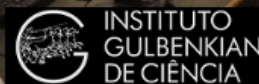


A photograph showing several people in a computer lab or training room. They are seated at desks with multiple computer monitors. Some individuals are looking at the screens, while others are engaged in discussion. The room has a casual, educational atmosphere.

# **GTPB**

The Gulbenkian Training Programme in Bioinformatics  
(Since 1999)

Pedro Fernandes, Organiser



# **Introduction to Bioinformatics**

**23-27 March 2020**

**Practical 1: Databases and Tools**

**Part i) – Aniridia viewed from The NCBI**

**Tuesday 28 April 2020**

## Investigating the gene(s) associated with Aniridia at the NCBI

As a starting point for this exercise, imagine you have a vital interest in discovering and investigating the main human gene responsible for the terrible disease of the eye, **Aniridia**. There are many ways (including **google!**) you could discover what this gene might be. I choose to delve into the vast seas of knowledge so generously proffered by the **The National Center for Biotechnology Information (NCBI)**.

So, begin by going to the **Home Page** of the **The National Center for Biotechnology Information (NCBI)** ("<http://www.ncbi.nlm.nih.gov/>").

You will arrive at a page offering access to the many **NCBI** resources available to you. Currently, you only require to search for genes, specifically those that relate to **Aniridia**, so first set the database selection field of the **Search** facility at the top of your page to **Gene**, set the **Search** field to **Aniridia** and click on the **Search** button.

A fine list of genes will emerge, including those sought. However, our interest is specific to **Human**, so the search should really be organism specific. To do this, one needs to execute an **Advanced** search. So, click on the **Advanced** button of the **Search** tool.

Now you can specify the precise field(s) of the annotation you wish to interrogate. In this case, set the **Disease/Phenotype** field to **Aniridia** and the **Organism** field to **Human**. As the two conditions are linked by **AND**, both must be true for any gene to be listed.

Click on the pretty **Search** button.

Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> <a href="#">WT1</a> ID: 7490	WT1 transcription factor [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (32387775..32435539, complement)	AWT1, GUD, NPHS4, WAGR, WIT-2, WT33	607102
<input type="checkbox"/> <a href="#">PAX6</a> ID: 5080	paired box 6 [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (31784792..31817961, complement)	AN, AN1, AN2, ASGD5, D11S812E, FVH1, MGDA, WAGR	607108
<input type="checkbox"/> <a href="#">TRIM44</a> ID: 54765	tripartite motif containing 44 [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (35662692..35818007)	AN3, DIPB, HSA249128, MC7	612298
<input type="checkbox"/> <a href="#">ELP4</a> ID: 26610	elongator acetyltransferase complex subunit 4 [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (31509767..31790324)	AN, AN2, C11orf19, PAX6NEB, PAXNEB, dJ68P15A.1, hELP4	606985
<input type="checkbox"/> <a href="#">DEL11P13</a> ID: 100528024	Wilms tumor, aniridia, genitourinary anomalies and mental retardation syndrome [ <i>Homo sapiens</i> (human)]		C11DELP13, WAGR	194072

Just a few genes survive. All should really be examined, but this is just an exercise, so trust me ... it is **PAX6** that is the most interesting gene<sup>1</sup>, in this context. This is the one to follow up by clicking on the link to its details.

<sup>1</sup> This despite **WT1** being at the top of the list? This is a relatively new promotion for **WT1**. For years it has been but a close second to **PAX6**. Whilst congratulations are clearly in order, this elevation is jolly inconvenient for the story I wish to reveal. So ... I intend to ignore it!

**PAX6** paired box 6 [ *Homo sapiens* (human) ]


Gene ID: 5080, updated on 5-Apr-2020

From the very top of the page, one learns that the NCBI specific identifier for this **Gene** is a simple number (**5080**, to be precise). Effective, if rather bland and indicative of a tragic lack of flare and imagination!

Table of contents
Summary
Genomic context
Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Pathways from PubChem
Interactions
General gene information
Markers, Clone Names, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links
Locus-specific Databases

There is much information about the gene **PAX6** on this page. One can slide up and down to drink in all the wonders on offer, or there is a **Table of contents** in the top right corner that will transport you directly to the section of your desires.

In the **Summary section**, one discovers that the fine fellows of the pretentiously labelled, **Hugo Gene Nomenclature Committee (HGNC)** suggest the name “**paired box 6**”, to be truncated to the less cumbersome **Symbol** “**PAX6**” when less formal address is deemed appropriate.



Summary

**Official Symbol**

PAX6 provided by [HGNC](#)

**Official Full Name**

paired box 6 provided by [HGNC](#)

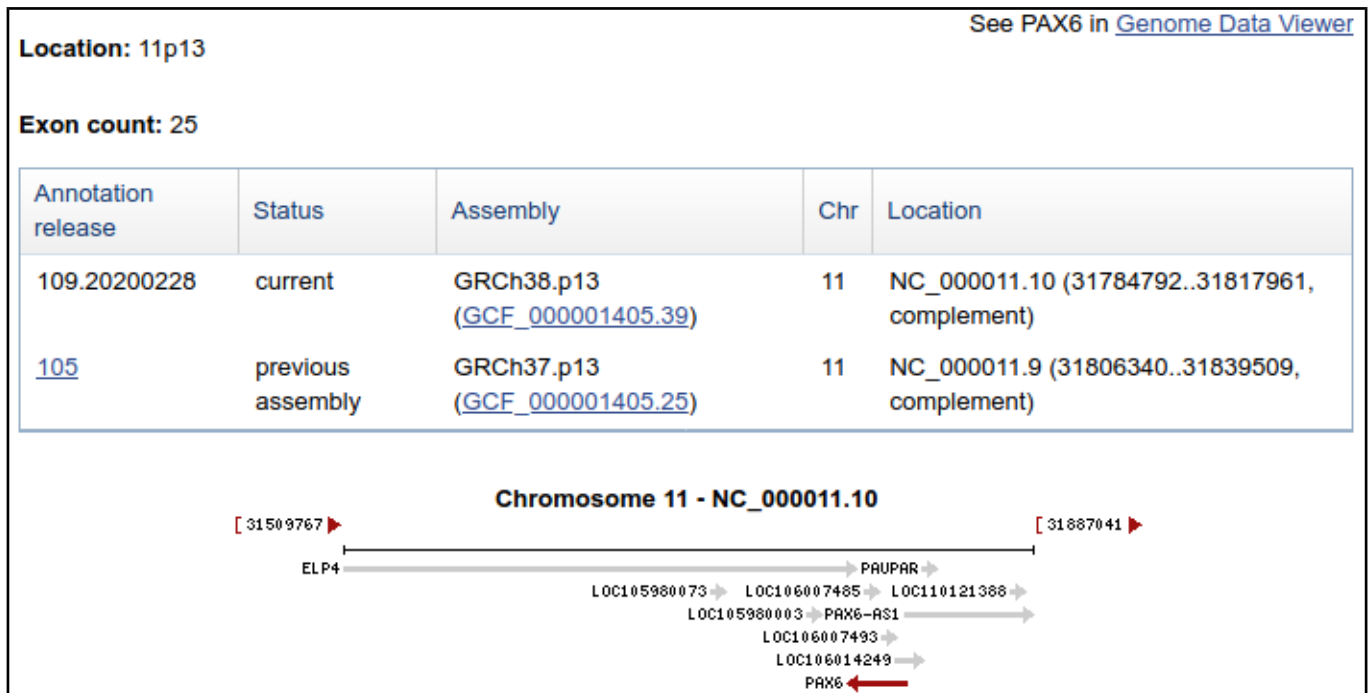
Also, from the **Summary section** one can conclude (concentrating on the features that pertain to this exercise) that:

**Summary** This gene encodes paired box protein Pax-6, one of many human homologs of the *Drosophila melanogaster* gene *prd*. In addition to a conserved paired box domain, a hallmark feature of this gene family, the encoded protein also contains a homeobox domain. Both domains are known to bind DNA and function as regulators of gene transcription. Activity of this protein is key in the development of neural tissues, particularly the eye. This gene is regulated by multiple enhancers located up to hundreds of kilobases distant from this locus. Mutations in this gene or in the enhancer regions can cause ocular disorders such as aniridia and Peter's anomaly. Use of alternate promoters and alternative splicing results in multiple transcript variants encoding different isoforms. Interestingly, inclusion of a particular alternate coding exon has been shown to increase the length of the paired box domain and alter its DNA binding specificity. Consequently, isoforms that carry the shorter paired box domain regulate a different set of genes compared to the isoforms carrying the longer paired box domain. [provided by RefSeq, Mar 2019]

- There are two major domains, a **paired domain** and a **homeobox**, both of which **bind DNA**.
- The gene is a **homologue** of a **Drosophila Melangaster** gene called **prd**.
- This gene is “*key in the development of neural tissues, particularly the eye*”, as eyes are almost universal, it is not surprising that **PAX6** has **homologues** in a wide range of organisms and that **prd** is not the only **PAX6** homologue of the fly.
- The gene regulates **Transcription** (i.e. is a **Transcription Factor**).
- “... *alternative promoters and alternative splicing results in multiple transcript variants encoding different isoforms.*”.
- “... *inclusion of a particular alternative coding exon has been shown to increase the length of the paired box domain and alter its DNA binding specificity*”.

All of these observations will be investigated in the exercises that follow.

From the **Genomic context** section it can be seen that:



- **PAX6** is situated on **Chromosome 11**, band **p13**.
- **PAX6** is on the complementary strand relative to that chosen to represent **Chromosome 11**.
- **ELP4** (another human gene listed as associated with **Aniridia**) is very close, on the opposite strand to **PAX6**.
- There are **25** exons for **PAX6**.
- A number of other features are recorded here. Most are not genes and so we will ignore them for now. However, do note the feature **PAX6-AS1**. This is a **non-coding RNA** that will play a small part in the dance to follow.
- Note also the feature **PAUPAR**. This is its only appearance that we come across at the **NCBI**, but it does seem to have a marginally higher profile elsewhere. Both **PAX6-AS1** and **PAUPAR** are reported as **PAX6 antisense RNAs**, which informs one only slightly.
- Note how the location of the **PAX6** gene has moved slightly between the **current assembly** of the **Human Genome** and the **previous assembly**.

Annotation release	Status	Assembly	Chr	Location
109.20200228	current	GRCh38.p13 ( <a href="#">GCF_000001405.39</a> )	11	NC_000011.10 (31784792..31817961, complement)
<a href="#">105</a>	previous assembly	GRCh37.p13 ( <a href="#">GCF_000001405.25</a> )	11	NC_000011.9 (31806340..31839509, complement)

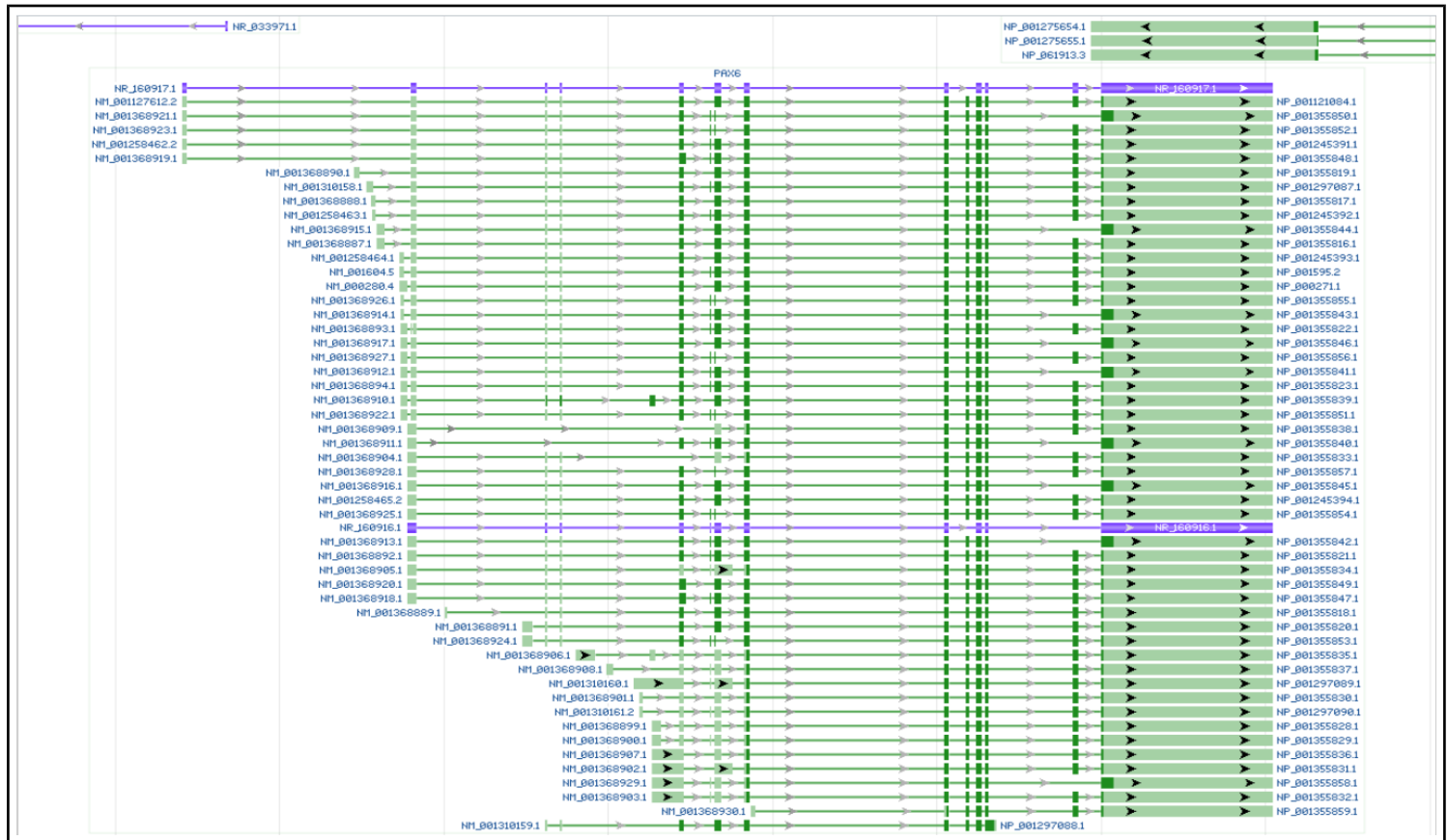
This demonstrates clearly that, of course the **Human Genome** is not **DONE! FINISHED! FIXED forever!** It is a consensus of the **Genomes** of a number of individuals and is recalculated regularly. **Genes** therefore appear to “move”, even change shape! It is even the case that some areas of the **Genome**, particularly around **Centromeres**, have proven exceptionally difficult to sequence and are, even now, represented in the sequence databases as long runs of Ns (N indicating the presence of a base of unknown type).

**25 Exons?** Jolly good, but I really wanted to know how many **Transcripts** there were according to the **NCBI**. That is, how many different ways it is thought that nature spliced the **25** exons together. I would also like to discover how many distinct **isoforms** the **NCBI** imagines to result from however many **Transcripts**. I proceed with impatience!

All of these observations will play a role in the exercises that follow.



Move down to the **Genome regions, transcripts and products** section. The **PAX6** genomic region, as interpreted by the **NCBI Genome Database**, is displayed for your delectation.



The whole width of the display represents the entire **PAX6** region of **Chromosome 11**. Each line represents a **PAX6 Transcript**.

The top **PAX6** line represents one of the two the **non-coding Transcripts** that this database associates with **PAX6**. A **non-coding Transcript** has a name (**Accession Code**) that begins **NR\_** (**Non-coding RNA**). The **Accession Code** is displayed to the left of the transcript line (in this case, **NR\_160917**, the **.1** at the end is the version number). The **Accession Code** for the **Protein Product** of the **Transcript** is displayed at the right hand end of the **Transcript** line (in this case it is blank, of course). The pretty **blue** blobs represent the **Introns**, the equally attractive **blue** lines joining the blobs, represent the **Exons**.

The choice of the first two letters of the **Accession codes** you see here reflect the status of the **Genes**, **Transcripts**, or **Proteins** they represent. Here we see, **NR\_**, **NM\_** and **NP\_** representing **non-coding Transcripts**, **curated coding Transcripts** and **protein products of NM\_ Transcripts** respectively. There are, for example, also **Accessions codes** prefixed by **XR\_**, **XM\_** and **XP\_** representing **predicted non-coding Transcripts**, **predicted coding Transcripts** and **protein products of XM\_ Transcripts** respectively. A full list of **RefSeq Accession code** prefixes can be found [Here](#).

All the **PAX6 Transcripts** shown here, excepting the two **blue NR\_ Transcripts**, are **curated coding Transcripts**. Each **coding Transcript** is represented by a **Transcript line** showing **CoDing Sequence (CDS)** **Introns** as **dark green** blobs, **Untranslated Regions (UTRs)** in **Introns** as **lighter green** blobs, joined together by **green** lines representing the **Exons**.

Note that each **coding Transcript** is associated with a unique **Protein Product**, the **Accession Code** of which is displayed at the right hand end of each **coding Transcript** line. This does *not* mean that every **coding Transcript** generates a different **Protein Product**. It just means that this database finds it convenient to represent **Protein Products** as if they were all distinct. There are, in fact, far fewer **Protein Isoforms** than there are **coding Transcripts**, as we will discover.

Note the three **curated coding Transcripts** in the top right hand corner of the graphic.

Hover over the any one of them and an large grey box full of fascinating facts will bounce forth from nowhere!

It should be clear that these are three **Transcripts** of the **ELP4** gene that was noted when looking at the **Genomic context** section.

Gene: ELP4  
Title: elongator acetyltransferase complex subunit 4  
RNA title: mRNA-elongator acetyltransferase complex subunit 4, transcript variant 1  
Protein title: elongator complex protein 4 isoform 3  
Merged features: NM\_001275655.1 and NP\_001275655.1  
Location: 31,509,767..31,790,324  
[Length]  
Span on NC\_000011.10: 280,558 nt  
Aligned length: 8,521 nt  
CDS length: 1,608 nt  
Protein length: 535 aa  
Download: [NP\\_001275655.1](#), [NM\\_001288726.2](#)  
Links & Tools  
View GenID: [26610 \(ELP4\)](#)  
View HGNC: [1171](#)  
View MIM: [606985](#)  
BLAST Protein: [NP\\_001275655.1](#)  
BLAST mRNA: [NM\\_001288726.2](#)  
BLAST Genome-specific: [NC\\_000011.10 \(31,509,767..31,790,324\)](#)  
FASTA View: [NC\\_000011.10 \(31,509,767..31,790,324\)](#)  
GenBank View: [NC\\_000011.10 \(31,509,767..31,790,324\)](#)

Note also an enigmatic **non-coding Gene** with just one **Transcript** called, endearingly, **NR\_0339711**. Hover over the **NR\_0339711** transcript line and a new box of tasteful grey will sally forth telling all there is to know about the enigmatic **NR\_0339711**!

A swift glance will be sufficient for you to see that **NR\_0339711** is simply the **PAX6-AS1** gene we first met in the **Genomic context** section, in rather thin disguise. **NR\_0339711** being the **Accession code** for the **non-coding RNA** product of the gene called **PAX6-AS1**.

There does not appear to be a wealth of information about the noble gene **PAX6-AS1**? Its Title "**PAX6 antisense RNA 1**" would seem to be the all there is to say? Well, I suppose that leaves plenty of good things for future investigators to research? But ... remember ... when the vital role for **PAX6-AS1** is revealed, you saw it first **HERE**!

Gene: PAX6-AS1  
Title: PAX6 antisense RNA 1  
Location: 31,816,566..31,887,041  
Length: 70,476 nt  
ncRNA: NR\_033971.1  
ncRNA\_class: lncRNA  
Title: PAX6 antisense RNA 1  
Location: 31,816,566..31,887,041  
[Length]  
Span on NC\_000011.10: 70,476 nt  
Aligned length: 1,656 nt  
Sequence length: 1,656 nt  
Download: [NR\\_033971.1](#)  
Links & Tools  
View GenID: [440034 \(PAX6-AS1\)](#)  
View HGNC: [53448](#)  
View GenID: [440034 \(PAX6-AS1\)](#)  
BLAST Genome-specific: [NC\\_000011.10 \(31,816,566..31,887,041\)](#)  
BLAST Genome-specific: [NC\\_000011.10 \(31,816,566..31,887,041\)](#), [NR\\_033971.1](#)  
FASTA View: [NC\\_000011.10 \(31,816,566..31,887,041\)](#), [NR\\_033971.1](#)  
GenBank View: [NC\\_000011.10 \(31,816,566..31,887,041\)](#), [NR\\_033971.1](#)  
BLAST mRNA: [NR\\_033971.1](#)  
FASTA View: [NC\\_000011.10 \(31,816,566..31,887,041\)](#), [NR\\_033971.1](#)  
GenBank View: [NC\\_000011.10 \(31,816,566..31,887,041\)](#), [NR\\_033971.1](#)  
Graphical View: [NR\\_033971.1](#)

In passing, there is no sign of the other **PAX6 antisense RNA**, **PAUPAR**, mentioned above? This can only be because region of the **PAUPAR** gene (as computed by the NCBI) does not overlap that of **PAX6**. This is not entirely clear from the more approximate representation of the **Genomic context** section.

Our first objective, to determine the number of **Transcripts** the **NCBI** suggests **PAX6** might have, remains unrequited!

We seek a number that varies wildly according to the definition of “**Transcript**” used by the **NCBI**, the quality of evidence required by the **NCBI** before they accept a **Transcript** exists and the volume of experimental evidence which increases as more research is completed (amongst other things!). Only a year ago, the evidence suggested just **11 PAX6 Transcripts**, now it is clear, at a glance, that there are many many more!

OK, so you could count the number of **Transcript** lines from the graphic? But I am far too nice a person to suggest you do that! Happily, the answer is readily available elsewhere.

Move to the **NCBI Reference Sequences (RefSeq)** section. Here you will find a numbered list of all the *mRNA and Protein(s)*.

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. [Explain](#)

Genomic

1. **NG\_008679.1 RefSeqGene**

Range

5001..38170

Download

[GenBank](#), [FASTA](#), [Sequence Viewer \(Graphics\)](#), [LRG\\_720](#)

mRNA and Protein(s)

1. **NM\_000280.4 → NP\_000271.1 paired box protein Pax-6 isoform a**

[See identical proteins and their annotated locations for NP\\_000271.1](#)

Status: REVIEWED

Description

Transcript Variant: This variant (1) initiates from the B (P1) promoter and encodes Isoform a. Variants 1, 3, 6, 7, and 12-16 all encode the same isoform (a).

Source sequence(s)

[BP394576\\_DA078958\\_M93650\\_Z83307](#)

Consensus CDS

[CCDS31451.1](#)

UniProtKB/Swiss-Prot

[P26367](#)

UniProtKB/TrEMBL

[Q66S1](#)

Related

[ENSP00000495109.1](#), [ENST00000643871.1](#)

Conserved Domains (2) [summary](#)

[smart00351](#)

Location:4 → 128

PAX; Paired Box domain

[pfam00046](#)

Location:214 → 267

Homeobox; Homeobox domain

2. **NM\_001127612.2 → NP\_001121084.1 paired box protein Pax-6 isoform a**

[See identical proteins and their annotated locations for NP\\_001121084.1](#)

Status: REVIEWED

Description

Transcript Variant: This variant (3) differs in the 5' UTR compared to variant 1. It initiates from the A (P0) promoter. Variants 1, 3, 6, 7, and 12-16 all encode the same isoform (a).

Source sequence(s)

[AK314470\\_BE221553\\_RM557761\\_BM725029\\_BP394398\\_BP394576\\_BU072567\\_BX089704\\_BX114225\\_CA397536\\_DA183294\\_Z83307](#)

Consensus CDS

[CCDS31451.1](#)

UniProtKB/Swiss-Prot

[P26367](#)

50. **NM\_001368930.1 → NP\_001355859.1 paired box protein Pax-6 isoform o**

[See identical proteins and their annotated locations for NP\\_001355859.1](#)

Status: REVIEWED

Source sequence(s)

[Z83307](#)

Related

[ENSP00000492769.1](#), [ENST00000638965.1](#)

Conserved Domains (1) [summary](#)

[pfam00046](#)

Location:13 → 66

Homeobox; Homeobox domain

51. **NM\_001604.5 → NP\_001595.2 paired box protein Pax-6 isoform b**

[See identical proteins and their annotated locations for NP\\_001595.2](#)

Status: REVIEWED

Description

Transcript Variant: This variant (2) uses an alternate splice site in the 5' UTR, and includes an alternate in-frame exon in the 5' coding region, compared to variant 1. It initiates from the B (P1) promoter. The encoded isoform (b, also known as 5a) is longer than isoform a. Variants 2, 4, 5, 8, and 17-19 all encode the same isoform (b).

Source sequence(s)

[BP394576\\_BX640762\\_CV569250\\_DA078958\\_DA141443\\_Z83307](#)

Consensus CDS

[CCDS31452.1](#)

UniProtKB/Swiss-Prot

[P26367](#)

UniProtKB/TrEMBL

[F1T0F8](#)

Related

[ENSP00000404100.1](#), [ENST00000419022.6](#)

Conserved Domains (2) [summary](#)

[smart00351](#)

Location:4 → 142

PAX; Paired Box domain

[pfam00046](#)

Location:228 → 281

Homeobox; Homeobox domain

RNA

1. **NR\_160916.1 RNA Sequence**

Status: REVIEWED

Source sequence(s)

[Z83307\\_Z95332](#)

2. **NR\_160917.1 RNA Sequence**

Status: REVIEWED

Source sequence(s)

[Z83307\\_Z95332](#)

Basic Bioinformatics

7 of 14

05:26:48



There are a few other observations to make before leaving the **NCBI Reference Sequences (RefSeq)** section. Primarily, details that will be expanded upon later, as these exercises progress.

Consensus CDS	<a href="#">CCDS31451.1</a>
UniProtKB/Swiss-Prot	<a href="#">P26367</a>
UniProtKB/TrEMBL	<a href="#">Q66SS1</a>

**Note first**, that many of the **coding Transcripts** are associated with **2 UniProtKB Proteins**. One from the **SwissProt** section of **UniProtKB** and one from the **TrEMBL** section. At first glance, this duplication might seem illogical. Of course, how this can happen sensibly will be clear to you, as you have all listened attentively to my pre-exercise videos!!? If not, explanation will be found in brief [here](#), and reinforced in the exercises that follow.

1. <a href="#">NM_000280.4</a> → <a href="#">NP_000271.1</a> paired box protein Pax-6 isoform a
2. <a href="#">NM_001127612.2</a> → <a href="#">NP_001121084.1</a> paired box protein Pax-6 isoform a
3. <a href="#">NM_001258462.2</a> → <a href="#">NP_001245391.1</a> paired box protein Pax-6 isoform b
25. <a href="#">NM_001368905.1</a> → <a href="#">NP_001355834.1</a> paired box protein Pax-6 isoform d

Look next at the **Title lines**, (first numbered lines) of a few of the **coding Transcripts**. It is clear that **Isoform** names are of the form **isoform x**, where **x** is a letter (starting with 'a' and progressing on towards 'z' as far as is required). Clearly, you could count how many **isoforms** there are ... but *please do not!* There are far better ways to do this.

34. <a href="#">NM_001368914.1</a> → <a href="#">NP_001355843.1</a> paired box protein Pax-6 isoform g
Status: REVIEWED
Description
Transcript Variant: This variant (35), as well as variants 33 and 34, encodes isoform g.

Look now at the **Description** field of any **coding Transcript**. See that a list of all the **Transcript Variants** coding for the same **isoform** are listed. Clearly, this would enable you to determine the number of

**Transcripts** that code for each **isoform** (e.g. from the illustrated example it can be seen that **3 Transcripts** code for **isoform g**)... but *please do not!* There are far better ways to do this.

Description
Transcript Variant: This variant (2) uses an alternate splice site in the 5' UTR, and includes an alternate in-frame exon in the 5' coding region, compared to variant 1. It initiates from the B (P1) promoter. The encoded isoform (b, also known as 5a) is longer than isoform a. Variants 2, 4, 5, 8, and 17-19 all encode the same isoform (b).

Look now at the **Description** field of any **isoform b Transcript** (e.g. Number 51). Note that "**isoform 5a**" is an alternative name for **isoform b** (the relevance of this will become apparent later). Note also, that **isoform b** (aka **isoform 5a**) is reported to be longer than **isoform a**.

Click on the link to the **Protein** for any **isoform a Transcript** (e.g. [NP\\_000271.1](#)).

1. <a href="#">NM_000280.4</a> → <a href="#">NP_000271.1</a> paired box protein Pax-6 isoform a
<a href="#">See identical proteins and their annotated locations for NP_000271.1</a>

paired box protein Pax-6 isoform a [Homo sapiens]
NCBI Reference Sequence: <a href="#">NP_000271.1</a>
<a href="#">Identical Proteins</a> <a href="#">FASTA</a> <a href="#">Graphics</a>
Go to: ☐
LOCUS <a href="#">NP_000271</a> 422 aa Linear PRI 05-APR-2020
DEFINITION <a href="#">paired box protein Pax-6 isoform a [Homo sapiens]</a> .

Note the the length of the protein is **422 Amino Acids**.

Click back to the **NCBI Reference Sequences (RefSeq)** display.

Click on the link to the **Protein** for any **isoform b Transcript** (e.g. [NP\\_001245391.1](#)).

3. <a href="#">NM_001258462.2</a> → <a href="#">NP_001245391.1</a> paired box protein Pax-6 isoform b
<a href="#">See identical proteins and their annotated locations for NP_001245391.1</a>

paired box protein Pax-6 isoform b [Homo sapiens]
NCBI Reference Sequence: <a href="#">NP_001245391.1</a>
<a href="#">Identical Proteins</a> <a href="#">FASTA</a> <a href="#">Graphics</a>
Go to: ☐
LOCUS <a href="#">NP_001245391</a> 436 aa Linear PRI 05-APR-2020
DEFINITION <a href="#">paired box protein Pax-6 isoform b [Homo sapiens]</a> .

Note the the length of the protein is **436 Amino Acids**.

Click back to the **NCBI Reference Sequences (RefSeq)** display.

Finally, compare the **Conserved Regions** of any **isoform a Transcript** with those of any **isoform b Transcript**.

Conserved Domains (2) <a href="#">summary</a>
<a href="#">smart00351</a> Location: 4 → 128
PAX; Paired Box domain
<a href="#">pfam00046</a> Location: 214 → 266
Homeobox; Homeobox domain

Both **Isoform a** and **Isoform b** are recorded as having two domains. A **Paired Box Domain** at the beginning, and a **Homeobox Domain** further along.

Conserved Domains (2) <a href="#">summary</a>
<a href="#">smart00351</a> Location: 4 → 142
PAX; Paired Box domain
<a href="#">pfam00046</a> Location: 228 → 280
Homeobox; Homeobox domain

Both **Paired Box Domains** are detected by a matches in the **SMART** database. Both **Homeobox Domains** are detected by matches in the **Pfam** database. Other **Domain Databases** will certainly provide supporting

evidence, but reference to just one match is sufficient here.

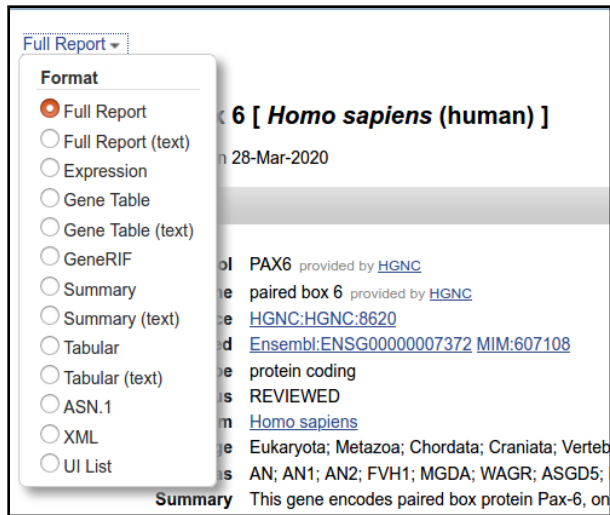
From the location information, the **Paired Box** of **Isoform b** appears to include an extra **14** amino acids.



All the information to discover, clumsily, the number of distinct **isoforms**, and the number of **Transcripts** generating each of those **isoforms**, has been located. Job done? Well yes, but before retiring in triumph, let us look at a couple of alternative sources for this information involving simple textual sources.

These straight text information sources can more easily be interrogated programmatically from the command line to discover the counts we seek. Some indication of how this might be achieved will be offered in a session at the end of the full week version of this training course.

Arguably the most obvious information sources is the **Textual Gene Table** for **PAX6**.



Move to the top of the **PAX6 Gene** page and click on menu link currently set to **Full Report**.

Select the option **Gene Table (Text)**.

Details of all the **PAX6 Transcripts** are displayed in tabular form. First the two **non-coding RNAs**.

PAX6 paired box 6[Homo sapiens]				
Gene ID: 5080, updated on 28-Mar-2020				
Reference GRCh38.p13 Primary Assembly NC_000011.10 (minus strand) from: 31817961 to: 31784792				
RNA transcript variant 53 NR_160917.1, 12 exons, total annotated spliced exon length: 6797				
Exon table for RNA NR_160917.1				
Genomic Interval Exon	Gene Interval Exon	Exon Length	Intron Length	
31817961-31817809	1-153	153	6793	
31811015-31810828	6947-7134	188	3902	
31806925-31806849	11037-11113	77	386	
31806462-31806402	11500-11560	61	3567	
31802834-31802704	15128-15258	131	927	
31801776-31801561	16186-16401	216	704	
31800856-31800691	17106-17271	166	5902	
31794788-31794630	23174-23332	159	827	
31793802-31793652	24160-24310	151	98	
31793553-31793438	24409-24524	116	2577	
31790860-31790710	27102-27252	151	690	
31790019-31784792	27943-33170	5228		
RNA transcript variant 52 NR_160916.1, 11 exons, total annotated spliced exon length: 6641				
Exon table for RNA NR_160916.1				
Genomic Interval Exon	Gene Interval Exon	Exon Length	Intron Length	
31811121-31810828	6841-7134	294	3902	
31806925-31806849	11037-11113	77	386	
31806462-31806402	11500-11560	61	3567	
31802834-31802704	15128-15258	131	791	
31801912-31801871	16050-16091	42	94	
31801776-31801561	16186-16401	216	704	
31800856-31800691	17106-17271	166	5902	
31794788-31794630	23174-23332	159	827	
31793802-31793652	24160-24310	151	98	
31793553-31793438	24409-24524	116	3418	
31790019-31784792	27943-33170	5228		

Followed by information for each of the **51 coding Transcripts**.

mRNA transcript variant 24 NM_001368903.1, 10 exons, total annotated spliced exon length: 7282 protein isoform d NP_001355832.1, 7 coding exons, annotated AA length: 286							
Exon table for mRNA NM_001368903.1 and protein NP_001355832.1							
Genomic Interval Exon	Genomic Interval Coding	Gene Interval Exon	Gene Interval Coding	Exon Length	Coding Length	Intron Length	
31803673-31802704	14289-15258	970	791				
31801912-31801871	16050-16091	42	94				
31801776-31801561	16186-16401	216	704				
31800856-31800691	31800805-31800691	17106-17271	166	115	5902		
31794788-31794630	31794788-31794630	23174-23332	159	159	515		
31794114-31794032	31794114-31794032	23848-23930	83	83	229		
31793802-31793652	31793802-31793652	24160-24310	151	151	98		
31793553-31793438	31793553-31793438	24409-24524	116	116	2577		
31790860-31790710	31790860-31790710	27102-27252	151	151	690		
31790019-31784792	31790019-31789934	27943-33170	5228	86			
mRNA transcript variant 51 NM_001368930.1, 7 exons, total annotated spliced exon length: 6011 protein isoform o NP_001355859.1, 6 coding exons, annotated AA length: 221							
Exon table for mRNA NM_001368930.1 and protein NP_001355859.1							
Genomic Interval Exon	Genomic Interval Coding	Gene Interval Exon	Gene Interval Coding	Exon Length	Coding Length	Intron Length	
31800661-31800539	17301-17423	123	5750				
31794788-31794630	31794788-31794630	23174-23332	159	79	515		
31794114-31794032	31794114-31794032	23848-23930	83	83	229		
31793802-31793652	31793802-31793652	24160-24310	151	151	98		
31793553-31793438	31793553-31793438	24409-24524	116	116	2577		
31790860-31790710	31790860-31790710	27102-27252	151	151	690		
31790019-31784792	31790019-31789934	27943-33170	5228	86			
mRNA transcript variant 9 NM_001310159.1, 9 exons, total annotated spliced exon length: 1393 protein isoform c NP_001297088.1 (CCDS86190.1), 8 coding exons, annotated AA length: 401							
Exon table for mRNA NM_001310159.1 and protein NP_001297088.1							
Genomic Interval Exon	Genomic Interval Coding	Gene Interval Exon	Gene Interval Coding	Exon Length	Coding Length	Intron Length	
31806925-31806849	11037-11113	77	386				
31806462-31806402	31806411-31806402	11500-11560	61	10	3567		
31802834-31802704	31802834-31802704	15128-15258	131	131	927		
31801776-31801561	31801776-31801561	16186-16401	216	216	704		
31800856-31800691	31800856-31800691	17106-17271	166	166	5902		
31794788-31794630	31794788-31794630	23174-23332	159	159	515		
31794114-31794032	31794114-31794032	23848-23930	83	83	229		
31793802-31793652	31793802-31793652	24160-24310	151	151	98		
31793553-31793205	31793553-31793264	24409-24757	349	290			

Notice that for every **coding Transcript** there is a line specifying the **isoform** that corresponds to the **Transcript**. This time, the **isoforms** only have different names if they represent different protein products.

mRNA transcript variant 24 NM\_001368903.1, 10 exons, total annotated spliced exon length: 7282  
protein isoform d NP\_001355832.1, 7 coding exons, annotated AA length: 286

**Isoform** names can be swiftly be confirmed to be of the form **isoform x**, where **x** is a letter (starting with ‘a’ and progressing on towards ‘z’ as far as is required) determining the particular **isoform**.

Name	#
isoform a:	9
isoform b:	7
isoform c:	1
isoform d:	13
isoform e:	1
isoform f:	1
isoform g:	3
isoform h:	3
isoform i:	2
isoform j:	1
isoform k:	1
isoform l:	6
isoform m:	1
isoform n:	1
isoform o:	1

So ... all you have to do is to trawl through the tables and see how much of the alphabet had to be used! Easy! But ... **PLEASE DO NOT DO THIS!!!** I will tell you, there are **15 isoforms**. They are called **isoform a**, **isoform b** ... **isoform o**.

Finally, there remains query number three, which is to determine how many **Transcripts** generate each of the **15 isoforms**? Again, easy! The answer lurks in the tables, you need only to read through for an hour or two and then you have the answer (and a headache). Once more ... **PLEASE DO NOT DO THIS!!!** I give you the answer.

Counting the number of **Transcripts** for each **isoform** requires one to count how many times an **isoform** name occurs. Much more ugly than just reading the **Description** fields of the **NCBI Reference Sequences (RefSeq)** display, **BUT** much easier to code, as you will see.

Alternatively, you might move back to the **Genomic regions, transcripts, and products** section and click on the **GenBank** link just above the graphic.

Here you see the portion of the **RefSeq** entry for the entirety of **Chromosome 11** that covers the **PAX6** gene region. As you can see, the **Chromosome 11 RefSeq** entry is **NC\_000011**. ‘C’ for **Chromosome**, of course. As previously, the number after the ‘.’ is a version number.

Notice there is no permanent **RefSeq** entry for the genomic region for each **Gene**. **Such** are dynamically generated as required from the single entry for the **Chromosome**.

One purpose for looking at this entry is to ensure everyone has delighted in viewing at least one example of a **GenBank Format** sequence. This format was originally defined for use with the **GenBank** database. I suggest the format really explains itself, but if you disagree, try the **Sample GenBank Record**, which provides links to clear explanation of all the possible features.

Also, the idea was to demonstrate that you could compute the answers to the questions posed by the exercise from the contents of this **RefSeq** entry as well as from the **Gene Table (Text)**.

Try searching for all lines that contain “mRNA” followed by 4 spaces (type **CtrlF** and a search box will appear at the bottom of the page).

You should find **54** hits, suggesting the presence of **54** transcripts that generate **mRNAs** perhaps?

Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly	
NCBI Reference Sequence: NC_000011.10	
<a href="#">FASTA</a> <a href="#">Graphics</a>	
LOCUS	NC_000011 33170 bp DNA linear CON 02-MAR-2020
DEFINITION	Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly.
ACCESSION	<a href="#">NC_000011</a> REGION: complement(31784792..31817961)
VERSION	NC_000011.10
DBLINK	BioProject: <a href="#">PRJNA168</a> Assembly: <a href="#">GCF_000001405.39</a>
KEYWORDS	RefSeq.
SOURCE	Homo sapiens (human)
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo. 1 (bases 1 to 33170)
REFERENCE	
AUTHORS	Taylor,T.D., Noguchi,H., Totoki,Y., Toyoda,A., Kuroki,Y., Dewar,K., Lloyd,C., Itoh,T., Takeda,T., Kim,D.W., She,X., Barlow,K.F., Bloom,T., Bruford,E., Chang,J.L., Cuomo,C.A., Eichler,E., Fitzgerald,M.G., Jaffe,D.B., LaButti,K., Nicol,R., Park,H.S., Seaman,C., Sougnez,C., Yang,X., Zimmer,A.R., Zody,M.C., Birren,B.W., Nusbaum,C., Fujiyama,A., Hattori,M., Rogers,J., Lander,E.S. and Sakaki,Y.
TITLE	Human chromosome 11 DNA sequence and analysis including novel gene identification
JOURNAL	Nature 448 (7083), 497-500 (2006)
PUBMED	<a href="#">16554811</a>
REFERENCE	2 (bases 1 to 33170)
CONSTRM	International Human Genome Sequencing Consortium
TITLE	Finishing the euchromatic sequence of the human genome
JOURNAL	Nature 431 (7011), 931-945 (2004)
PUBMED	<a href="#">15496913</a>
REFERENCE	3 (bases 1 to 33170)
AUTHORS	Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R.,

```

mRNA
join(1..153,6947..7134,11037..11113,11500..11560,
15128..15258,16050..16091,16186..16200,17106..17271,23174..23332,
23848..23930,24160..24310,24409..24524,27102..27252,
27943..33170)
/gene="PAX6"
/feature_synonym="AN; AN1; AN2; ASG05; D115812E; FVH1; MGDA;
WAGR"
/product="paired box 6, transcript variant 3"
/note="Derived by automated computational analysis using
gene prediction method: BestRefSeq."
/transcript_id="NM_00107612.2"
/db_xref="GeneID:5080"
/db_xref="HGNC:8620"
/db_xref="MIM:607108"
mRNA
join(1..153,6947..7134,11037..11113,11500..11560,
15128..15258,16050..16091,16186..16200,17106..17271,
23174..23332,23848..23930,24160..24310,24409..24524,
27943..33170)
/gene="PAX6"
/feature_synonym="AN; AN1; AN2; ASG05; D115812E; FVH1; MGDA;
WAGR"
/product="paired box 6, transcript variant 42"

```

```

complement(27638...>33170)
/ gene="ELP4"
/ gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
PAXN6B; PAXN6B"
/product="elongator acetyltransferase complex subunit 4,
transcript variant 2"
/ note="Derived by automated computational analysis using
gene prediction method: BestRefSeq."
/transcript_id="NM_001288725.2"
/db_xref="GeneID:26618"
/db_xref="HGNC:HGNC:1171"
/db_xref="MIM:606985"
complement(27638...>33170)
/ gene="ELP4"
/ gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
PAXN6B; PAXN6B"
/product="elongator acetyltransferase complex subunit 4,
transcript variant 3"
/ note="Derived by automated computational analysis using
gene prediction method: BestRefSeq."
/transcript_id="NM_001288726.2"
/db_xref="GeneID:26618"
/db_xref="HGNC:HGNC:1171"
/db_xref="MIM:606985"
complement(27638...>33170)
/ gene="ELP4"
/ gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
PAXN6B; PAXN6B"
/product="elongator acetyltransferase complex subunit 4,
transcript variant 1"
/ note="Derived by automated computational analysis using
gene prediction method: BestRefSeq."
/transcript_id="NM_019040.5"
/db_xref="Ensembl:ENST00000640961.2"
/db_xref="GeneID:26618"
/db_xref="HGNC:HGNC:1171"
/db_xref="MIM:606985"

```

```

CDS
join(17157. -17271, 23174. -23332, 23848. -23930, 24160. -24310,
24489. -24524, 27943. -28332)
/genes="PAX6"
/genes_synonym="AN; AN1; AN2; ASG05; D1S81B2; FVH1; MGDA;
WAGR"
/notes="isoform n is encoded by transcript variant 50;
Derived by automated computational analysis using gene
prediction method: BestRefSeq."
/codon_start=1
/product="paired box protein Pax-6 isoform n"
/protein_id="NP_001352588.1"
/db_xref="GeneID:5088"
/db_xref="HGNC:HGNC:8620"
/db_xref="MIM:607188"
/translation="MGADGMYDKLRMLNGQTGSWTRPGWYPTSGVPTGQDCQQO
EGGGNTNISISSNGDEEAQMRLQLKRLQRHTSFSTQEIIEALEKFEFTHYPDVF
ARELAKALDTPEARLQVFNWNRRAKREELKNRGRAGNTSHPIPTSSSSFTSYV
QPIPRPTTPVSSFTSGGN.GRTTAL.TATYSLPPMPS.FTRAMLPGQSFPLVCQF
KPFPEVNLCLNTGDTGPKSCKKKKKKKERKYCNVNSVDYDTFTVLSSGKKHMLLEPL
O
FYKVLVYCTTGGGDLKQGLPYTGTISVGNLHFSGITFTFHFLVNHLYVYMKK
RTM"
CDS
join(17157. -17271, 23174. -23332, 23848. -23930, 24160. -24310,
24489. -24524, 27192. -27252, 27943. -28028)
/genes="PAX6"
/genes_synonym="AN; AN1; AN2; ASG05; D1S81B2; FVH1; MGDA;
WAGR"
/notes="isoform d is encoded by transcript variant 22;
Derived by automated computational analysis using gene
prediction method: BestRefSeq."
/codon_start=1
/product="paired box protein Pax-6 isoform d"
/protein_id="NP_001352583.1"
/db_xref="GeneID:5088"
/db_xref="HGNC:HGNC:8620"

```

You might have expected **51**, given previous investigations *BUT* ... remember that this “**PAX6**” region also includes **3 ELP4 coding Transcripts**. So, with a bit of thought, mission accomplished as far as counting the **Transcripts** is concerned? **51 PAX6** transcripts plus **3 ELP4** transcripts equals **54** transcripts of unspecified origin, after all.

Now try searching for “**PAX-6 isoform**”. Lo and behold! **51** hits and the naming scheme for the **isoforms** as expected? I suggest we are there!

Of course, we discuss extremely sloppy strategies to answer questions of rather dubious worth here, but it is the principles, the possibilities that are of interest in this context.

Once again, *please do not try to work out anything from you displays*. The answers offered a page back still apply.

Just one question remains. How did I determine the answers I strenuously requested you not to waste time working out? Well ... I most certainly did not spend ages reading through web displays, text tables or the **GenBank** format sequence!! I spent just long enough to see **HOW** the queries could be answered, and then I downloaded both the plain text data representations to my computer and wrote simple programs to extract the information I wanted.

Pretty clever eh? ... Well, not really. I do not generally do clever things. With some small instruction, copying data from sites such as the **NCBI** and composing small programs (scripts) to analyse that data is trivial. We hope to convince you that this is true in the final stage of this course of instruction.

Hopefully, you will see the importance of acquiring minimal programming skills. The general truth being that, if you wish merely to superficially **browse** the data/information offered by sites such as the **NCBI**, then use a **browser**. However, if you wish to meaningfully **interrogate** that data/information, you will almost inevitably need to use more powerful, if less beautiful, tools.

You may, with some justice say, *“But when would we ever want to ask the questions suggested in these exercises?”*. Maybe never, but the fact remains. Whatever questions you **do** want to ask, a browsing approach alone will rarely suffice, particularly if you wish to examine large sets of data.

Time for a break folks? Next we will look at, basically the same story, as told by the **Ensembl** database.

**DPJ – 2020.04.28**



**Supplementary notes and discussion arising from the Instruction Text.**

The intention is to provide extra instruction and discussion not essential to the purpose of the exercise. The “Appendices” are for “interest only”. They can all be skipped if you are short of time.

a full consideration of some issues skimmed over in the exercise proper.

If you are attending a “supervised” presentation of the exercise, I would hope to have conducted a live discussion of all these issues to an extent that reflects:

Some of the “**Discussion Points**” are rather long and rambling. You have been warned.

Can a single coding **Transcript** correspond to more than one **UniProtKB Protein**?

Any given **Transcript** will have just one CoDing Sequence (CDS) and so can only code for a single **Protein**.

As touched upon in one of the introductory videos for this exercise, **UniProtKB** strives to be **non-redundant**, that is, no **Protein** should be represented more than once.

Therefore, all coding **Transcripts** should correspond to *one and only one* **Uniprot** entry.

However, as the video explained, **UniprotKB** has two sections:

**Swiss-Prot** Comprised of **Proteins** that have been fully “*Manually annotated*” with “*information extracted from literature and curator-evaluated computational analysis*”.

**TrEMBL** Comprised of **Proteins** determined only by “*Computational analysis*”. These “*await full manual annotation*”. After such “*full manual annotation*” a **TrEMBL** entry will be discovered to be nonsense (and deleted), a duplicate of something already in the **Swiss-Prot** section (and deleted), or a truly worthy newly discovered **Protein** deserving of instant promotion to the **Swiss-Prot** section.

So, it might be the case that a protein could exist, for a short time, *both* in **TrEMBL** *and* in **Swiss-Prot**. As soon as the **TrEMBL** version is properly examined and determined to be a duplicate of an extant **Swiss-Prot** entry, it will be eliminated. However, should a **coding Transcript** be annotated during the time of duplication, some **coding Transcripts** might well appear, **erroneously**, to match *two* proteins.

That is what has happened here in the case of several **coding Transcripts**.

Of course, none of this will be new to you as you watched my lovely video carefully? Only *very very BAD* people who skipped the video will read the above as novel information.

[Click Here to Return to the Exercise →](#)

**DPJ – 2020.04.28**