMed
hormone activity	NAS	PubMed
identical protein binding	IPI	PubMed

Process	Evidence Code	Pubs
ER to Golgi vesicle-mediated transport	TAS	
G protein-coupled receptor signaling pathway	IDA	PubMed
activation of protein kinase B activity	IDA	PubMed
acute-phase response	IDA	PubMed

Component	Evidence Code	Pubs
Golgi lumen	TAS	
Golgi membrane	TAS	
endoplasmic reticulum lumen	TAS	
endoplasmic reticulum-Golgi intermediate compartment membrane	TAS	

Exercise 1.1: NCBI databases (GenBank)

(2) Look for insulin in **GenBank** (Nucleotide) and answer the following questions:

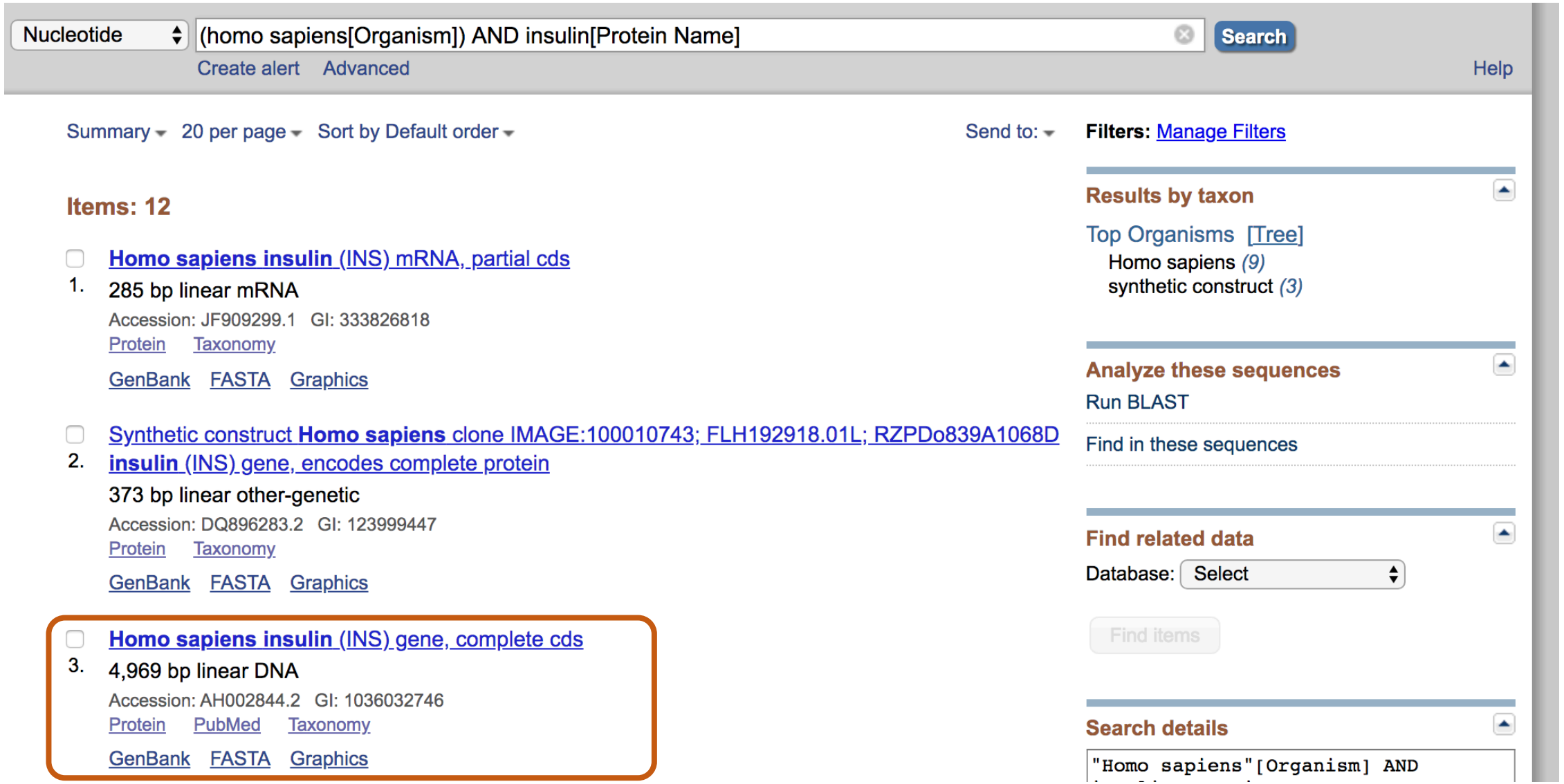
- What is the accession number of insulin. (And what is an accession number?!) What is the point of this identifier? What would happen if a new (or corrected) version of the same sequence is added to the data base?

- What is the version of the gene assembly?

- Who put this information on Genbank, and can you find out something more about the research that was done to produce this data?

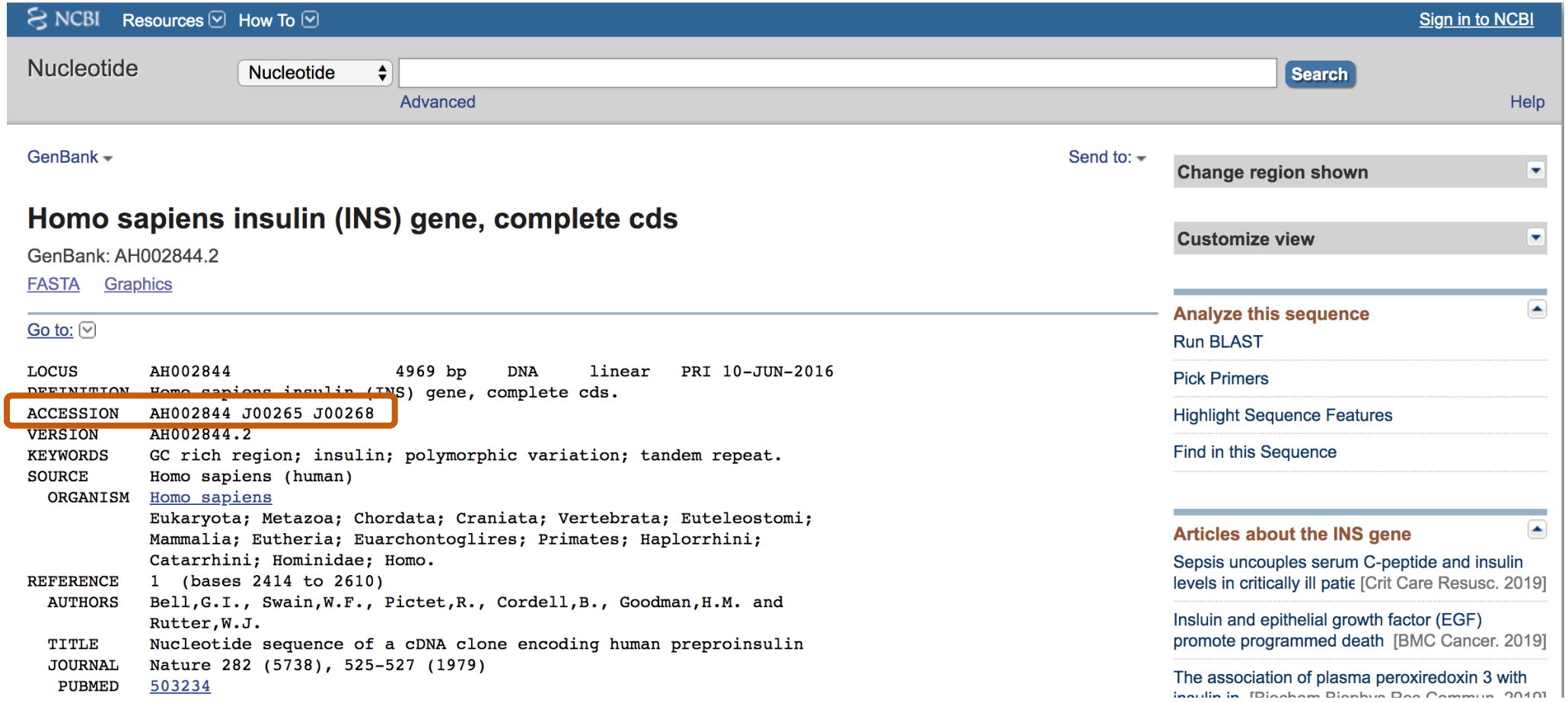
- Find the FASTA sequence of the gene. What is a FASTA file and how is it structured?

Exercise 1.1: NCBI databases (GenBank)



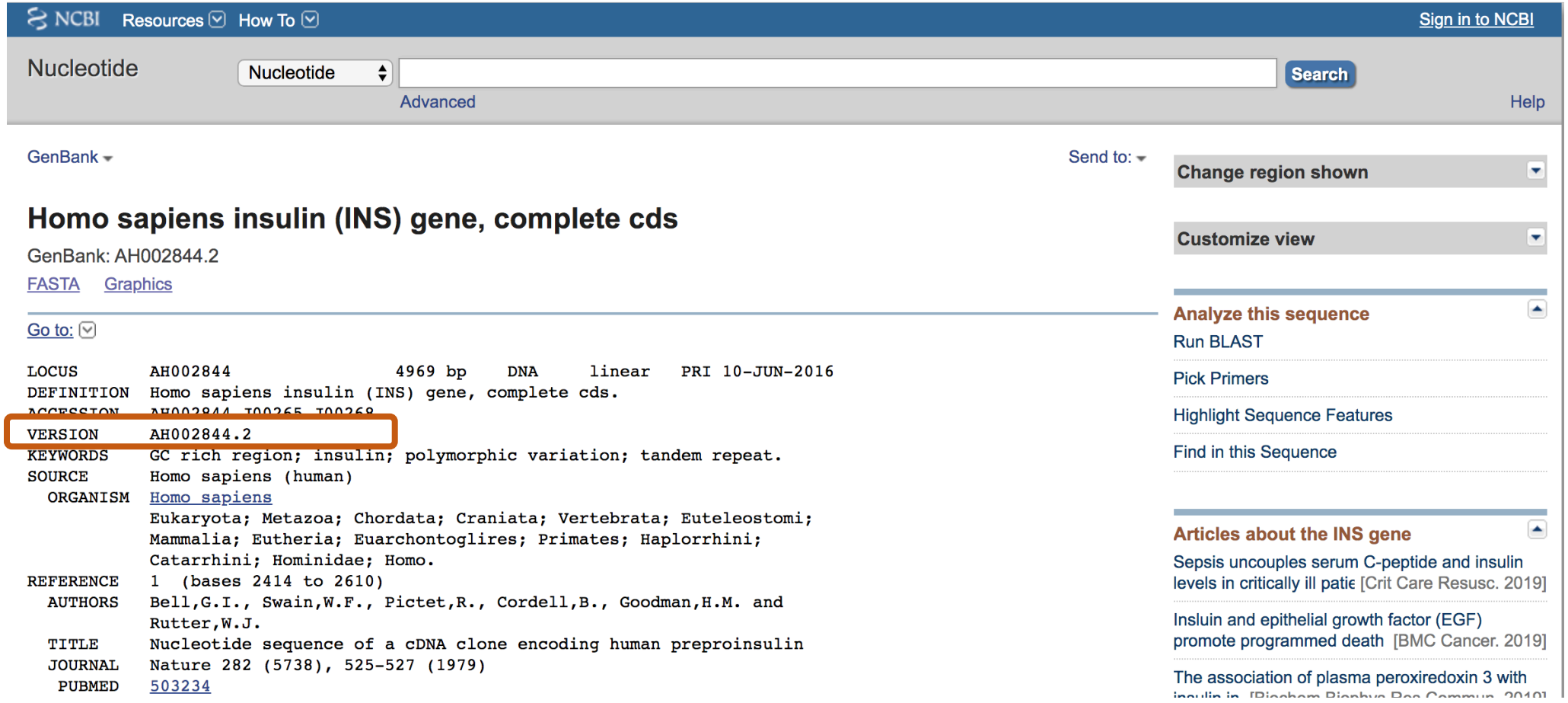
Exercise 1.1: NCBI databases (GenBank)

- What is the accession number of insulin. (And what is an accession number?!) What is the point of this identifier? What would happen if a new (or corrected) version of the same sequence is added to the data base?



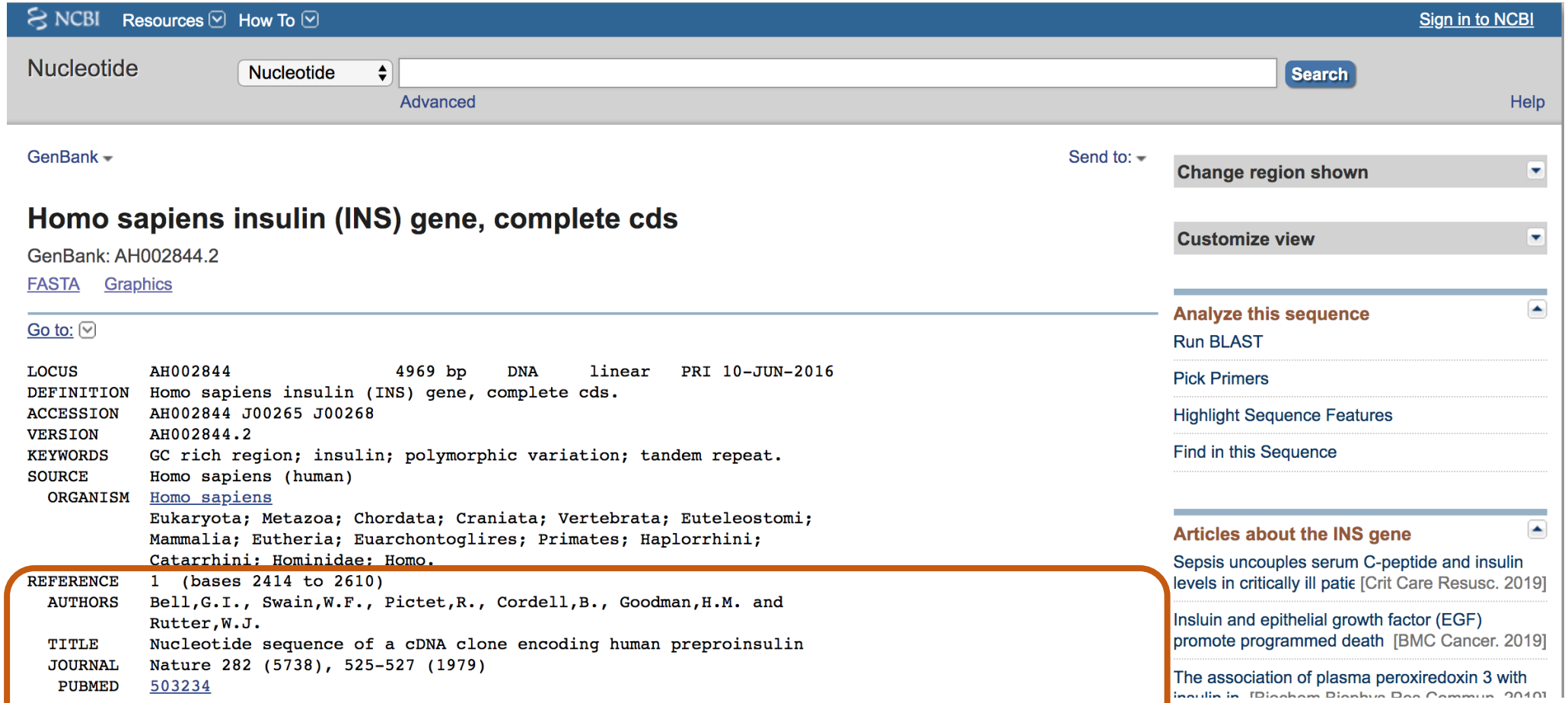
Exercise 1.1: NCBI databases (GenBank)

- What is the version of the gene assembly?



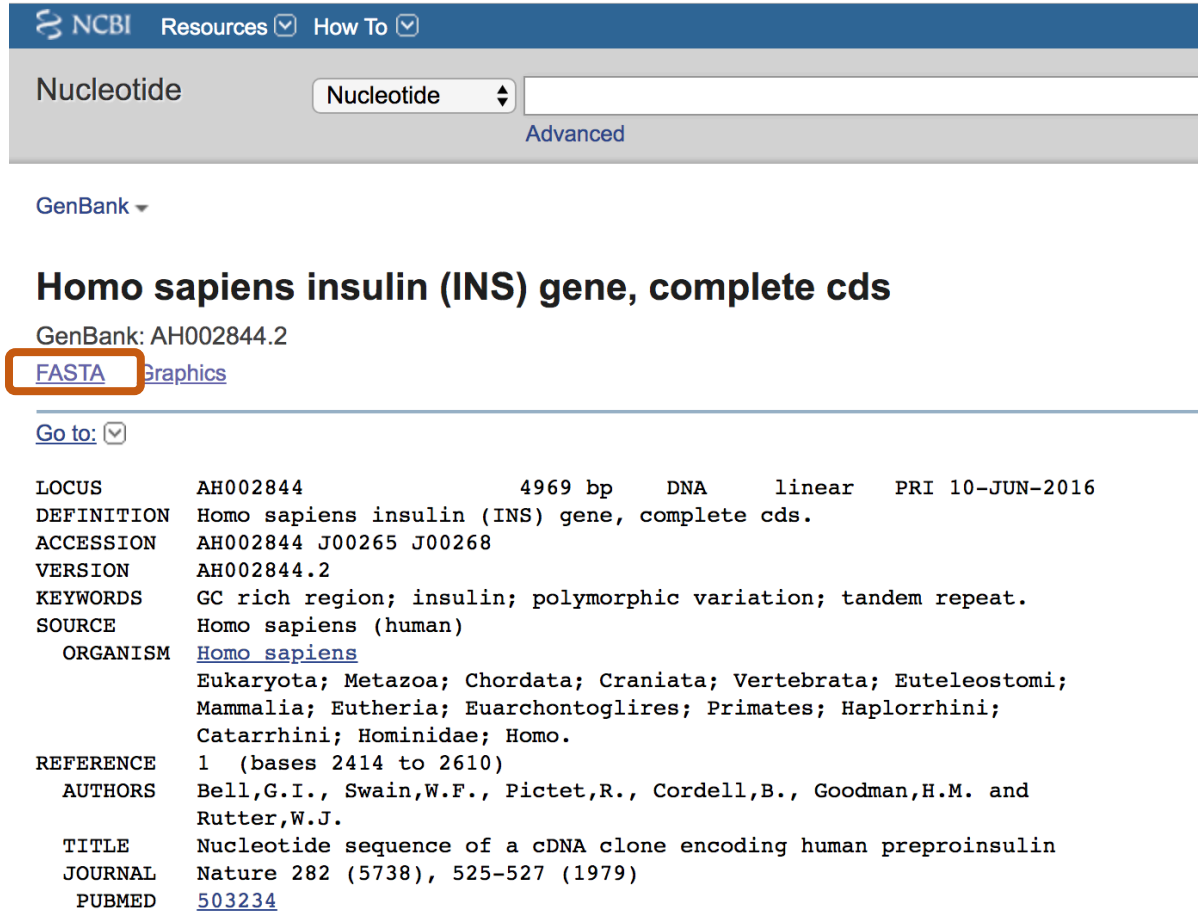
Exercise 1.1: NCBI databases (GenBank)

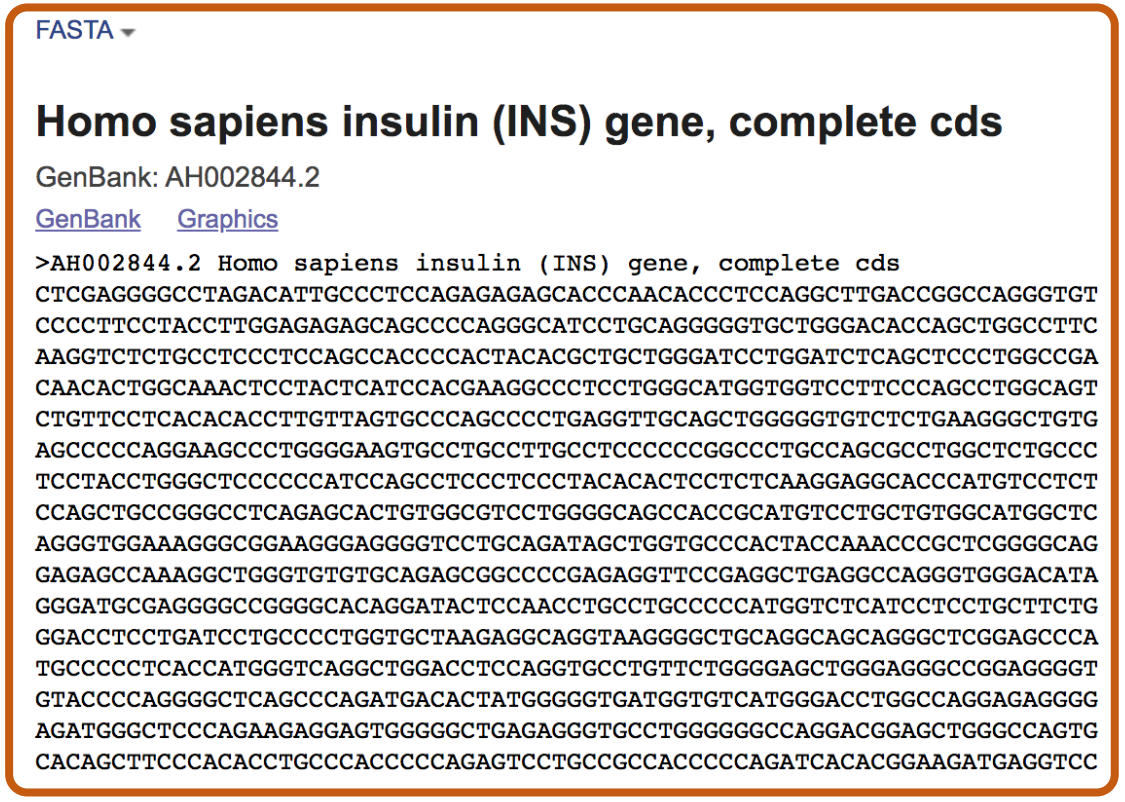
- Who put this information on Genbank, and can you find out something more about the research that was done to produce this data?



Exercise 1.1: NCBI databases (GenBank)

- Find the FASTA sequence of the gene.





Exercise 1.1: NCBI databases (GenBank)

- What is a FASTA file and how is it structured?

The FASTA format

>Gene_1 Very descriptive description

TATATATATGGTAGTAGGTAAAT

>Gene_2 This gene has some unknown bases in the middle

TACGATCGTCGCTAGCTAGCTAC

TGATGTCTNNNNGATCGATGCTT

CTAGCTAGCTAGCTGATCGATGC

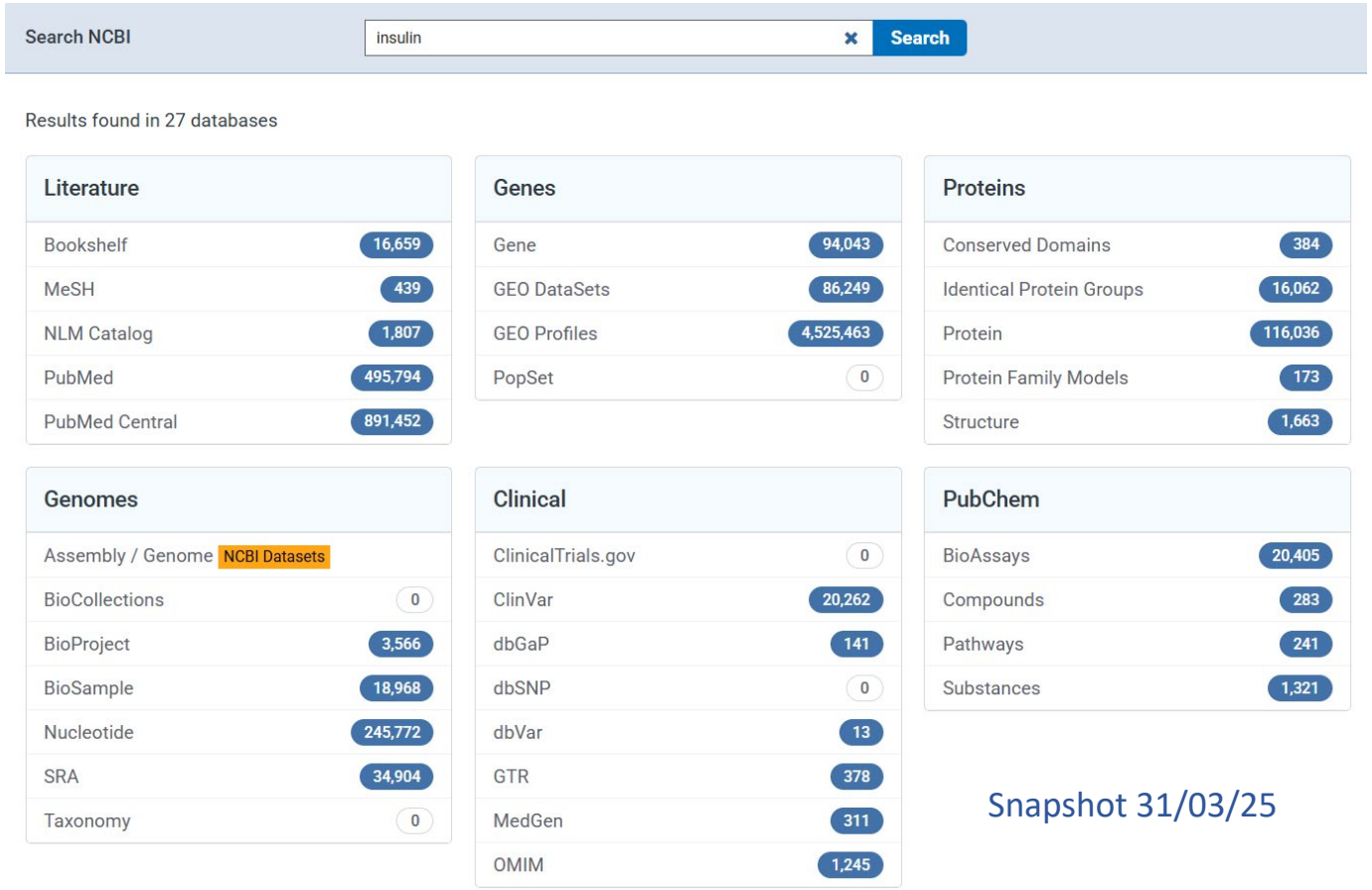
>Gene_3

TATAGTCTGCTGCTGATCGACGA

TGCTAGTCGTGATCGATGCTAGC

Exercise 1.1: NCBI databases (all databases)

(3) Finally, search for insulin in all NCBI databases. What other types of additional information can you find here?



Snapshot 31/03/25

Exercise 1.1: NCBI databases (dbSNP)

(4) In this exercise, we will study the human BRCA1 gene. Mutations in this gene have been associated with breast cancer:

- How many deletions for this gene can you find in dbSNP that are pathogenic?
- Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP

Exercise 1.1: NCBI databases (dbSNP)

How many deletions for this gene can you find in dbSNP that are pathogenic?

Snapshot 31/03/25

Exercise 1.1: NCBI databases (dbSNP)

Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP

rs80357502

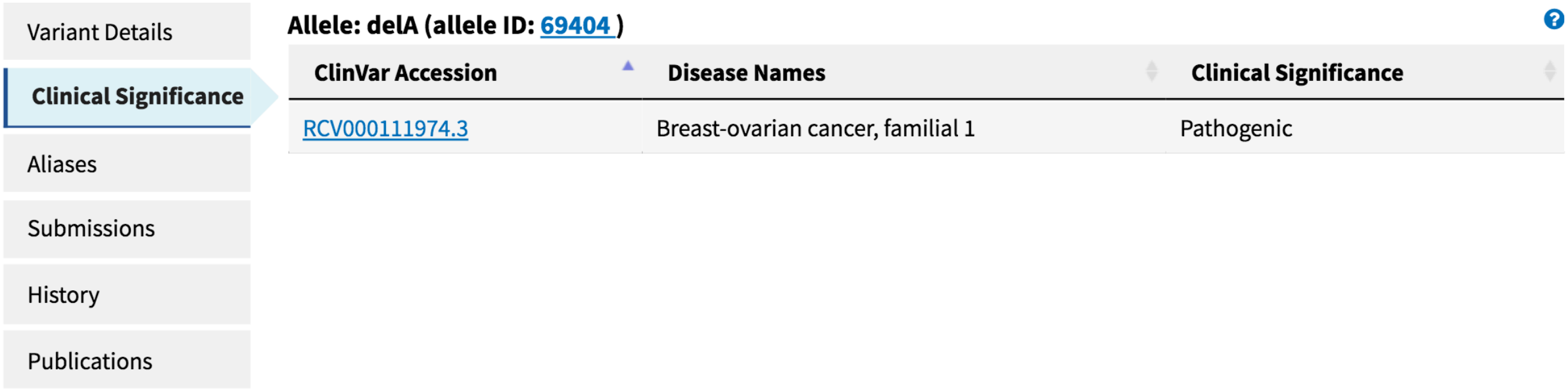
Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr17:43092551 (GRCh38.p14)	**Gene : Consequence**	BRCA1 : Frameshift Variant
Alleles	delA	**Publications**	1 citation
Variation Type	Deletion		LitVar2 4
Frequency	None	**Genomic View**	See rs on genome

Gene: BRCA1, BRCA1 DNA repair associated (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
BRCA1 transcript variant 1	NM_007294.3:c.2980del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 2	NM_007300.4:c.2980del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 3	NM_007297.4:c.2839del	C [TGT] > V [GT]	Coding Sequence Variant
BRCA1 transcript variant 4	NM_007298.3:c.	N/A	Intron Variant
BRCA1 transcript variant 5	NM_007299.4:c.	N/A	Intron Variant
BRCA1 transcript variant 6	NR_027676.1:n.3116del	N/A	Non Coding Transcript Variant
breast cancer type 1 susceptibility protein isoform 1	NP_009225.1:p.Cys994fs	C (Cys) > V (Val)	Frameshift
breast cancer type 1 susceptibility protein isoform 2	NP_009231.2:p.Cys994fs	C (Cys) > V (Val)	Frameshift
breast cancer type 1 susceptibility protein isoform 3	NP_009228.2:p.Cys947fs	C (Cys) > V (Val)	Frameshift

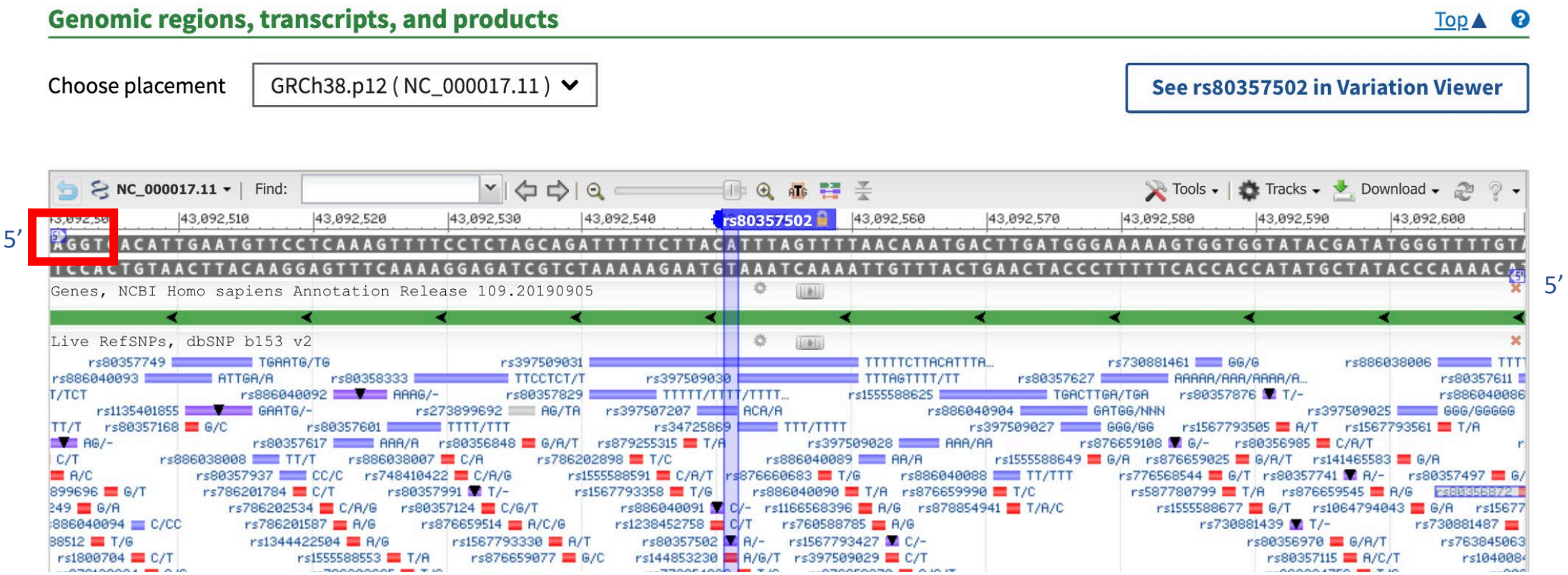
Exercise 1.1: NCBI databases (dbSNP)

Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP



Exercise 1.1: NCBI databases (dbSNP)

Take a closer look at one of the selected deletions to see what kind of information is stored in dbSNP



Exercise 1.2: UniProt

- Navigate to http://www.uniprot.org/ and search again for human insulin.

What additional information can you find here compared to Genbank?

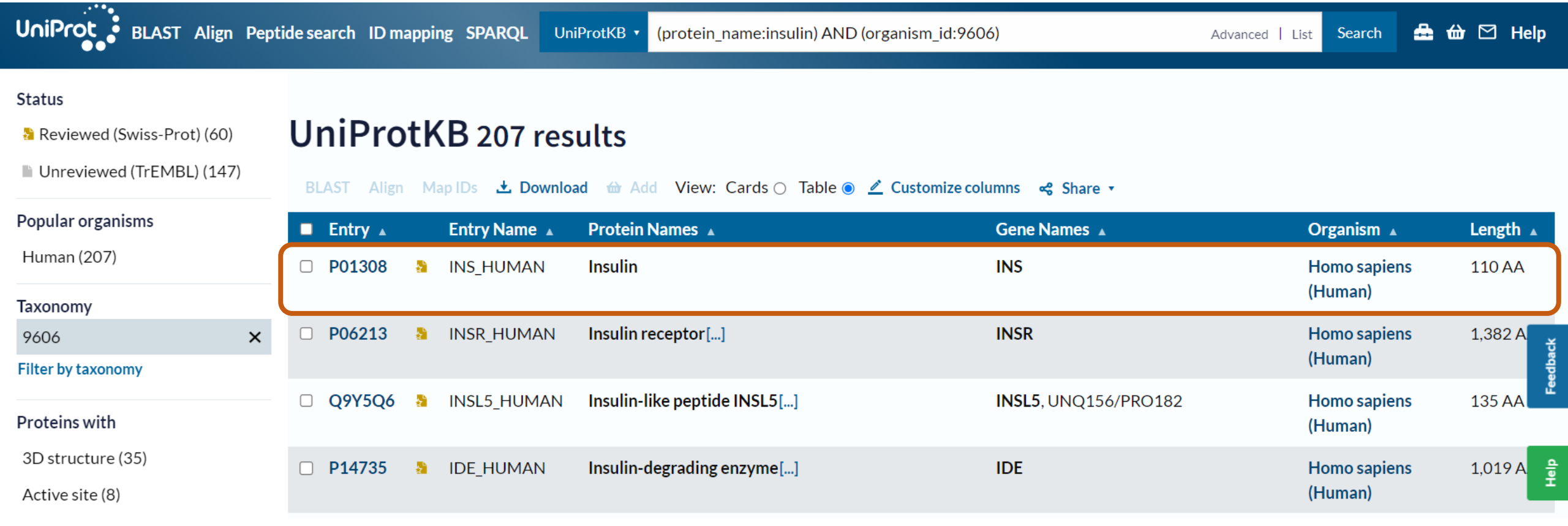
- Try to find out which protein is associated with Alexander disease.

- UniProt isn't only able to return single protein entries, you can also look up whole proteomes if you want. Search for the proteome of the zebrafish (Danio rerio) or your favorite model organism. How many proteins does it contain? How many of those are reviewed?

Exercise 1.2: UniProt

- Navigate to http://www.uniprot.org/ and search again for human insulin.

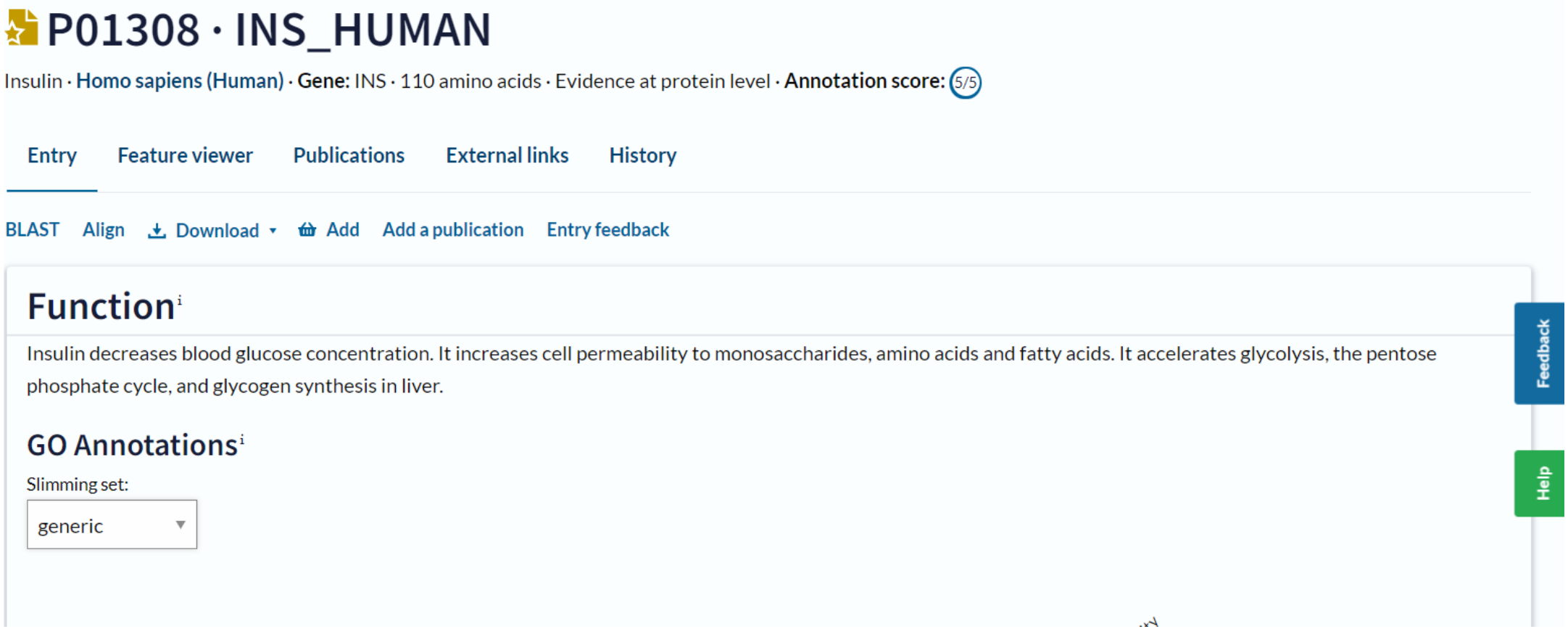
What additional information can you find here compared to Genbank?



Exercise 1.2: UniProt

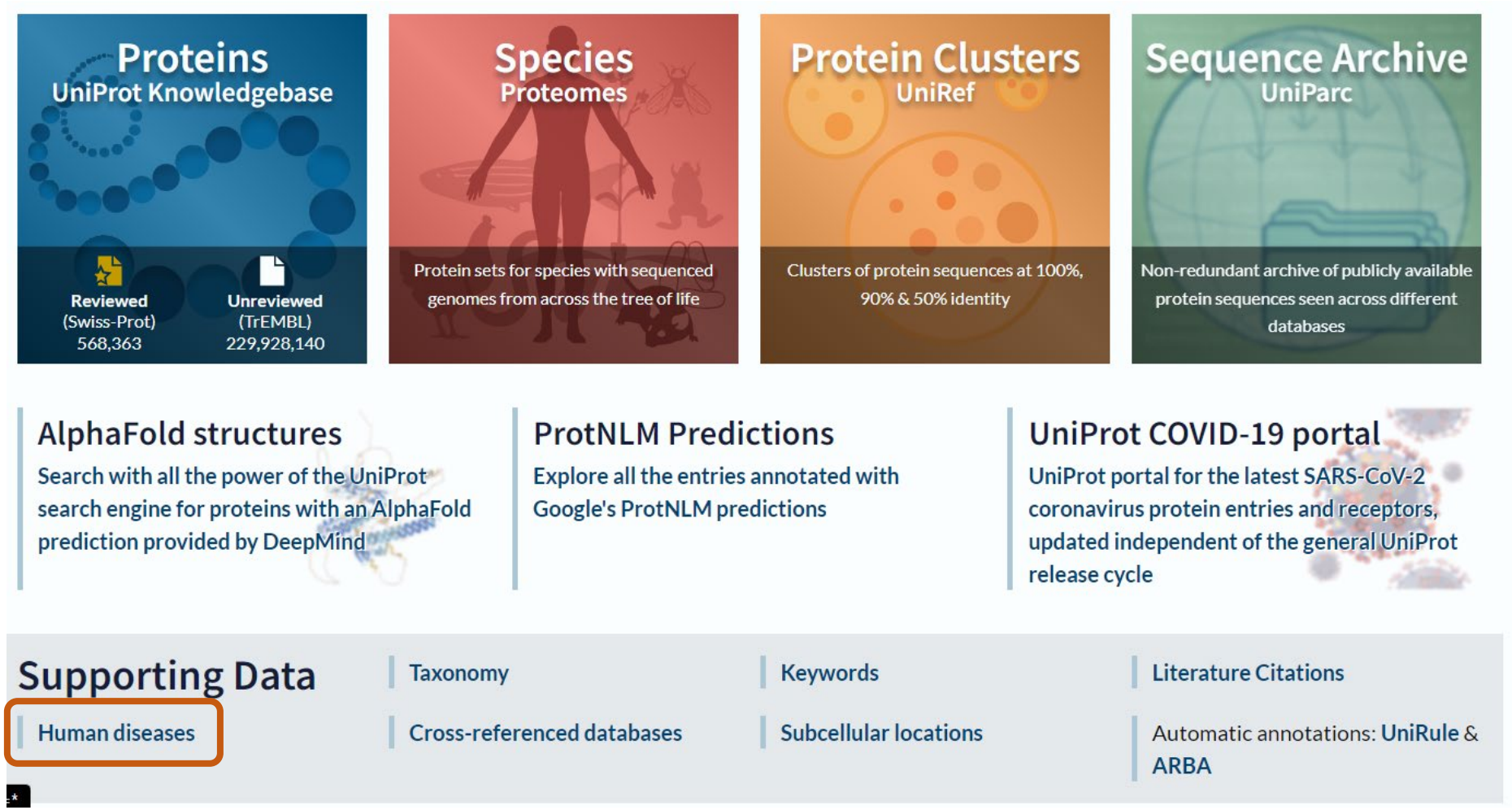
- Navigate to http://www.uniprot.org/ and search again for human insulin.

What additional information can you find here compared to Genbank?



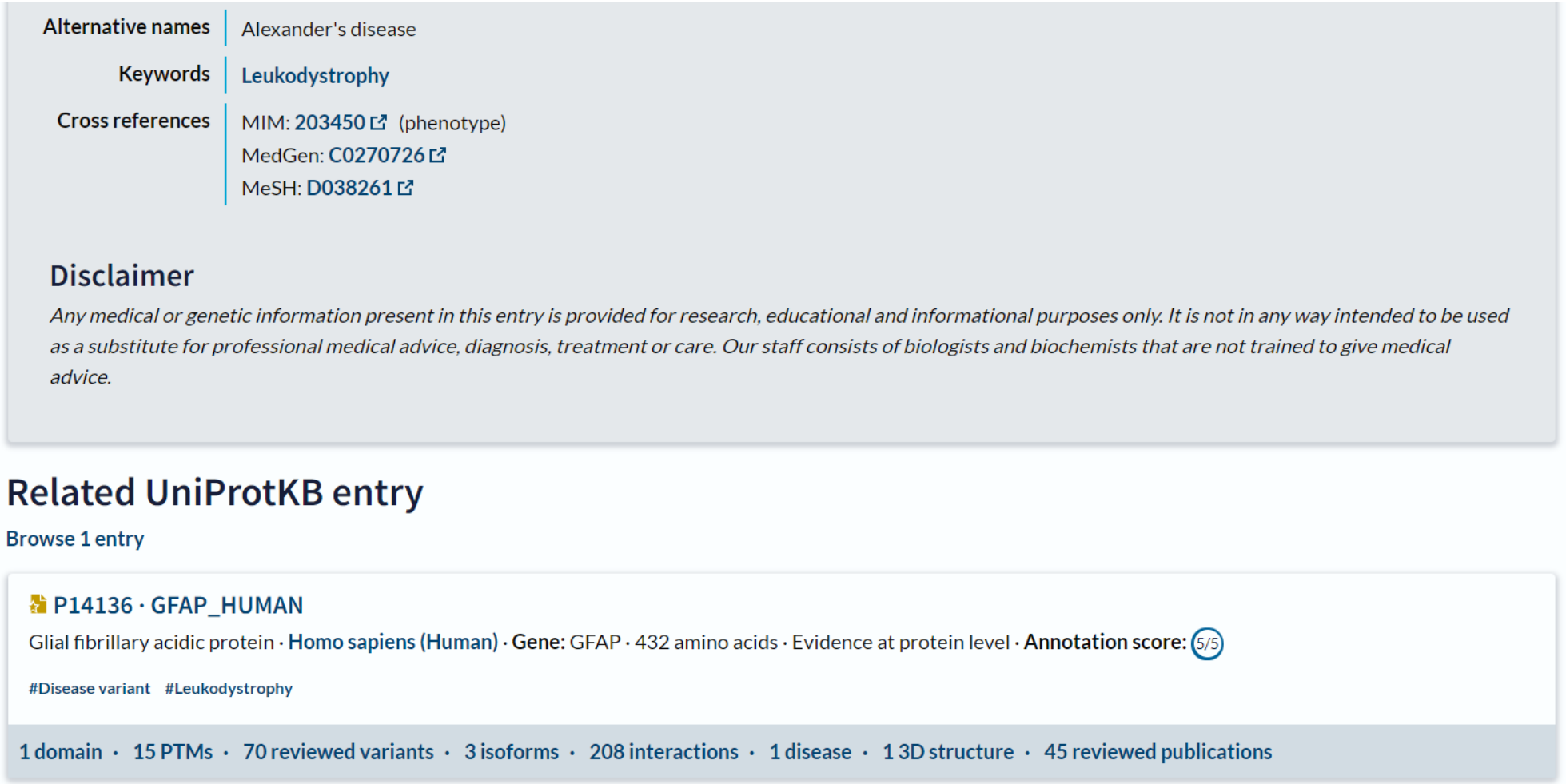
Exercise 1.2: UniProt

- Try to find out which protein is associated with Alexander disease.



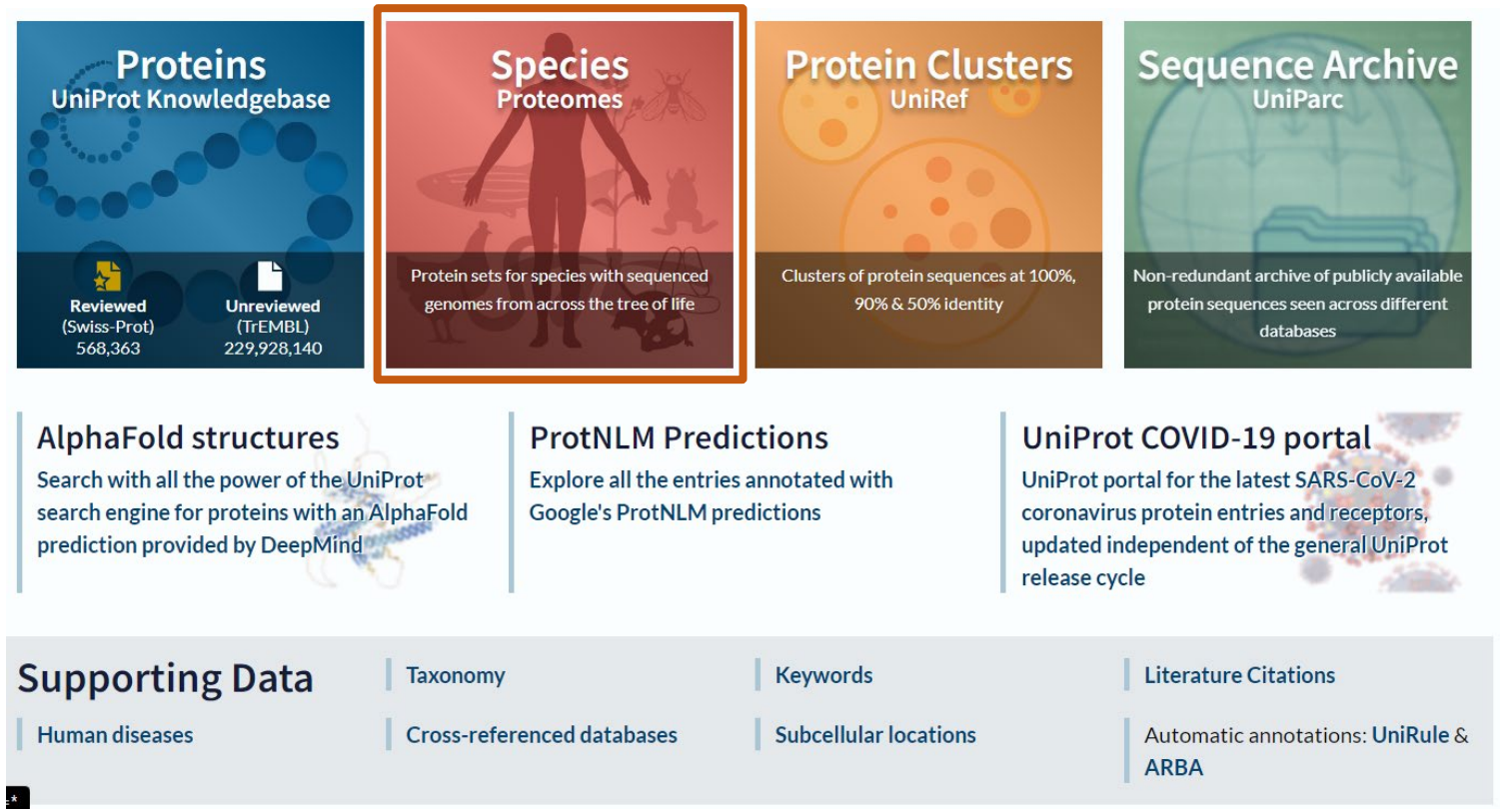
Exercise 1.2: UniProt

- Try to find out which protein is associated with Alexander disease.



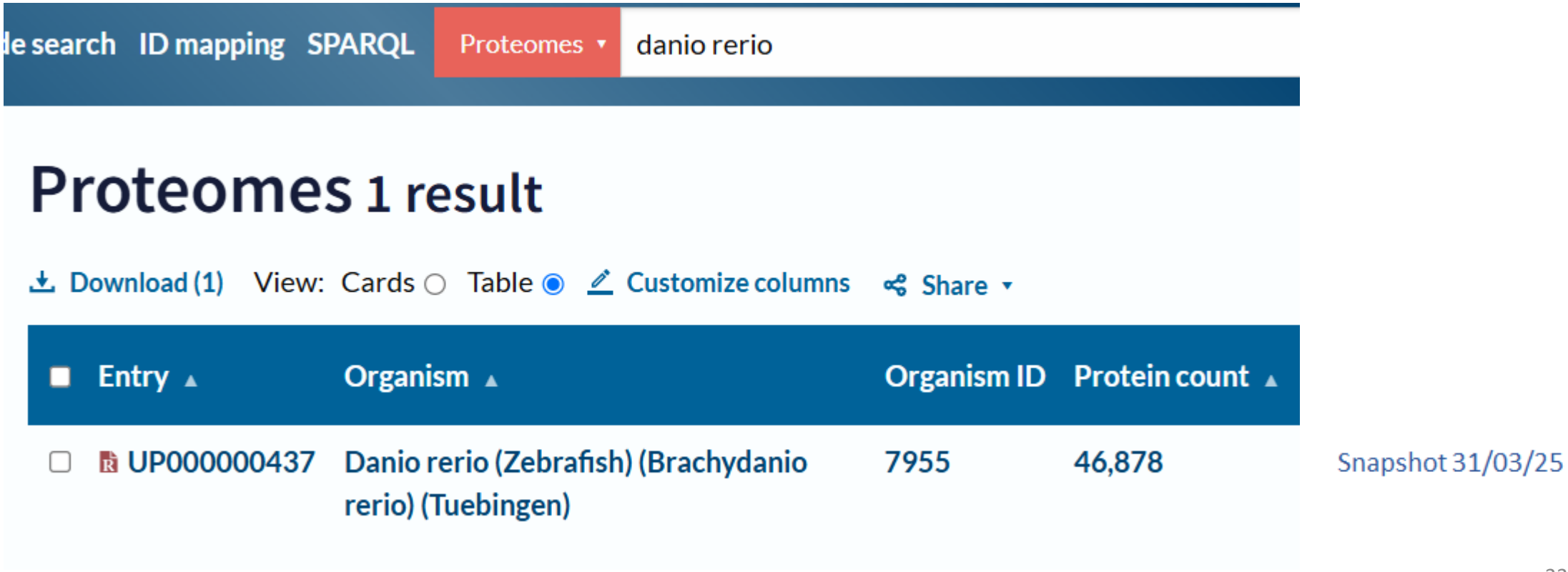
Exercise 1.2: UniProt

- Search for the proteome of the zebrafish (Danio rerio) or your favorite model organism. How many proteins does it contain? How many of those are reviewed?



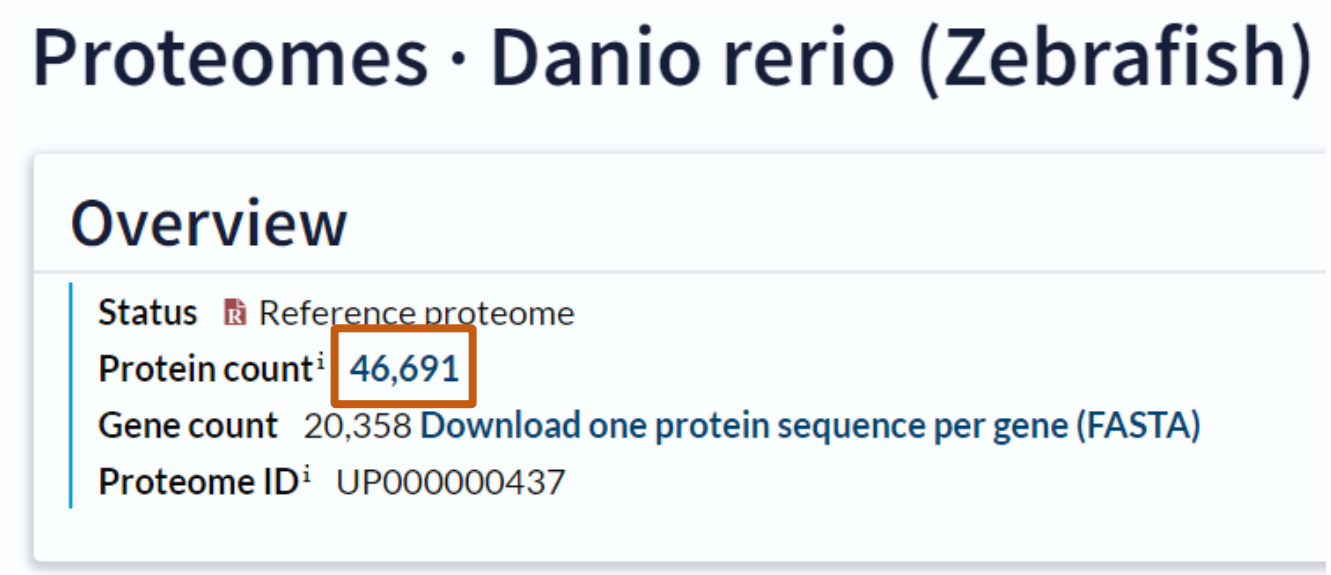
Exercise 1.2: UniProt

- Search for the proteome of the zebrafish (Danio rerio) or your favorite model organism. How many proteins does it contain? How many of those are reviewed?

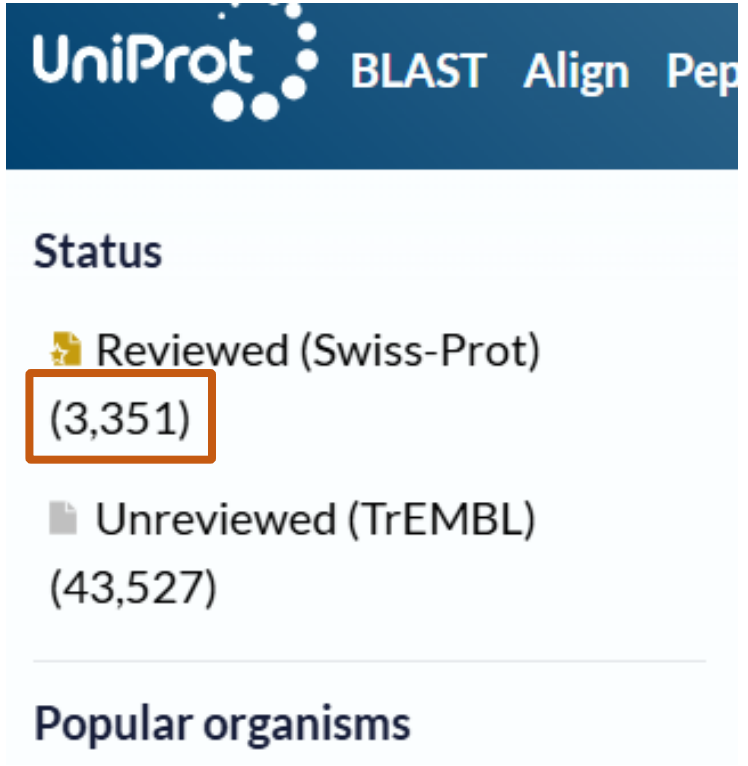


Exercise 1.2: UniProt

- Search for the proteome of the zebrafish (Danio rerio) or your favorite model organism. How many proteins does it contain? How many of those are reviewed?



Snapshot 31/03/25



At the end of this chapter, you should:

- know the databases discussed in the syllabus and what information they contain
- be able to access the databases and retrieve the desired information
- understand how the different databases relate to each other

Section 2

Pairwise alignment

Exercise 2.1: Getting familiar with dot plots

- Use the emboss dottup tool:

https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup

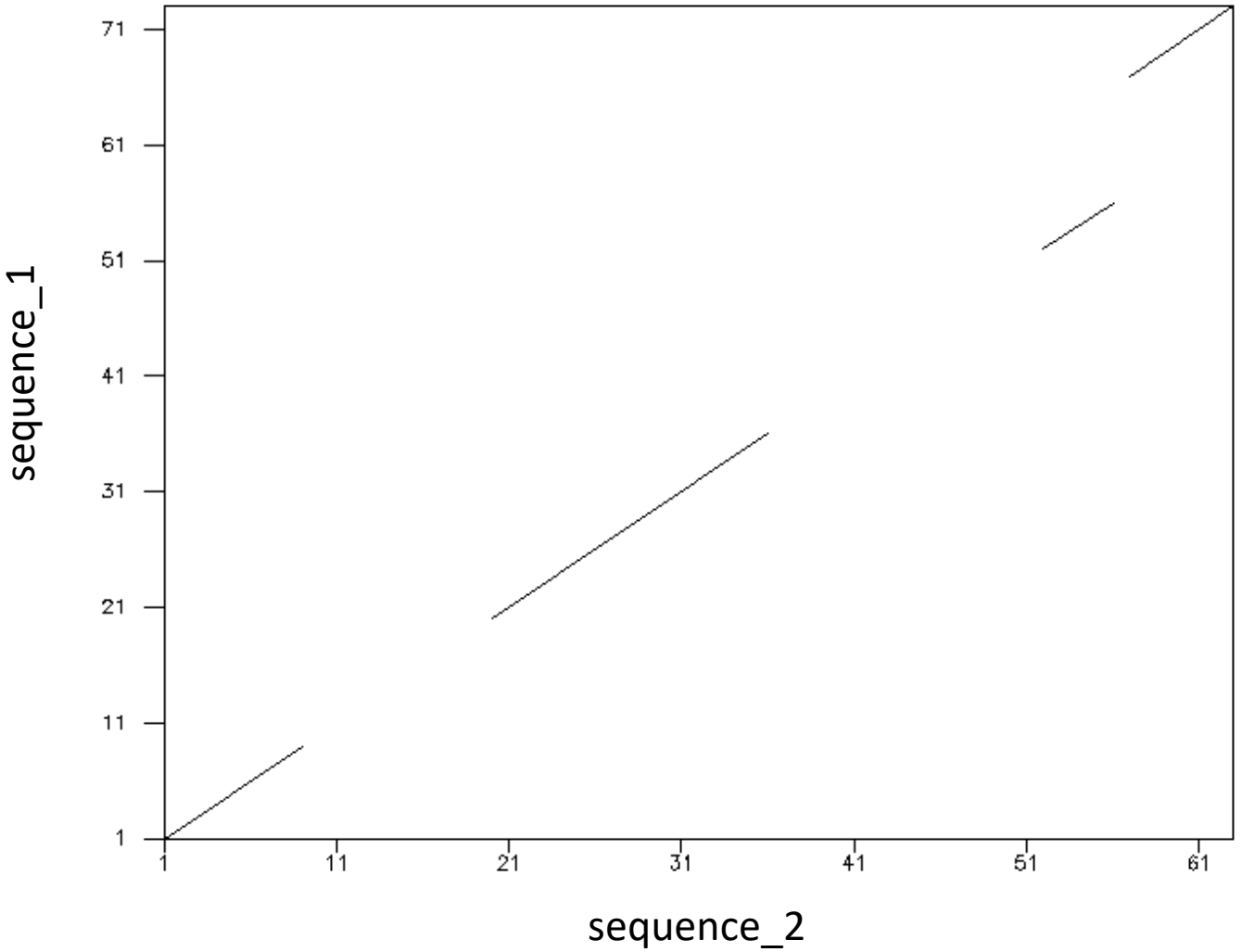
- Make the 4 different alignments described in the course notes (***use a wordsize of 5 for all plots***)

Exercise 2.1: Getting familiar with dot plots

1

>sequence_1
MSETAPAAPASAPAPAEKTPVKK
KARKSAGAAKRKASGPPVSELIT
KAVAASKERSKKALAAAGYDGVS
LAAL

>sequence_2
MSETAPAAPKLIASSKPLPPVKKK
ARKSAGAAKRKAAASWTRRPLA
STVNWSKERSGVSLAAL

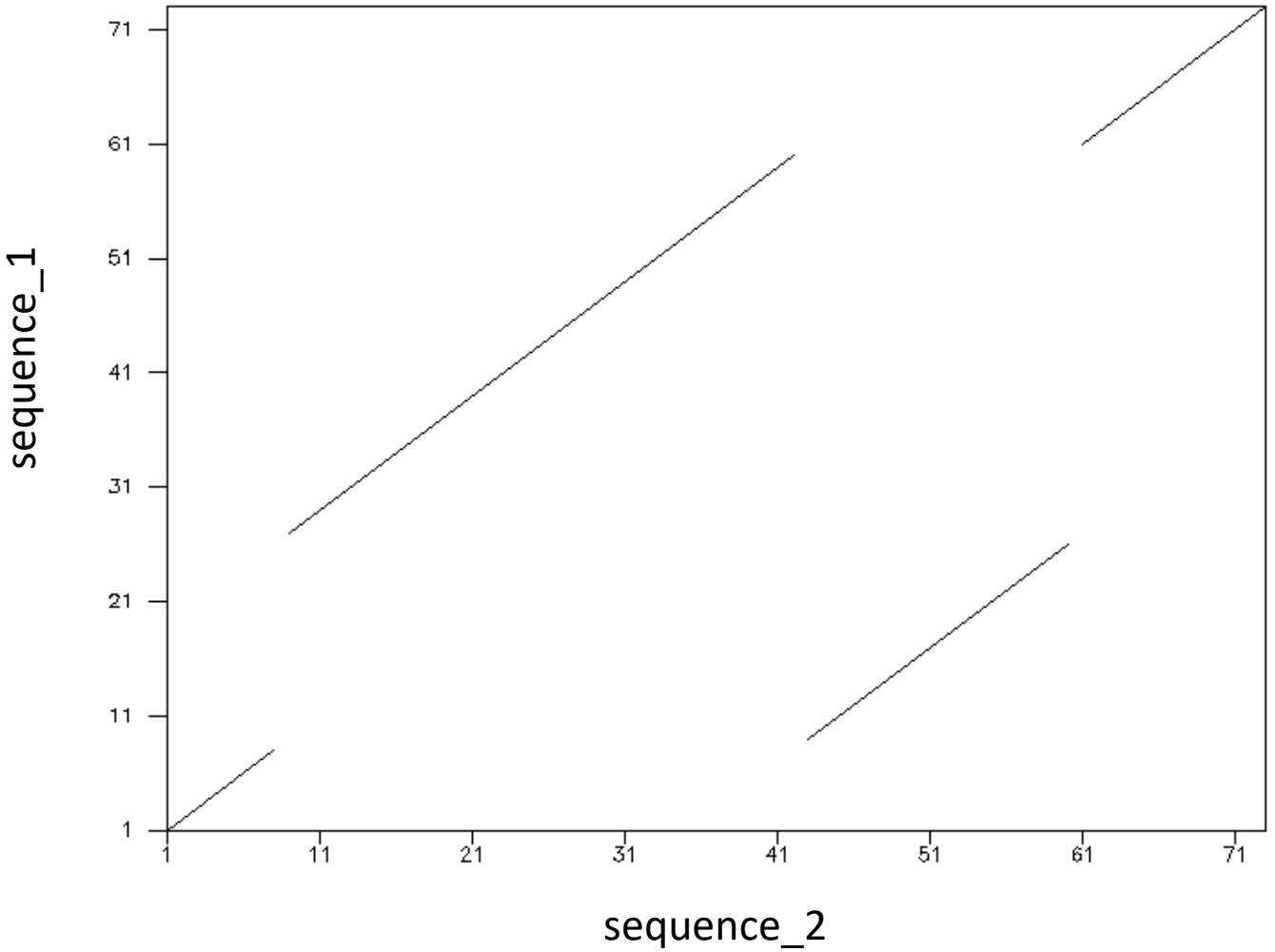


Exercise 2.1: Getting familiar with dot plots

2

>sequence_1
MSETAPAAPASAPAPAEKTPVKKK
ARKSAGAAKRKASGPPVSELITKAV
AASKERSGVSLAALKKALAAAGYD

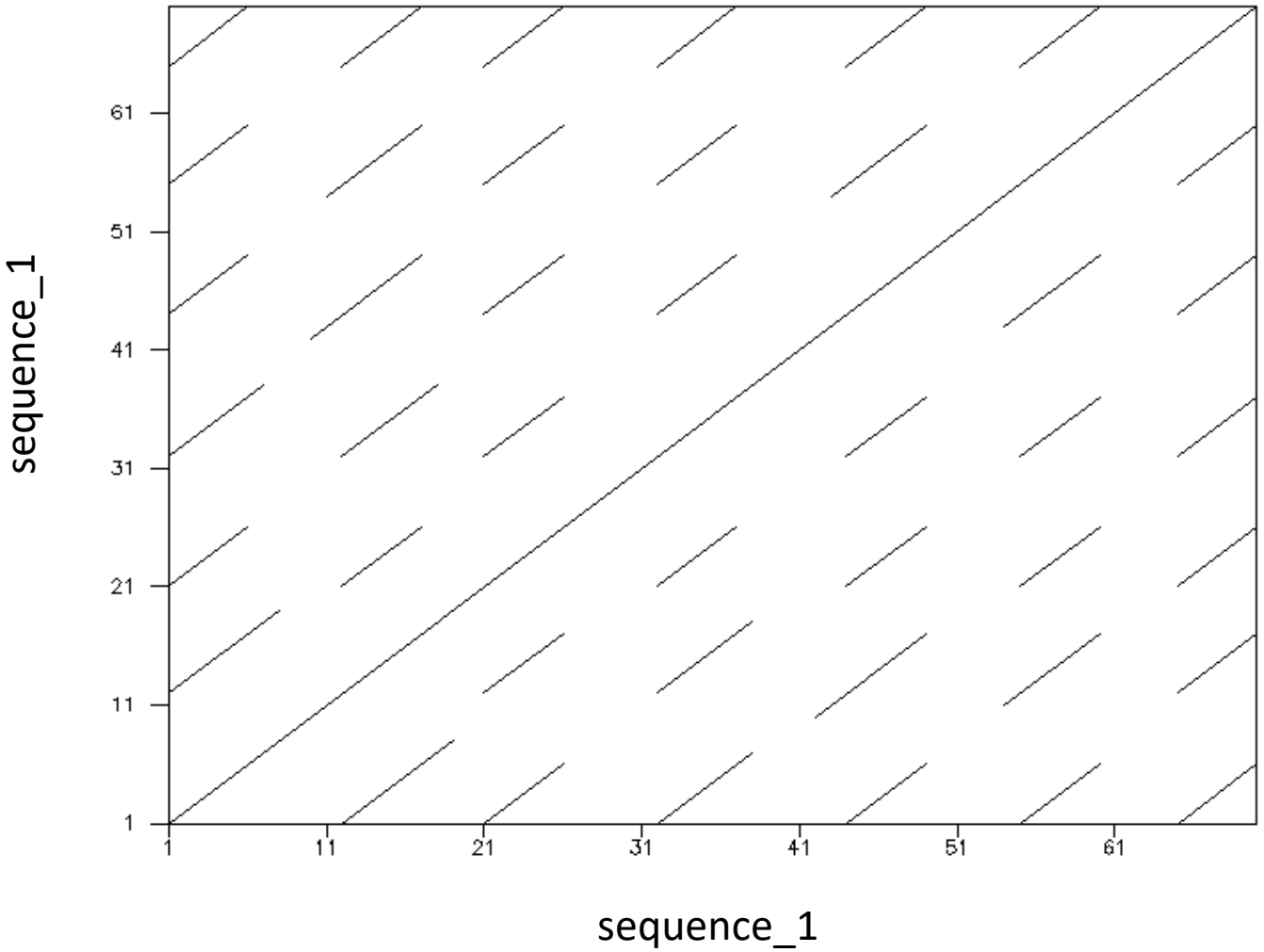
>sequence_2
MSETAPAAKSAGAAKRKASGPPVS
ELITKAVAASKERSGVSLPASAPAPA
EKTPVKKKARAALKKALAAAGYD



Exercise 2.1: Getting familiar with dot plots

3

>sequence_1
MSETAPAAPASMSETAPAARMSETA
PKSAGAMSETAPAKRKASMSETAPG
PPVSMSETAPELITMSETAP

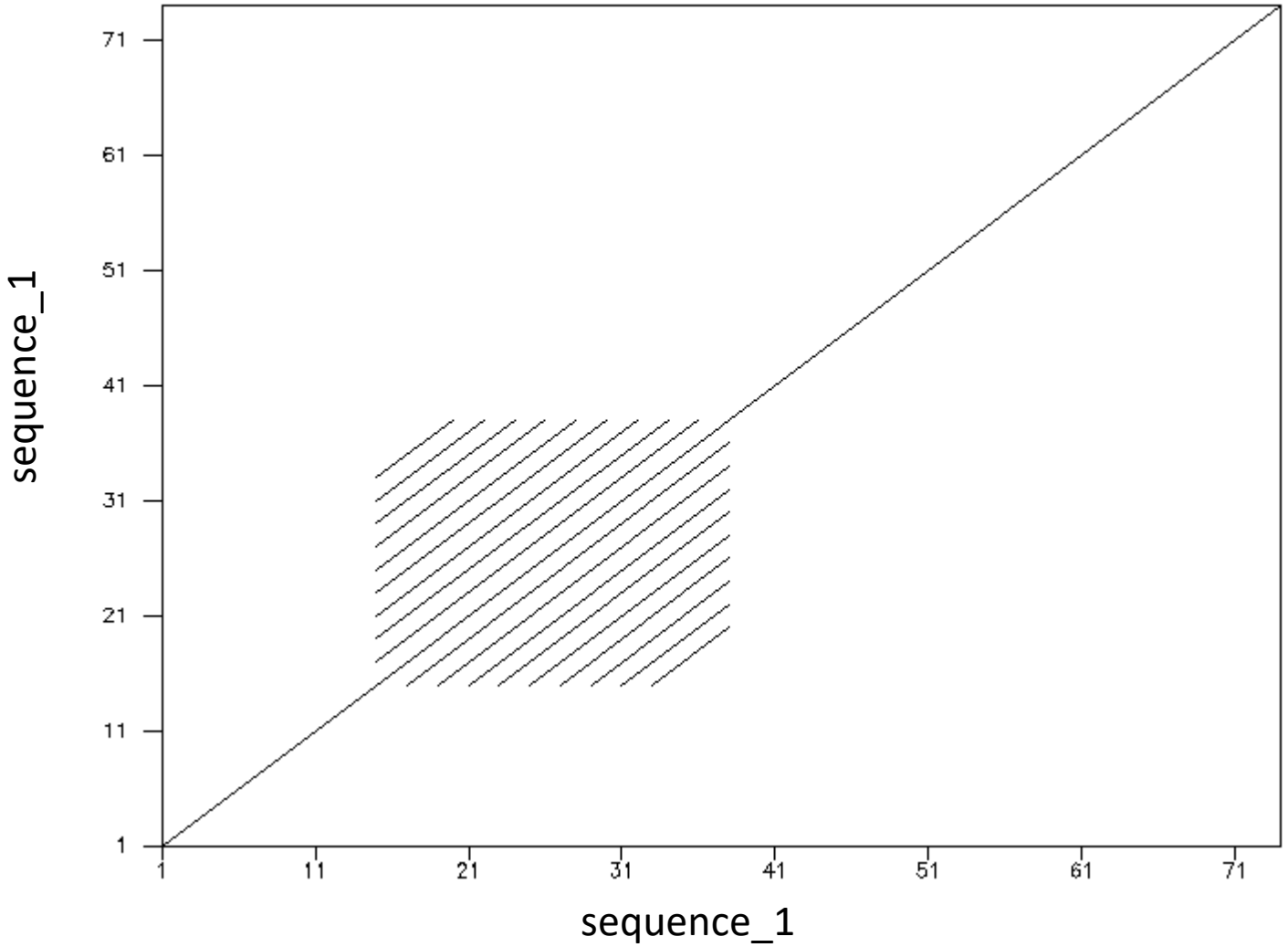


Exercise 2.1: Getting familiar with dot plots

4

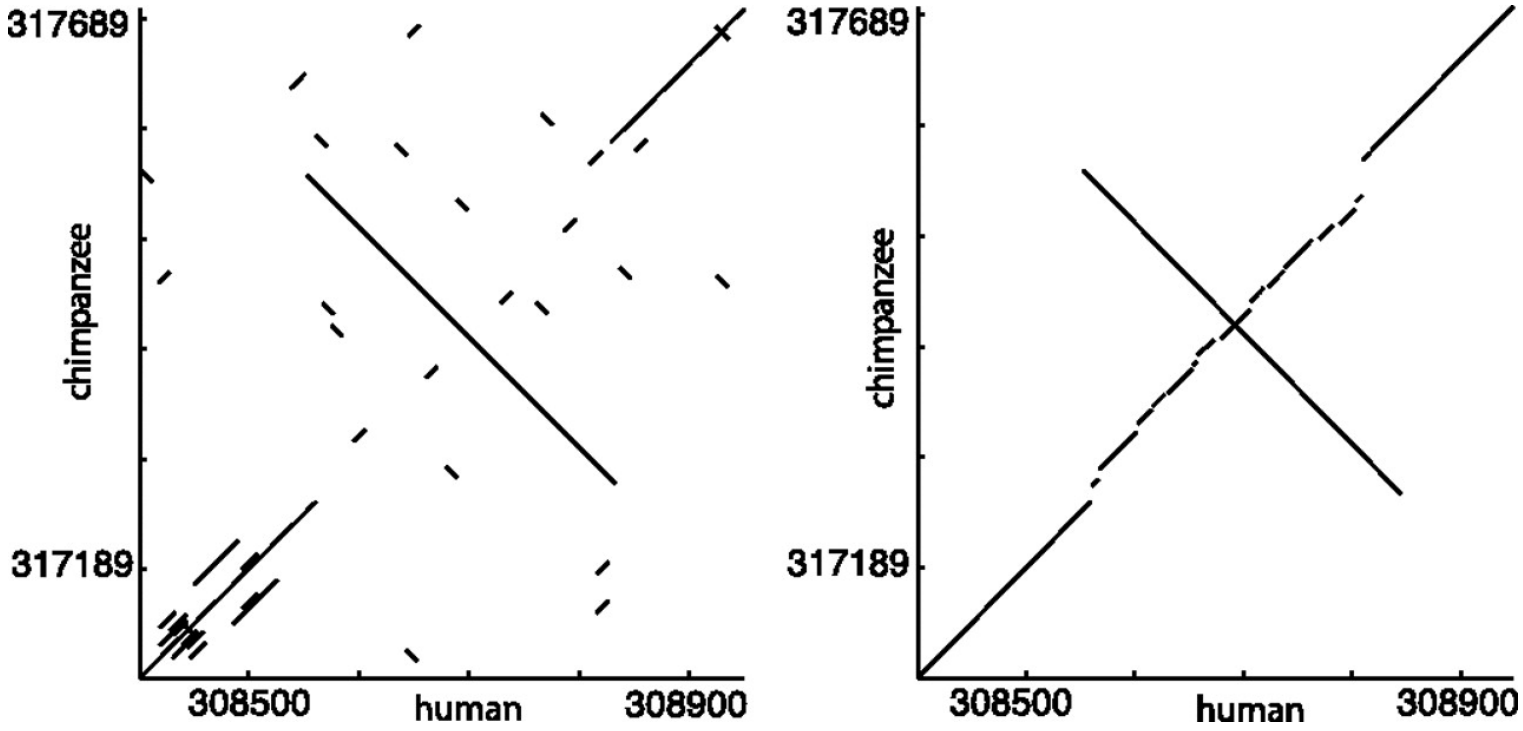
>sequence_1
ASAPAPAEKTPVKKKAKAKAKAKAKAKAK
AKAKAKAKARKSAGAAKRKASGPPVSELIT
KAVAASKERSGVSLA

>sequence_1
ASAPAPAEKTPVKKKAKAKAKAKAKAKAK
AKAKAKAKARKSAGAAKRKASGPPVSELIT
KAVAASKERSGVSLA



Exercise 2.1: Getting familiar with dot plots

Example of a dot plot with inversion

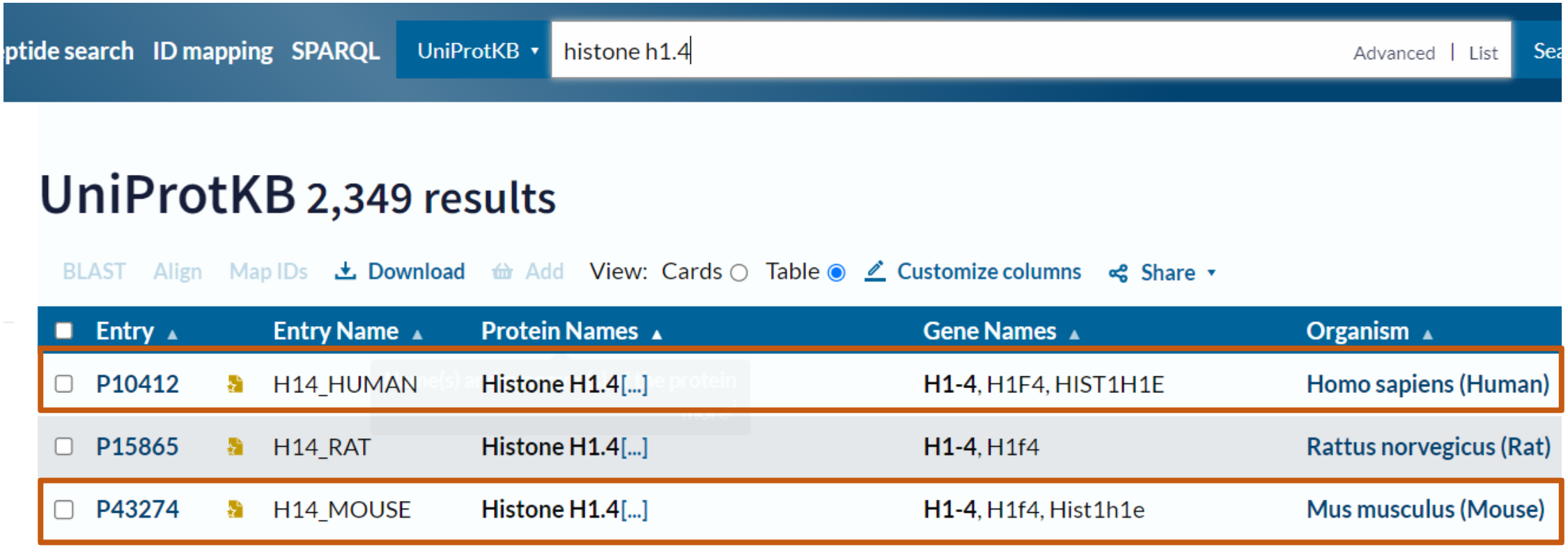


https://doi.org/10.1073/pnas.0603984103

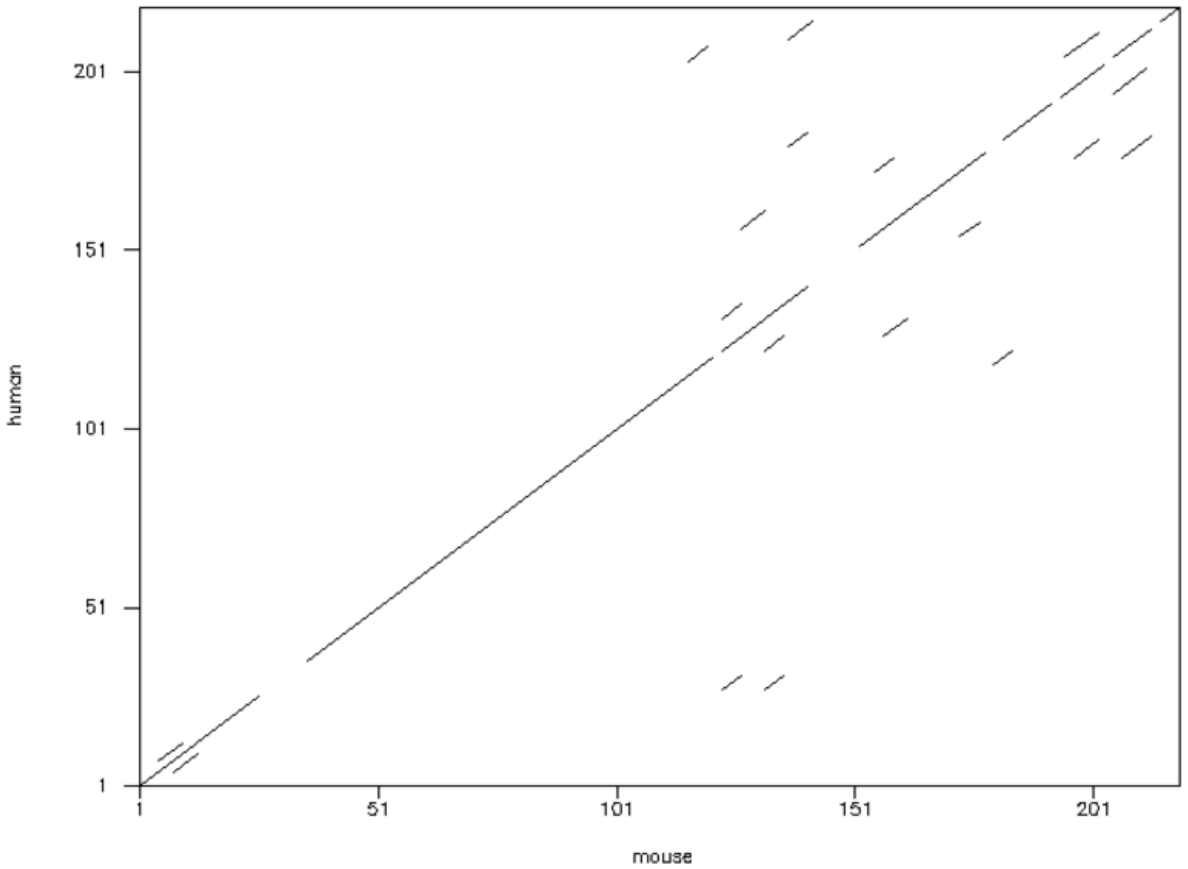
Exercise 2.2: Studying the influence of the word size on dot plots

- From the UniProt website, retrieve the protein sequence from histone H1.4 from human and mouse in FASTA format.
- Use the emboss dottup tool from https://www.ebi.ac.uk/jdispatcher/seqstats/emboss_dottup to construct a dotplot. Do you see similarities between both sequences?
- Vary the word size from 5 to 10. What happens and why?

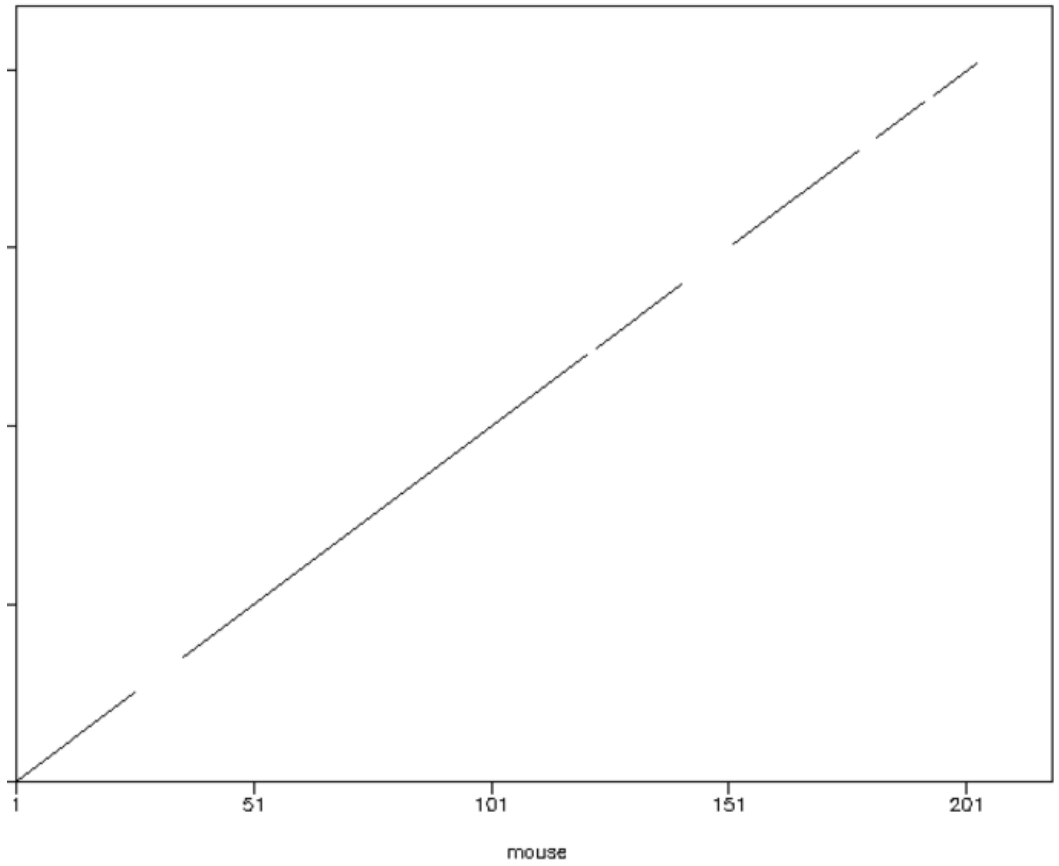
Exercise 2.2: Studying the influence of the word size on dot plots



Exercise 2.2: Studying the influence of the word size on dot plots



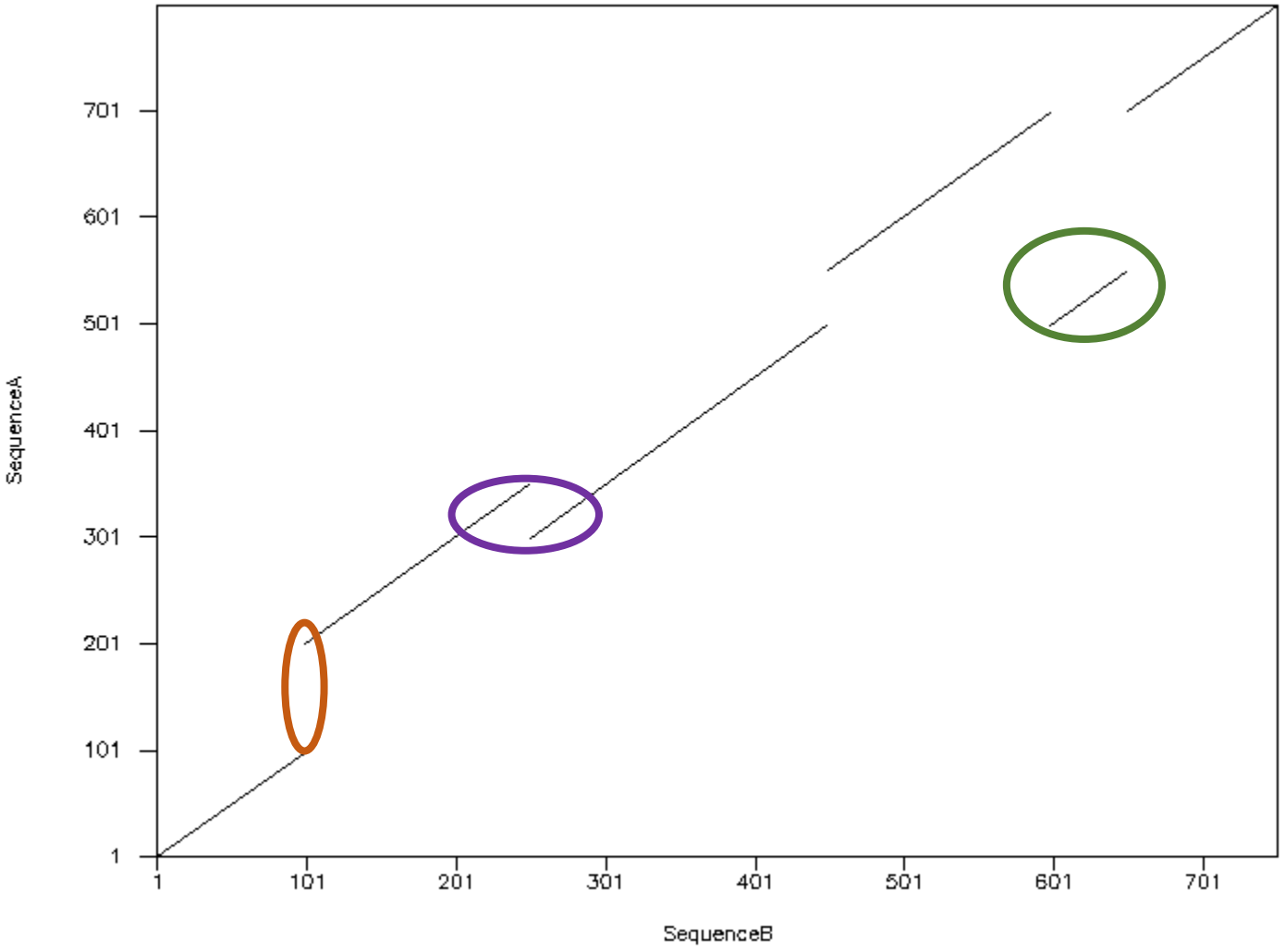
Wordsize 5



wordsize 10

Exercise 2.3: Read a dotplot

Draw the dotplot of sequences A en B(wordsize 10) and describe what happened to sequence A to get to sequence B.



- Region 100-200 deleted in B
- Region 300-350 in A
  is duplicated in B
- Region 500-550 in A
  is translocated to 600-650 in B.

Exercise 2.4: Exploring the NW and SW algorithms

Using the default parameters, perform a global and local alignment of following two sequences.

```
>sequence 1
GLVGEIIGR
>sequence 2
GAVGLIIVR
```

Explanation of the algorithms can be found in the main course material. It is important that you understand these algorithms and are able to calculate the score of a global and local alignment.

Exercise 2.5: Selecting the proper alignment tool

For this exercise you will need the Needle and Water tools at https://www.ebi.ac.uk/jdispatcher/psa
(Here, you can use the default settings of the alignment tools. You will get more insight in the parameters in the following exercise).

(1) Search for conserved residues in the human UDP-galactose transporter-related protein 1 and UDP-galactose transporter homolog 1 from Neosartorya fumigata.

(2) Compare the protein sequences of the human and mouse B-cell translocation gene 1 protein using a pairwise alignment. How do these sequences differ?

Exercise 2.5: Selecting the proper alignment tool (part 1)

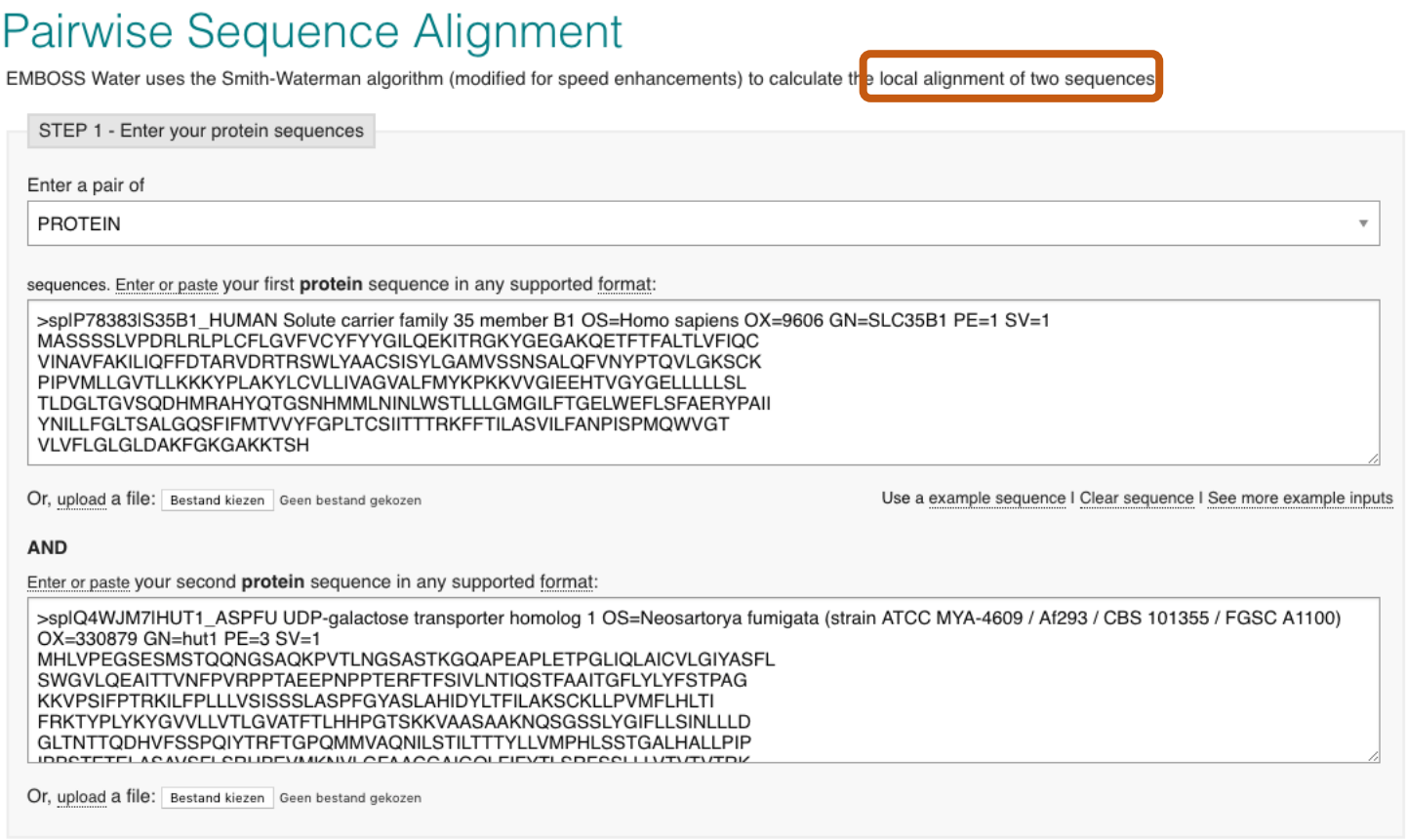
Step 1: retrieve the protein sequences from Uniprot

☐ P78383	S35B1_HUMAN	**Solute carrier family 35 member B1**, ATP/ADP exchanger ER, AXER, Endoplasmic reticulum ATP/ADP translocase, UDP-galactose transporter-related protein 1, UGTrel1	**SLC35B1**, UGTREL1	Homo sapiens (Human)

☐	Entry ▲	Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
☐	Q4WJM7	HUT1_ASPFU	**UDP-galactose transporter homolog 1**	**hut1**, AFUA_1G05440	Neosartorya fumigata (strain ATCC MYA-4609 / Af293 / CBS 101355 / FGSC A1100) (Aspergillus fumigatus)	415 AA

Exercise 2.5: Selecting the proper alignment tool (part 1)

Step 2: choose the right alignment method



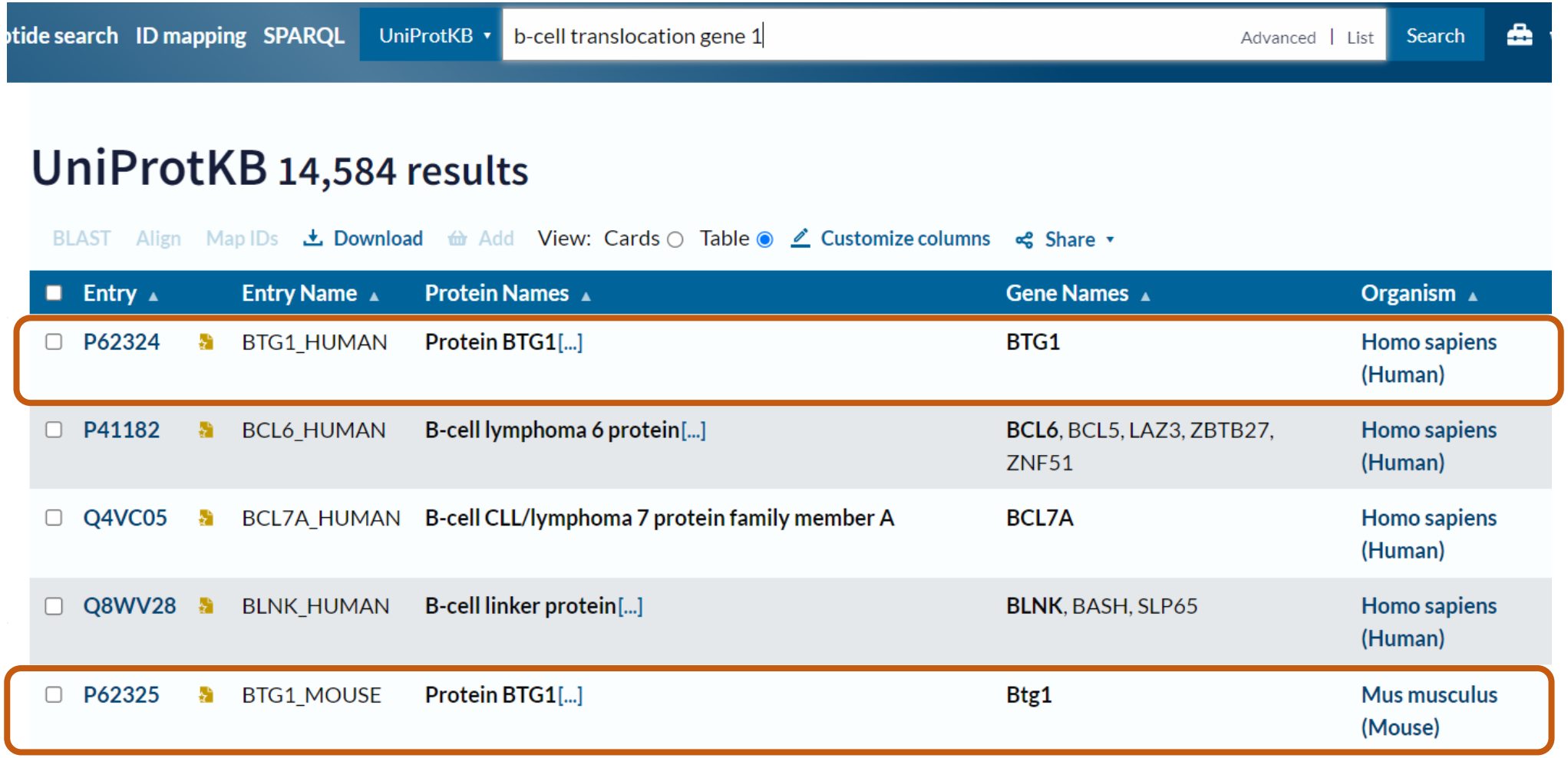
Exercise 2.5: Selecting the proper alignment tool (part 1)

Step 3: interpret the results

```
S35B1_HUMAN       9 PDRLRLPLCFLGVFVCYFYYGILQEKITRGKY--------GEGAKQETFT     50
                    |..::|.:|.||::..:..:|:|||.||...:        ......|.||
HUT1_ASPFU       43 PGLIQLAICVLGIYASFLSWGVLQEAITTVNFPVRPPTAEEPNPPTERFT     92

S35B1_HUMAN      51 FALTLVFIQCVINAVFAKILIQFFDTA-----RVDRTRSWLYAACSISYL     95
                    |::.|..||....|:...:.:.|...|     .:..||..|:....:|..
HUT1_ASPFU       93 FSIVLNTIQSTFAAITGFLYLYFSTPAGKKVPSIFPTRKILFPLLLVSIS    142

S35B1_HUMAN      96 GAMVS--SNSALQFVNYPTQVLGKSCKPIPVMLLGVTLLKKKYPLAKYLC    143
                    .::.|  ..::|..::|.|.:|.||||.:|||.|.:|:.:|.|||.||..
HUT1_ASPFU      143 SSLASPFGYASLAHIDYLTFILAKSCKLLPVMFLHLTIFRKTYPLYKYGV    192

S35B1_HUMAN     144 VLLIVAGVALF-MYKP---KKV-VGIEEHTVG---YGELLLLLSLTLDGL    185
                    |||:..|||.| ::.|   ||| ....::..|   ||..||.::|.||||
HUT1_ASPFU      193 VLLVTLGVATFTLHHPGTSKKVAASAAKNQSGSSLYGIFLLSINLLLDGL    242

S35B1_HUMAN     186 TGVSQDHMRAHYQ-----TGSNHMMLNINLWSTLLLGMGILF------TG    224
                    |..:|||:.:..|     ||. .||:..|:.||:|....:|.      ||
HUT1_ASPFU      243 TNTTQDHVFSSPQIYTRFTGP-QMMVAQNILSTILTTTYLLVMPHLSSTG    291

S35B1_HUMAN     225 ----------------ELWEFLSFAERYPAIIYNILLFGLTSALGQSFIF    258
                                  ||...:||..|:|.::.|:|.|....|:||.|||
HUT1_ASPFU      292 ALHALLPIPIPPSTETELASAVSFLSRHPEVMKNVLGFAACGAIGQLFIF    341

S35B1_HUMAN     259 MTVVYFGPLTCSIITTTRKFFTILASVILFANPISPMQWVGTVLVFLGLG    308
                    .|:..|..|....:|.|||..|:|.||..|.:.:|..||:|..|||.|:|
HUT1_ASPFU      342 YTLSRFSSLLLVTVTVTRKMLTMLLSVFWFGHTLSAGQWLGIGLVFGGIG    391

S35B1_HUMAN     309 LDAKFGKGAKKT    320
                    .:|...|..|::
HUT1_ASPFU      392 AEAVVQKREKQS    403
```

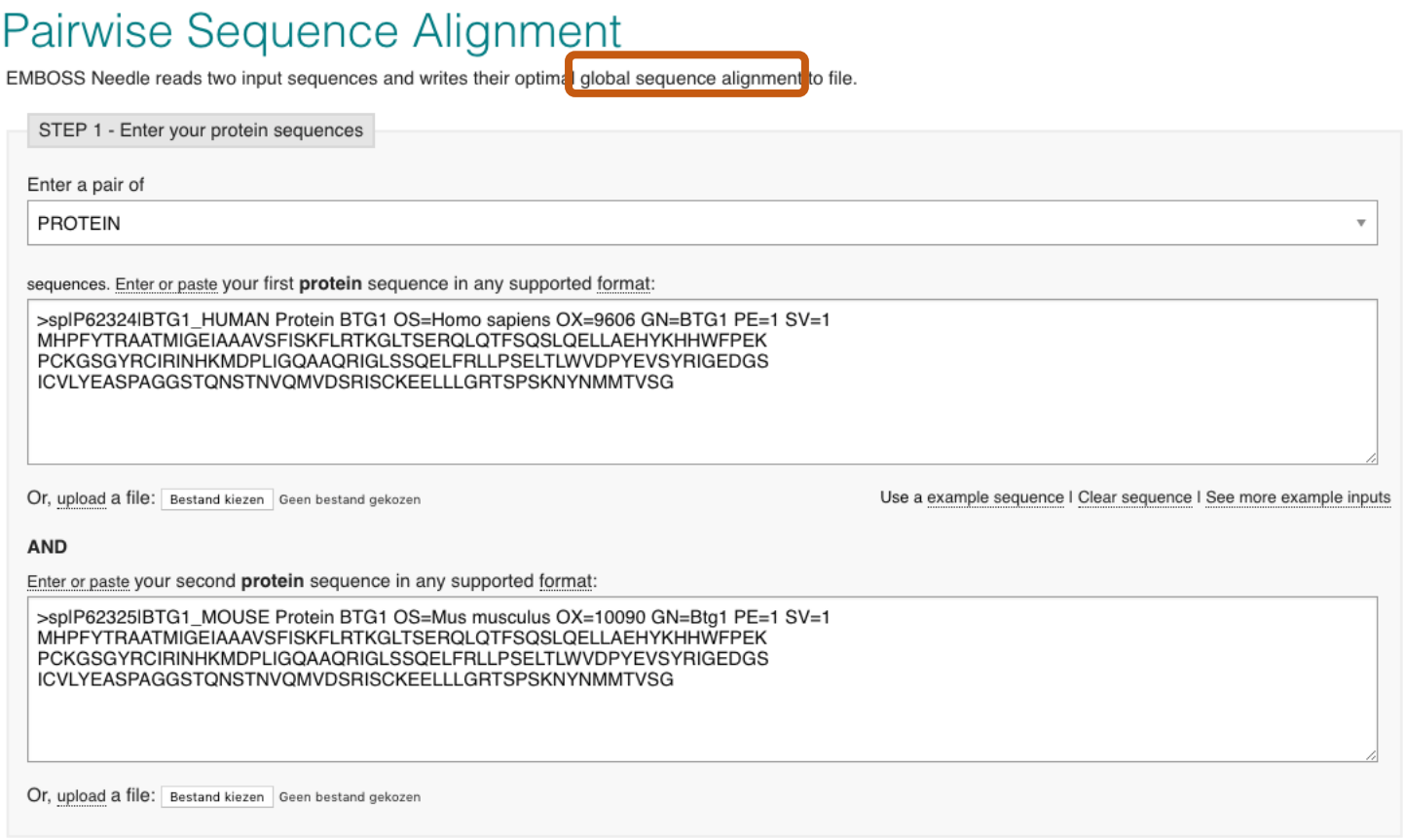
Exercise 2.5: Selecting the proper alignment tool (part 2)

Step 1: retrieve the protein sequences from Uniprot



Exercise 2.5: Selecting the proper alignment tool (part 2)

Step 2: choose the right alignment method



Exercise 2.5: Selecting the proper alignment tool (part 2)

Step 3: interpret the results

```
#=======================================
#
# Aligned_sequences: 2
# 1: BTG1_HUMAN
# 2: BTG1_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 171
# Identity:     171/171 (100.0%)
# Similarity:   171/171 (100.0%)
# Gaps:           0/171 ( 0.0%)
# Score: 892.0
#
#
#=======================================

BTG1_HUMAN         1 MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAE     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE         1 MHPFYTRAATMIGEIAAAVSFISKFLRTKGLTSERQLQTFSQSLQELLAE     50

BTG1_HUMAN        51 HYKHHWFPEKPCKGSGYRCIRINHKMDPLIGQAAQRIGLSSQELFRLLPS    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE        51 HYKHHWFPEKPCKGSGYRCIRINHKMDPLIGQAAQRIGLSSQELFRLLPS    100

BTG1_HUMAN       101 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDSRISCK    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
BTG1_MOUSE       101 ELTLWVDPYEVSYRIGEDGSICVLYEASPAGGSTQNSTNVQMVDSRISCK    150

BTG1_HUMAN       151 EELLLGRTSPSKNYNMMTVSG    171
                     |||||||||||||||||||||
BTG1_MOUSE       151 EELLLGRTSPSKNYNMMTVSG    171
```

Exercise 2.5: Selecting the proper alignment tool

Summary

Global alignment:

- align two sequences in their entirety (start to end)
- search for differences in similar sequences with similar length

Local alignment:

- searches for a partial overlap somewhere in the sequences
- find conserved residues in less similar sequences

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

- Align the human Histone H1.1 protein and the Histone-like nucleoid-structuring protein (DNA-binding protein H-NS) from E. coli using a local alignment tool.

- Different sets of parameters

- Influence of gap opening penalty, gap extension penalty, substitution matrix?

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

- Sequences:

>sp|Q02539|H11_HUMAN Histone H1.1 OS=Homo sapiens OX=9606 GN=HIST1H1A PE=1 SV=3
MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKKPAGPSVSELIVQAASSSKERGGVSLAALKKALAAAGY
DVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKASSVETK PGASKVATKTKATGASKKLKKATGASKKSV
KTPKKAKKPAATRKSSKNPKKPKTVKPKKV AKSPAKAKAVKPKAAKARVTKPKTAKPKKAAPKKK

>sp|P0ACF8|HNS_ECOLI DNA-binding protein H-NS OS=Escherichia coli (strain K12) OX=83333 GN=hns PE=1 SV=2
MSEALKILNNIRTLRAQARECTLETLEEMLEKLEVVVNERREEESAAAAEVEERTRKLQQYREMLIADGIDPNEL
LNSLAAVKSGTKAKRAQRPAKYSYVDENGETKTWTGQGRTPAVIK KAMDEQGKSLDDFLIKQ

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

blosum62, gap opening penalty of 10, gap extension penalty of 0.5

```
#=======================================
#
# Aligned_sequences: 2
# 1: H11_HUMAN
# 2: HNS_ECOLI
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 87
# Identity:      19/87 (21.8%)
# Similarity:    35/87 (40.2%)
# Gaps:           7/87 ( 8.0%)
# Score: 41.0
#
#
#=======================================

H11_HUMAN        118 ETKPGASKVATKTKATGASKKLKKATGA-------SKKSVKTPKKAKKPA    160
                     |....|::|..:|:.....:::..|.|.       |..:||:..|||:..
HNS_ECOLI         43 EESAAAAEVEERTRKLQQYREMLIADGIDPNELLNSLAAVKSGTKAKRAQ     92

H11_HUMAN        161 ATRKSSKNPKKPKTVKPKKVAKSPAKAKAVKPKAAKA    197
                     ...|.|...:..:|.......:::||..|....:..|:
HNS_ECOLI         93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKS    129
```

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

blosum62, **gap opening penalty of 5**, gap extension penalty of 0.5

```
#=======================================

H11_HUMAN          45 ELIVQ--------AASSSKERGGVSLAALKKALAAAGYDVEKNNSRIKLG      86
                      |::|.         ||:..:||.. .|...::.|.|.|  ::.|    :|
HNS_ECOLI          34 EVVVNERREEESAAAAEVEERTR-KLQQYREMLIADG--IDPN----EL-      75

H11_HUMAN          87 IKSLVSKGTLVQTKGTGASGSFKLNKKASSVETKPGA-SKV----ATKTK     131
                      :.||.:    |:: ||.|       |:|.    :|.. |.|    .|||
HNS_ECOLI          76 LNSLAA----VKS-GTKA-------KRAQ----RPAKYSYVDENGETKT-     108

H11_HUMAN         132 ATGASKK---LKKATGASKKSV    150
                      .||..:.   :|||.....||:
HNS_ECOLI         109 WTGQGRTPAVIKKAMDEQGKSL    130

#
# Length: 122
# Identity:      33/122 (27.0%)
# Similarity:    51/122 (41.8%)
# Gaps:          41/122 (33.6%)
# Score: 75.0
```

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

blosum62, gap opening penalty of 5, **gap extension penalty of 5**

```
H11_HUMAN         70 AAAGYDVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKASSVET    119
                     :||..:||:...:::...:.|::.| :...:...:..:.|...||...:
HNS_ECOLI         45 SAAAAEVEERTRKLQQYREMLIADG-IDPNELLNSLAAVKSGTKAKRAQ-     92

H11_HUMAN        120 KPGA-SKV-AT-KTKA-TGASK--K-LKKATGASKKSV    150
                     :|.. |.| .. :||. ||..:  . :|||.....||:
HNS_ECOLI         93 RPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKSL    130
```

```
# Length: 88
# Identity:      21/88 (23.9%)
# Similarity:    40/88 (45.5%)
# Gaps:           9/88 (10.2%)
# Score: 44.0
```

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

blosum30, gap opening penalty of 10, gap extension penalty of 0.5

```
H11_HUMAN         79 NNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKK-----ASSVETKPGA    123
                     ||.|.   :::.:.:.||...:.......:.:|::     |:::|::::.:
HNS_ECOLI          9 NNIRT---LRAQARECTLETLEEMLEKLEVVVNERREEESAAAAEVEERT     55

H11_HUMAN        124 SKVATKTKATGASKKLKKATGASKKSVKTPKKAKKPAATRKSSKNPKKPK    173
                     .|::...:..|....:....|..:||:..|||:.:...|.|...:..:
HNS_ECOLI         56 RKLQQYREMLIADGIDPNELLNSLAAVKSGTKAKRAQRPAKYSYVDENGE    105

H11_HUMAN        174 TVKPKKVAKSPAKAKAVKPKAAKA    197
                     |.......::||..|.::.::.|:
HNS_ECOLI        106 TKTWTGQGRTPAVIKKAMDEQGKS    129
```

```
# Length: 124
# Identity:      23/124 (18.5%)
# Similarity:    58/124 (46.8%)
# Gaps:           8/124 ( 6.5%)
# Score: 83.0
```

Exercise 2.6: Studying the effect of the parameters on pairwise alignments

blosum90, gap opening penalty of 10, gap extension penalty of 0.5

```
H11_HUMAN        118 ETKPGASKVATKTKATGASKKLKK------ATG----------ASKKSVK    151
                     |....|:.|..:|      :||::      |.|          |:.||..
HNS_ECOLI         43 EESAAAAEVEERT------RKLQQYREMLIADGIDPNELLNSLAAVKSGT     86

H11_HUMAN        152 TPKKAKKPA    160
                     ..|:|::||
HNS_ECOLI         87 KAKRAQRPA     95
```

```
# Length: 59
# Identity:       15/59 (25.4%)
# Similarity:     24/59 (40.7%)
# Gaps:           22/59 (37.3%)
# Score: 39.5
```

At the end of this chapter, you should:

- be able to construct and interpret dot plots
- understand the algorithms of pairwise alignment and the influence of the different parameters
- understand the fundamental differences between the pairwise alignment methods
- perform and interpret pairwise alignments (with online tools and manually)

Section 3

Multiple sequence alignment

Exercise 3.1: Progressive versus iterative methods

Let's try to make a multiple sequencing alignment of 5 sequences.

Use **Clustal Omega, T-Coffee, MUSCLE and MAFFT** to perform a multiple sequencing alignment. You might notice **differences between the alignments** depending on the tool you use. **Can you explain this? Which one do you think is the best alignment?**

What do the consensus symbols mean in the alignment?

An * (asterisk) indicates positions which have a single, fully conserved residue.
A : (colon) indicates conservation between groups of strongly similar properties - scoring > 0.5 in the Gonnet PAM 250 matrix.
A . (period) indicates conservation between groups of weakly similar properties - scoring =< 0.5 in the Gonnet PAM 250 matrix.

Exercise 3.1: Progressive versus iterative methods

Progressive methods

T-Coffee

```
CLUSTAL W (1.83) multiple sequence alignment

Sequence1       GARFIELDTHELASTF-ATCAT
Sequence2       GARFIELDTHEFAS----TCAT
Sequence3       GARFIELDTHEVERYFASTCAT
Sequence4       --------THEFA-----TCAT
Sequence5       GARFIELDTHEVAS----TCAT
                         ***.      ****
```

Clustal Omega

```
CLUSTAL O(1.2.4) multiple sequence alignment


Sequence4       --------THEFAT-----CAT    9
Sequence3       GARFIELDTHEVERYFASTCAT    22
Sequence1       GARFIELDTHELASTFAT-CAT    21
Sequence2       GARFIELDTHEFASTCAT----    18
Sequence5       GARFIELDTHEVASTCAT----    18
                         ***.
```

Iterative methods

MUSCLE

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Sequence4       --------THE----FA-TCAT
Sequence2       GARFIELDTHE----FASTCAT
Sequence3       GARFIELDTHEVERYFASTCAT
Sequence1       GARFIELDTHELASTFA-TCAT
Sequence5       GARFIELDTHE----VASTCAT
                         ***    .* ****
```

MAFFT

```
CLUSTAL format alignment by MAFFT FFT-NS-i (v7.429)


Sequence1       GARFIELDTHELASTFA-TCAT
Sequence2       GARFIELDTHE----FASTCAT
Sequence5       GARFIELDTHE----VASTCAT
Sequence3       GARFIELDTHEVERYFASTCAT
Sequence4       --------THE----FA-TCAT
                         ***    .* ****
```

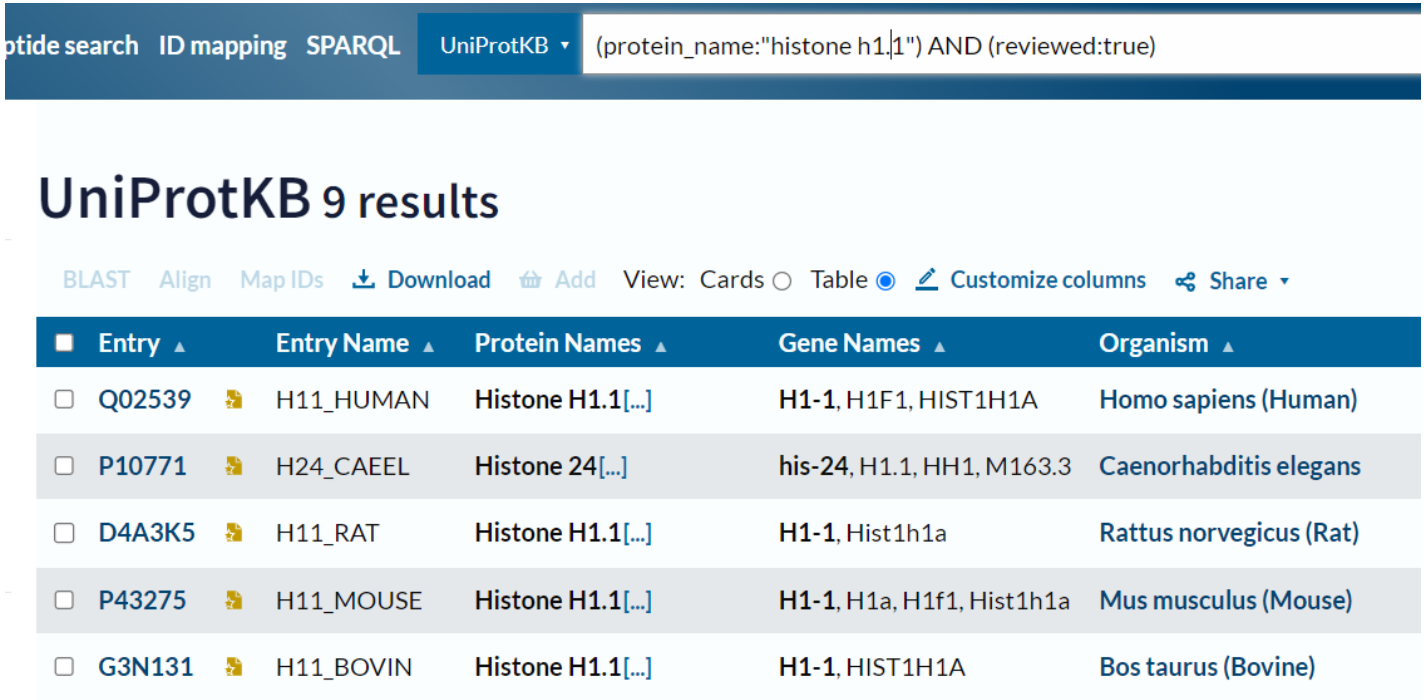
Exercise 3.2: Comparison of orthologous gene products

Perform a multiple alignment to compare the Histone H1.1 proteins from four mammals (i.e. mouse, rat, human and cow).

Are the sequences very similar or dissimilar?

Exercise 3.2: Comparison of orthologous gene products

Step 1: search the protein sequences on Uniprot



Step 2: Select the Human, Rat, Mouse and Bovin results and Download as Fasta (canonical) uncompressed

Exercise 3.2: Comparison of orthologous gene products

Step 2: Perform multiple alignment

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

sp|P43275|H11_MOUSE      MSETAPVAQAASTATEKPAAAKKTKKPAK-AAAPRKKPAGPSVSELIVQAVSSSKERSGV
sp|D4A3K5|H11_RAT        MSETAPVPQPASVAPEKPAATKKTRKPAK-AAVPRKKPAGPSVSELIVQAVSSSKERSGV
sp|Q02539|H11_HUMAN      MSETVPPAPAASAAPEKPLAGKKAKKPAKAAAASKKKPAGPSVSELIVQAASSSKERGGV
sp|G3N131|H11_BOVIN      MSEVALPAPAASTSPEKPSAGKKAKKPAKAAAAAKKKPAGPSVSELIVQAVSSSKERSGV
                         ***..  . .**.:.*** * **:.**** **...***************.******.**

sp|P43275|H11_MOUSE      SLAALKKSLAAAGYDVEKNNSRIKLGLKSLVNKGTLVQTKGTGAAGSFKLNKKA---ESK
sp|D4A3K5|H11_RAT        SLAALKKSLAAAGYDVEKNNSRIKLGLKSLVNKGTLVQTKGTGAAGSFKLNKKA---ESK
sp|Q02539|H11_HUMAN      SLAALKKALAAAGYDVEKNNSRIKLGIKSLVSKGTLVQTKGTGASGSFKLNKKASSVETK
sp|G3N131|H11_BOVIN      SLAALKKALAAAGYDVEKNNSRIKLGLKSLVGKGTLVQTKGTGASGSFKLNKKVASVDAK
                         *******:******************:****.***********:********.   ::*

sp|P43275|H11_MOUSE      AITTKVSVKAKASGAAKKPKKTAG-AAAKKTVKTPKKPKKPAVSKK-TSKSPKKPKVVKA
sp|D4A3K5|H11_RAT        ASTTKVTVKAKASGAAKKPKKTAG-AAAKKTVKTPKKPKKPAVSKKTSSKSPKKPKVVKA
sp|Q02539|H11_HUMAN      PGASKVATKTKATGASKKLKKATG--ASKKSVKTPKKAKKPAATRK-SSKNPKKPKTVKP
sp|G3N131|H11_BOVIN      PTATKVATKTKVTSASKKPKKASGAAAAKKSVKTPKKARKSVLTKK-SSKSPKKPKAVKP
                         . ::**:.*:*.:.*:** **::*  *:**:******..*.. :.* :**.*****.**.

sp|P43275|H11_MOUSE      KKVAKSPAKAKAVKPKASKAKVTKPKTPAKPKKAAPKKK
sp|D4A3K5|H11_RAT        KKVAKSPAKAKAVKPKAAKVKVTKPKTPAKPKKAAPKKK
sp|Q02539|H11_HUMAN      KKVAKSPAKAKAVKPKAAKARVTKPKT-AKPKKAAPKKK
sp|G3N131|H11_BOVIN      KKVAKSPAKAKAVKPKGAKVKVTKPKTAAKPKKAAPKKK
                         ****************.:*..****** ***********
```

At the end of this chapter, you should:

- be able to perform and interpret multiple alignments
- understand the fundamental differences between the discussed alignment methods

Overview question

Kallmann syndrome 6 is a condition that causes hypogonadotropic hypogonadism (HH) and an impaired sense of smell. Mutations in multiple genes have been associated with this disease, including FGF8. Imagine a family of three where the father has already been diagnosed with this syndrome. A genetic analysis concluded that he has a SNP in one of his FGF8 alleles (rs137852659). To see whether his son also carries this mutation, the region around this SNP was sequenced for both alleles. These sequences are stored in FGF8_sequences.fasta in Overview_question. In this exercise, you will examine these sequences to see whether his son has inherited the mutation.

Overview question

1. First, check whether the sequences contain any SNPs.

tip: to find a SNP, you will have to compare the sequences to the reference sequence

2.If you find a SNP, check whether it is identical to the SNP that was detected in one of the alleles of the father.

3. In addition, try to answer following questions using dbSNP:

- On which strand is the gene located?
- What is the influence of the SNP on the protein level?
- Can you find other genomic variations in this gene that are associated with a disease?

Overview question

(1) Identify all the SNPs in the sequences

Step 1: Get the reference sequence of the FGF8 gene

FASTA ▾

Homo sapiens fibroblast growth factor 8 (FGF8), RefSeqGene on

NCBI Reference Sequence: NG_007151.1

GenBank Graphics

```
>NG_007151.1:4957-10940 Homo sapiens fibroblast growth factor 8 (FGF8), RefSeqGene
on chromosome 10
AGCCAGCGGCCACCGGCTCCGGCAGCGGCGCACAGCGATTCGGTGCGGCGCGGCGAGCACGACGTTCCAC
GGGACCCGCGGAGCCGCGTCGTGATCGCCGCCGGCCTCCCGCACCCGCACCCTCTCCGCTCGCGCCCTGC
TCAGCGCGTCCTCCCGCGGCGGCCCGCGGGACGGCGTGACCCGCCGGGCTCTCGGTGCCCCGGGGCCGCG
CGCCATGGGCAGCCCCCGCTCCGCGCTGAGCTGCCTGTGAGTACCGCGCCGCCCTGCCCCGCCACCCGCC
CCCCGCGCCCCTCGCCTCACCCGCGCCTCTCTCTCCCGCCCGCTTTTGTCTCCCACAGGCTGTTGCACTT
GCTGGTCCTCTGCCTCCAAGCCCAGGTGAGGAGGGGTGCGCGGAGGCGGGGGCCGGGCGCGCCGGTGTGA
GACCCGGGTGGGCAGCGCCGGTCGGGGGCACCGGGACTGACTCTGCGGCCGGCGCGGGAGGGCTGAGGGC
ACCTTAGAAACCCAGCCCCGAGCCACCCCGAGGAGGGAGCTGAGGCACAGAGAGGTAGCACCCCTCCTGA
GGTCACACAGCGAGTGAGTGGCCAGGATAGAAACGAAGGTCTCGGAGCCCAGCACTGTCCCCCATGCATC
```

Send to: ▾

- Complete Record
- Coding Sequences
- Gene Features

Choose Destination

- File
- Clipboard
- Collections
- Analysis Tool

Download 1 item.

Format

FASTA

Show GI ☐

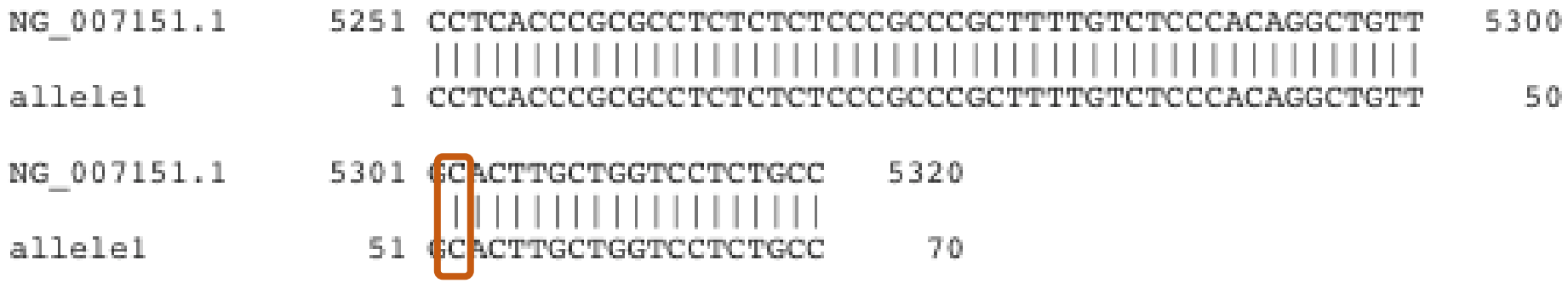
Create File

Overview question

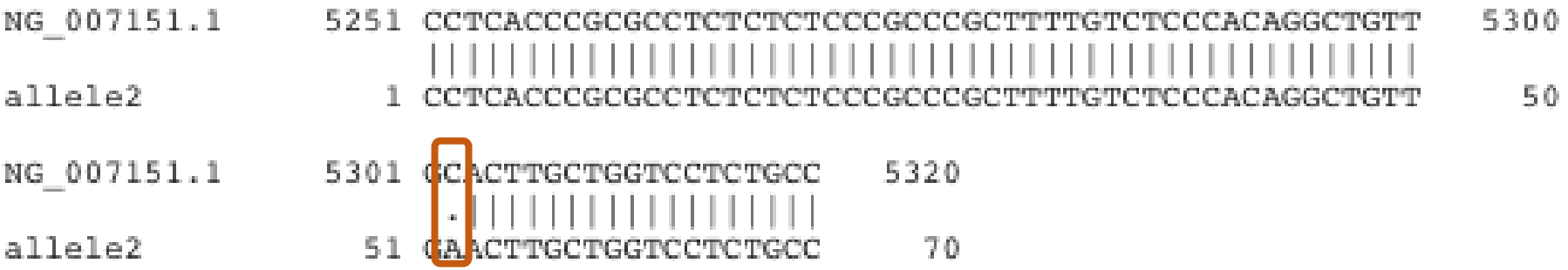
(1) Identify all the SNPs in the sequences

Step 2: Find the genetic variations through a local alignment

Alignment of allele 1 with the reference gene

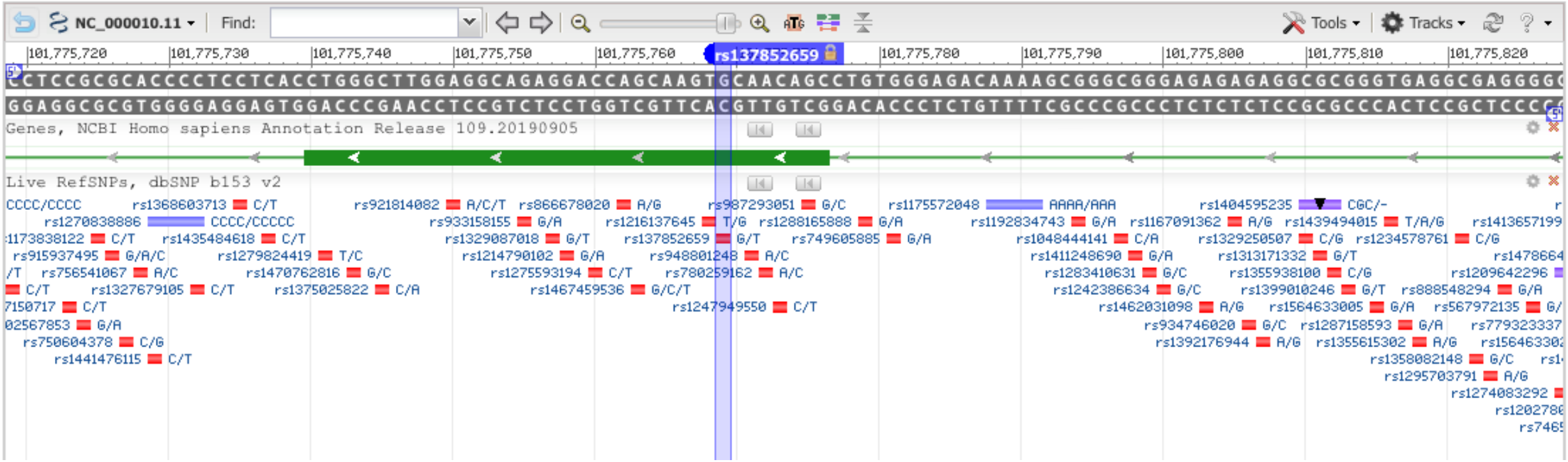


Alignment of allele 2 with the reference gene



Overview question

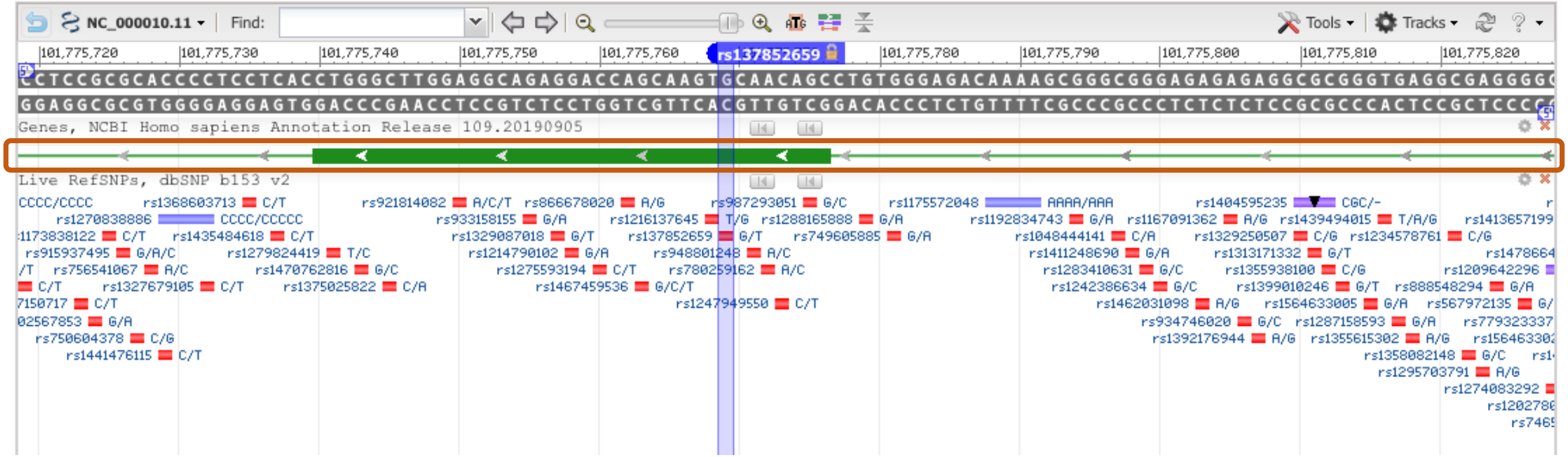
(2) Check whether the mutation in allele 2 was inherited from the father



Overview question

(3) Short questions:

On which strand is the gene located?



Overview question

(3) Short questions:

On which strand is the gene located?

Gene: FGF8, fibroblast growth factor 8 (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
FGF8 transcript variant A	NM_033165.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant B	NM_006119.5:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant E	NM_033164.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant F	NM_033163.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant G	NM_001206389.1:c.	N/A	5 Prime UTR Variant
fibroblast growth factor 8 isoform A precursor	NP_149355.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform B precursor	NP_006110.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform E precursor	NP_149354.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform F precursor	NP_149353.1:p.His14Asn	H (His) > N (Asn)	Missense Variant

Overview question

(3) Short questions:

What is the influence of the SNP on the protein level?

Gene: FGF8, fibroblast growth factor 8 (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
FGF8 transcript variant A	NM_033165.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant B	NM_006119.5:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant E	NM_033164.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant F	NM_033163.4:c.40C>A	H [CAC] > N [AAC]	Coding Sequence Variant
FGF8 transcript variant G	NM_001206389.1:c.	N/A	5 Prime UTR Variant
fibroblast growth factor 8 isoform A precursor	NP_149355.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform B precursor	NP_006110.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform E precursor	NP_149354.1:p.His14Asn	H (His) > N (Asn)	Missense Variant
fibroblast growth factor 8 isoform F precursor	NP_149353.1:p.His14Asn	H (His) > N (Asn)	Missense Variant

Clinical Significance	Reported in ClinVar
Gene : Consequence	FGF8 Missense Variant
Publications	1 citation
Genomic View	See rs on genome

Overview question

(3) Short questions:

Can you find other genomic variations in this gene that are associated with a disease?