Introduction to Bioinformatics

23-27 March 2020

Practical 1: Databases and Tools

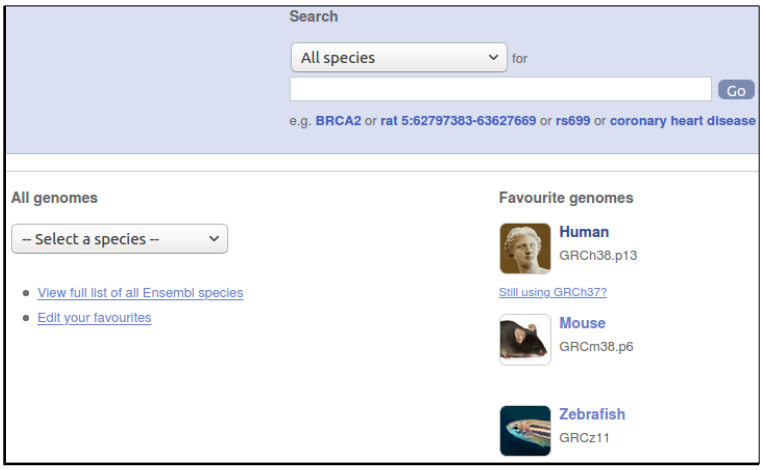
Part ii) – Aniridia viewed from Ensembl

Tuesday 28 April 2020

Investigating the gene(s) associated with Aniridia using Ensembl

As previously, the starting point for this exercise is to imagine you have a vital interest in discovering and investigating the main human gene responsible for the terrible disease of the eye, **Aniridia**. This time, however, we will investigate the topic as recorded by the **Genome Database** Ensembl.

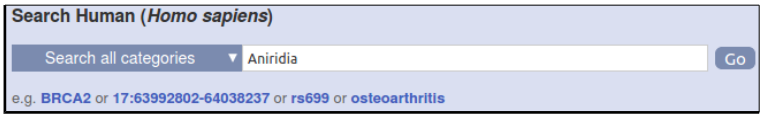
So, begin by going to the **Home Page** of Ensembl.



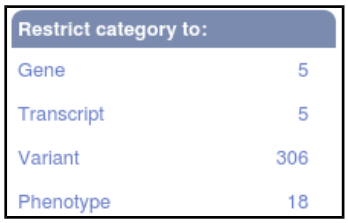
A good way to start would be to select a species, **Human** being the pertinent choice. Accordingly, click on the **Human** option in the list of **Favourite genomes**.

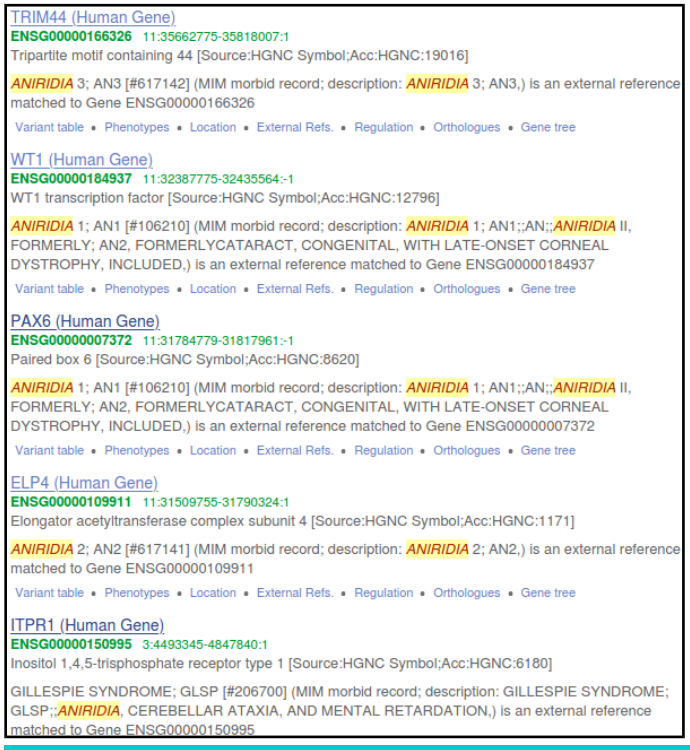
Why **Ensembl** should imagine my **Favourite genomes** are **Human**, **Mouse** and **Zebrafish** I cannot imagine? I have deep reservations concerning both **Mice** and **Humans**, and I have never been formally introduced to a **Zebrafish** and so am inclined to be of *"open mind"* for the present.

Following the well worn path of the **NCBI** investigation, enter **Aniridia** into the **Search Human (Homo Sapiens)** box and click on the **Go** button.



You will arrive at a page offering access to an assortment of data items that relate to **Aniridia**. These include a link to **5 Genes**. As it is the **Genes** that are of prime interest here, why not click on the aforementioned 5 Gene link in the **Restrict category to:** section?





A list of matching **Genes** bustles forth energetically. Admittedly in a different order, but **4** of the **5 Genes** selected were also chosen by the **NCBI**. Pretty good agreement so far. Once again, it is **PAX6** that is of most interest (although relegated from second to third position by **Ensembl**).

So, click on the PAX6 (Human Gene) link to see the **Ensembl** version of the details of **PAX6**.

First a quick glance at the top of the page. As you will assuredly recall from our meander through the wonders of the **NCBI**, the **Assembly** of the **Human Genome** analysed at the **NCBI** is **GRCh38.13**. As this is the most up to date version, it is not surprising that it is also the **Assembly** that **Ensembl** uses.

Both sites use the same, universally accepted **Assembly** of the **Human Genome**. This **Assembly** is updated regularly and the most recent version is the *"raw data starting point"* for both databases. The analysis of that shared *"starting point"* at the **NCBI** and at **Ensembl** is, however, completely independent. It is the difference in *"interpretation"* of the same **Assembly** that we investigate in the early stages of these exercises.

The conflicts begin at once! Both agree upon **Chromosome 11**, although **Ensembl** does not mention the band (becoming a legacy concept perhaps?), but the precise region of **Chromosome 11** assigned to **PAX6** by the **NCBI** is, as I feel sure you do not need me to remind you:

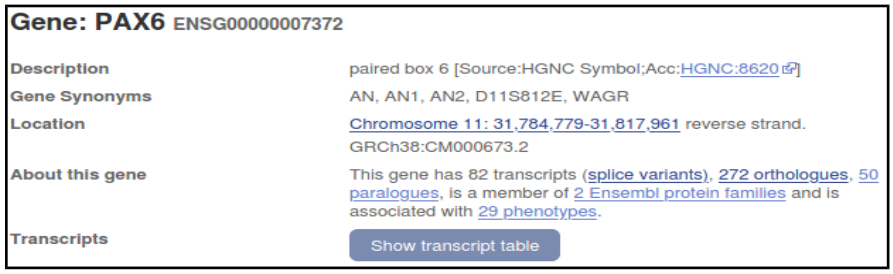
31,784,792 - 31,817,961

Whereas, **Ensembl** speculates:

31,784,779 - 31,817,961

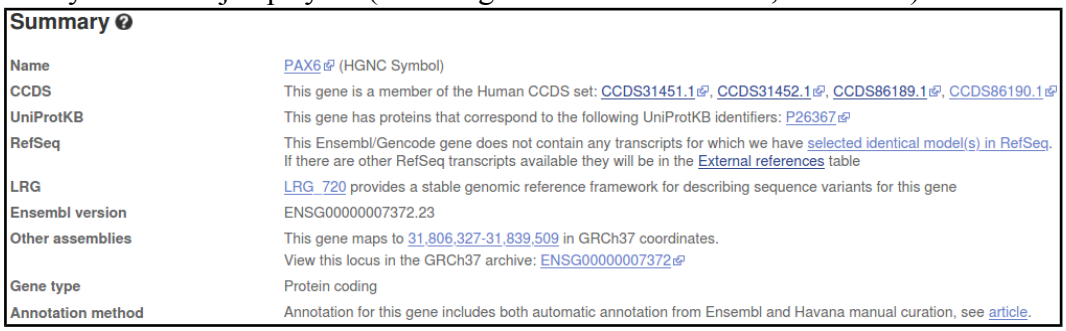
Maybe **Ensembl** sees a **PAX6** transcript that reaches a bit farther than its most adventurous **NCBI** counterpart?

At the top of the **Gene:** tab, **Ensembl** declares the internal name for this **Gene** as a generously zeroed, **ENSG00000007372** (equivalent to **5080** at the **NCBI**, as surely you recall). Why so many leading zeroes? Well … there was a time when it was thought there would be one or two more **Genes** hiding away in the **Human Genome** than turned out to be the case. Better safe than sorry, opined the **Ensembl** designers. **ENSG**? Stands for **ENS**embl **G**ene, but you worked that one out for yourselves, I am certain.



Determining the number of **Transcripts** is easy. **82**, it boldly declares in the **About the gene section** at the top of the **Gene: PAX6** tab.

The CCDS (Consensus CoDing Sequence) Database is comprised of **CDS** (**Co**ding **S**equences, currently just for **Human** and **Mouse**) that are agreed by all the major players (including **Ensembl** and **NCBI**, of course) concerned with **Genome Analysis**.



In the **Summary** section, this gene claims membership of four **CCDS families**.

A **CCDS family** being a collection of **Gene transcripts** sharing the same **CCDS** approved **CoDing Sequence (CDS)**. Put another way, **Transcripts** coding for the same **isoform**? I reason that **Ensembl** predicts **FOUR *quality*** isoforms for **PAX6**.

Feel free to click around further here if you have the inclination, but … be warned that here the story gets complex beyond the needs of this simple exercise. Simple questions leading to answers that require much disentangling. Such is the way of things.

Were you now to click on the **Show transcript table** button, you would be offered a cornucopia of further detail.

At the top of the table are all the best quality **Protein coding Transcripts** that belong to **CCDS Database Families**.

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq Match	Flags
PAX6-269	ENST00000640368.1	6975	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-282	ENST00000643871.1	6944	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	GENCODE basic APPRIS ALT1
PAX6-244	ENST00000638914.3	6922	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:1 GENCODE basic APPRIS P4
PAX6-231	ENST00000606377.6	6901	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-209	ENST00000419022.6	6888	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:1 GENCODE basic APPRIS P4
PAX6-254	ENST00000639386.2	6509	286aa	Protein coding	CCDS86189	A0A1W2PRA8	-	TSL:5 GENCODE basic
PAX6-203	ENST00000379109.7	3182	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:2 GENCODE basic APPRIS ALT1
PAX6-235	ENST00000638629.1	2844	286aa	Protein coding	CCDS86189	A0A1W2PRA8	-	TSL:2 GENCODE basic
PAX6-272	ENST00000640610.1	2730	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS ALT1
PAX6-258	ENST00000639916.1	2622	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS ALT1
PAX6-207	ENST00000379129.7	2614	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-202	ENST00000379107.7	2579	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-208	ENST00000379132.8	2576	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:5 GENCODE basic APPRIS ALT1
PAX6-281	ENST00000640975.1	2553	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:1 GENCODE basic APPRIS P4
PAX6-256	ENST00000639409.1	2450	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:1 GENCODE basic APPRIS P4
PAX6-248	ENST00000639034.2	1794	401aa	Protein coding	CCDS86190	D1KF47	-	TSL:5 GENCODE basic
PAX6-201	ENST00000241001.13	1736	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS ALT1
PAX6-217	ENST00000481563.6	1587	286aa	Protein coding	CCDS86189	A0A1W2PRA8	-	TSL:5 GENCODE basic
PAX6-257	ENST00000639548.1	1417	286aa	Protein coding	CCDS86189	A0A1W2PRA8	-	TSL:5 GENCODE basic
PAX6-263	ENST00000640125.1	1375	286aa	Protein coding	CCDS86189	A0A1W2PRA8	-	TSL:5 GENCODE basic
PAX6-252	ENST00000639109.1	2734	277aa	Protein coding	-	A0A1W2PQ31	-	CDS 3' incomplete TSL:5
PAX6-251	ENST00000639079.1	2608	263aa	Protein coding	-	A0A1W2PQJ8	-	CDS 3' incomplete TSL:5
PAX6-260	ENST00000639943.1	2584	314aa	Protein coding	-	A0A1W2PPM5	-	CDS 3' incomplete TSL:5

There follows many more "*Protein coding*" **Transcripts**, predicted to exist by **Ensembl**, but not matching any of the agreed **Co**ding **S**equences (**CDS**) in the **CCDS Database**.

The last **Transcript** recorded as coding for a protein declares itself as "*Nonsense mediated decay*"? The official answer to your question is revealed below, but I have to admit that my understanding remains short of full enlightenment.

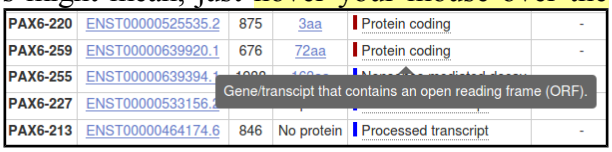
This **Transcript** codes for a protein of **163 amino acids** that is recorded in the **Uniprot Databases** with an **Accession Code** of A0A1W2PQW3. This, I opine, puts it a mite closer to rationality than the **Transcript** two entries above that codes for a protein of **3 amino acids** that has no **Uniprot** entry?

PAX6-220	ENST00000525535.2	875	3aa	Protein coding	-	-	-	CDS 3' incomplete TSL:5
PAX6-259	ENST00000639920.1	676	72aa	Protein coding	-	A0A1W2PR58	-	CDS 3' incomplete TSL:5
PAX6-255	ENST00000639394.1	1988	163aa	Nonsense mediated decay	-	A0A1W2PQW3	-	TSL:5
PAX6-227	ENST00000533156.2	848	No protein	Processed transcript	-	-	-	TSL:5
PAX6-213	ENST00000464174.6	846	No protein	Processed transcript	-	-	-	TSL:5
PAX6-222	ENST00000530373.6	785	No protein	Processed transcript	-	-	-	TSL:5
PAX6-223	ENST00000530714.6	650	No protein	Processed transcript	-	-	-	TSL:5
PAX6-266	ENST00000640251.1	649	No protein	Processed transcript	-	-	-	TSL:5
PAX6-229	ENST00000534353.5	540	No protein	Processed transcript	-	-	-	TSL:4
PAX6-253	ENST00000639203.1	532	No protein	Processed transcript	-	-	-	TSL:5
PAX6-233	ENST00000638278.1	417	No protein	Processed transcript	-	-	-	TSL:NA
PAX6-274	ENST00000640617.1	412	No protein	Processed transcript	-	-	-	TSL:4
PAX6-278	ENST00000640819.1	368	No protein	Processed transcript	-	-	-	TSL:5
PAX6-228	ENST00000533333.5	6173	No protein	Retained intron	-	-	-	TSL:2
PAX6-216	ENST00000474783.2	4392	No protein	Retained intron	-	-	-	TSL:5
PAX6-214	ENST00000470027.7	3587	No protein	Retained intron	-	-	-	TSL:2
PAX6-264	ENST00000640172.1	2525	No protein	Retained intron	-	-	-	TSL:5

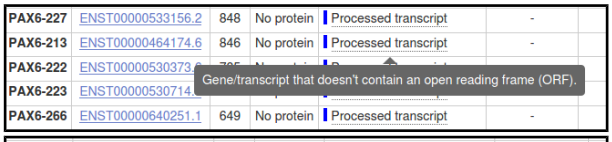
The remaining **Transcripts** are of **Biotype** "*Processed transcript*" or "*Retained intron*".

For a clearer explanation as to what the various **Biotype** possibilities might mean, just hover your mouse over the terms and a **pop up** will **pop up** and explain.

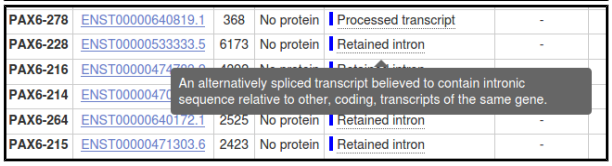
"*Protein coding*" is really self explanatory. A **Transcript** with an **Open Reading Frame** that gives rise to a protein.



"*Processed transcript*". A **Transcript** that generates an **RNA** that does not code for a protein.



"*Retained intron*". A "*Protein coding*" **Transcript** whose function has been disrupted because one or more introns have not been successfully spliced out. An **RNA** is produced, but no protein.



And what of the enigmatic "*Nonsense mediated decay*"? No **pop up**, **pops up**! and we are left to wonder. But, fear not, a suitable **pop up** is in the planning stage. In the meantime, explanations of all possible **Biotype** values are available online. Here one can discover that the elaboration of "*Nonsense mediated decay*" is:

> "*If the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site then it is tagged as NMD. If the variant does not cover the full reference coding sequence then it is annotated as NMD if NMD is unavoidable i.e. no matter what the exon structure of the missing portion is the transcript will be subject to NMD.* "

Gulp!

PAX6-276	ENST00000640735.1	2060	No protein
PAX6-277	ENST00000640766.1	1797	283aa
PAX6-278	ENST00000640819.1	368	No protein
PAX6-279	ENST00000640872.1	2367	281aa
PAX6-280	ENST00000640963.1	1267	382aa
PAX6-281	ENST00000640975.1	2553	436aa
PAX6-282	ENST00000643871.1	6944	422aa

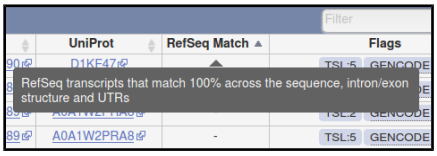
If you wished to check the **Transcript** count, you might click on the **Name** header to order the **Transcripts** by their identifying number. The numbers range from **201** to **282** confirming **82 Transcripts**. Jolly good! You might wonder why the numbering starts at **201**? "Where are the first **200 Transcripts**?" one might ponder. The answer is historical and of no real interest any more. As far as I am aware, there are no current plans to make the naming scheme less whimsical, so why not just accept what is.

Protein	Biotype	CCDS	UniProt	RefSeq Match
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
286aa	Protein coding	CCDS86189	A0A1W2PRA8	-
286aa	Protein coding	CCDS86189	A0A1W2PRA8	-
286aa	Protein coding	CCDS86189	A0A1W2PRA8	-
286aa	Protein coding	CCDS86189	A0A1W2PRA8	-
286aa	Protein coding	CCDS86189	A0A1W2PRA8	-
401aa	Protein coding	CCDS86190	D1KF47	-
277aa	Protein coding	-	A0A1W2PQ31	-

If you wished to checked the **CCDS** family associations of the quality coding **Transcripts**, you might click on the **CCDS** header to order the **Transcripts** by their **CCDS** family.

A trifle tedious, but it is now a simple matter to discover how many **Transcripts** correspond to each **CCDS** family (and so to each quality **isoform**).

Logically, each **CCDS** family should (of course?) correspond to just ***ONE*** entry in the protein sequence databases of Uniprot. However, two of the families displayed here relate to two proteins sequences each!! Why this is possible will become clearer when we discuss **Uniprot** in greater depth. If you cannot wait that long, click HERE for a brief explanation.



Hover over the **RefSeq Match** header. It will be revealed that this column exists to record **RefSeq Transcript** Sequences that match **Ensembl Transcript** predictions in every locational respect.

ALL **exons** and **introns**, including those in **U**n**T**ranslated **R**egions (**UTR**s), must be identical. A substantial objective of this exercise is to highlight the ways in which alternative examinations of the same data (a shared **Assembly** of the **Human Genome** in this case) can give rise to conflicting conclusions. Here is an example where the **NCBI** an **Ensembl** attempt to achieve a level of agreement.

Exclusively for the **Human Genome**, these two Institutes have set up The Matched Annotation from the NCBI and EBI (MANE) project, the objective of which is to identify one quality **Transcript** per **Protein Coding Locus** that is predicted identically by the differing analytical methods of both services.

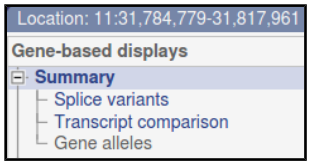
Click on the **RefSeq Match** header. This would bring all the **Transcripts** with **Refesq Matches** to the top of the column … if there were any! However, it is clear there are ***NO*** **RefSeq Transcripts** that exactly match ***ANY*** **Ensembl Transcripts** for **PAX6**! The **MANE** project is relatively new. Currently they claim to have found **Matching Transcripts** for around **70%** of **Human Protein Coding Genes**, but not yet **PAX6**.

To view less stringent associations between the **Transcript** predictions of the **NCBI** and **Ensembl**, click on the External references link to be found in the **PAX6 Summary** section, **RefSeq** subsection.

Slide down the page a wee bit and you will come to a table with many superficial similarities to the **Transcript Table** you have just been examining.

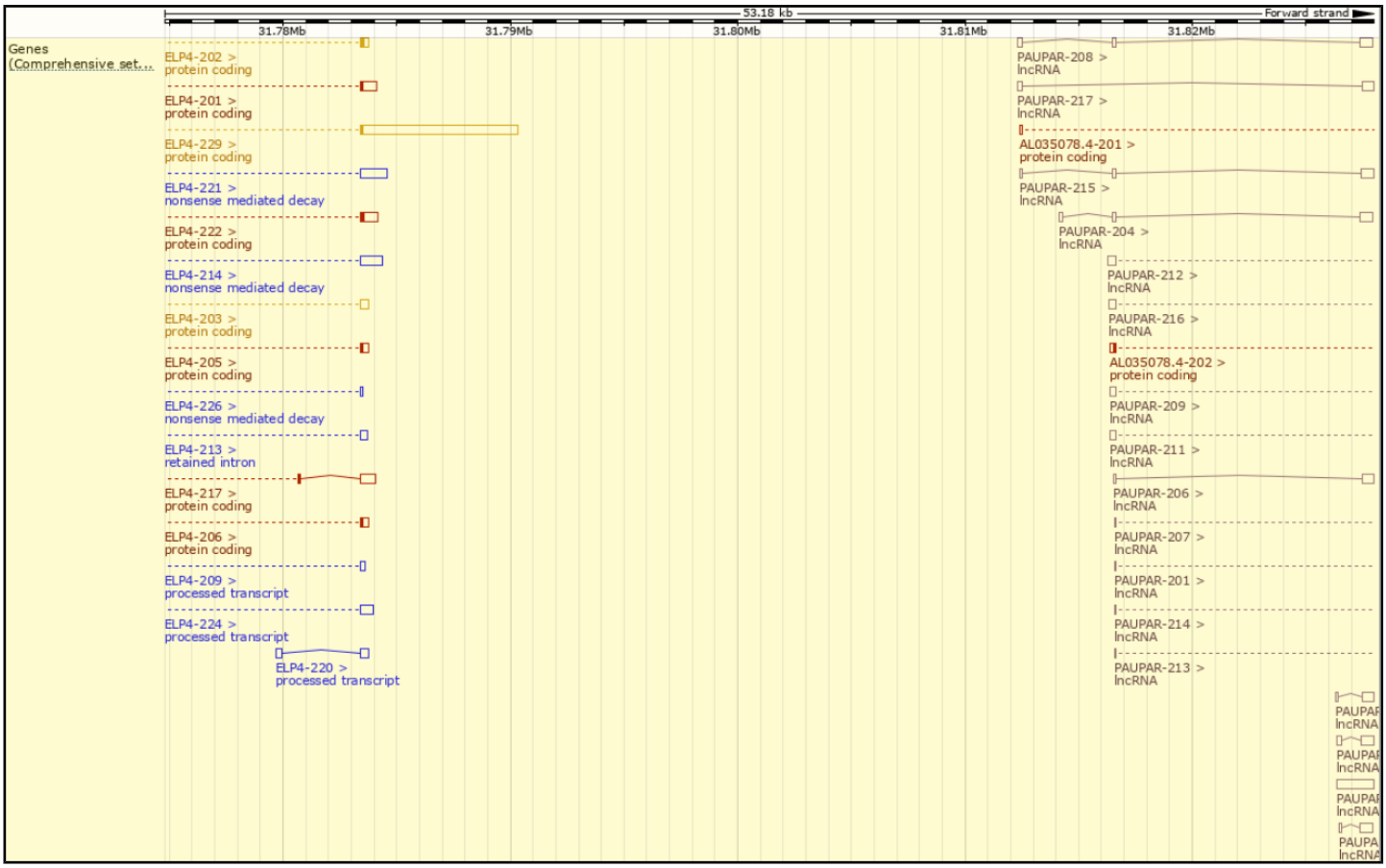
Transcript ID	Transcript name	CCDS	UniProtKB/ Swiss-Prot	RefSeq mRNA	Human Protein Atlas	PDB
ENST00000606377.6	PAX6-231	CCDS31452.1	P26367	NM_001368892.1 NM_001368918.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000640368.1	PAX6-269	CCDS31452.1	P26367	NM_001368893.1 NM_001368894.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000638914.3	PAX6-244	CCDS31452.1	P26367	NM_001258462.2 NM_001368919.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000419022.6	PAX6-209	CCDS31452.1	P26367	NM_001604.5	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000643871.1	PAX6-282	CCDS31451.1	P26367	NM_000280.4 NM_001258464.1 NM_001368887.1 NM_001368888.1 NM_001368890.1 NM_001368910.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000379132.8	PAX6-208	CCDS31451.1	P26367	NM_001368891.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000379109.7	PAX6-203	CCDS31451.1	P26367	NM_001368889.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000639916.1	PAX6-258	CCDS31451.1	P26367	NM_001258465.2 NM_001368920.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX
ENST00000639409.1	PAX6-256	CCDS31452.1	P26367	NM_001258463.1	CAB034143 CAB034143 HPA030775 HPA030775	2CUE 6PAX

This time, many good quality (**NM_**) **NCBI Transcripts** are referenced in the **RefSeq mRNA** column. In this table, **100%** matching is not required. The **RefSeq Transcripts** only have to agree enough to contribute credibility to the **Ensembl** assertion that one, or more, exons exist. The **RefSeq mRNAs** are supportive evidence for ***part(s)*** of the **Ensembl** prediction, not the same entire prediction using different analytical methods.



Next, back up a bit. Click on the **Summary** option from the **Gene-based displays** offered at the top of the **Location** tab of your current page.

Slide down a trifle until you come to a graphical representation of the **Transcripts** of all **Genes** in the **PAX6** region.



The general arrangement is very similar to that seen at the NCBI site. **Transcripts** are represented horizontally with "*blobs*" representing the **Exons** joined by lines representing the **Introns**.

"*blobs*" that are coloured in code for **Protein**. "*blobs*" that are empty do not code for **Protein**.

The top part of the display represents the forward strand of **Chromosome 11**, that does not include **PAX6**. **Ensembl** does not reverse their displays to reflect the strand of the **Gene** in focus, as does the **NCBI**.

This top strand paints a different and more complex picture than was seen at the **NCBI**. Specifically:

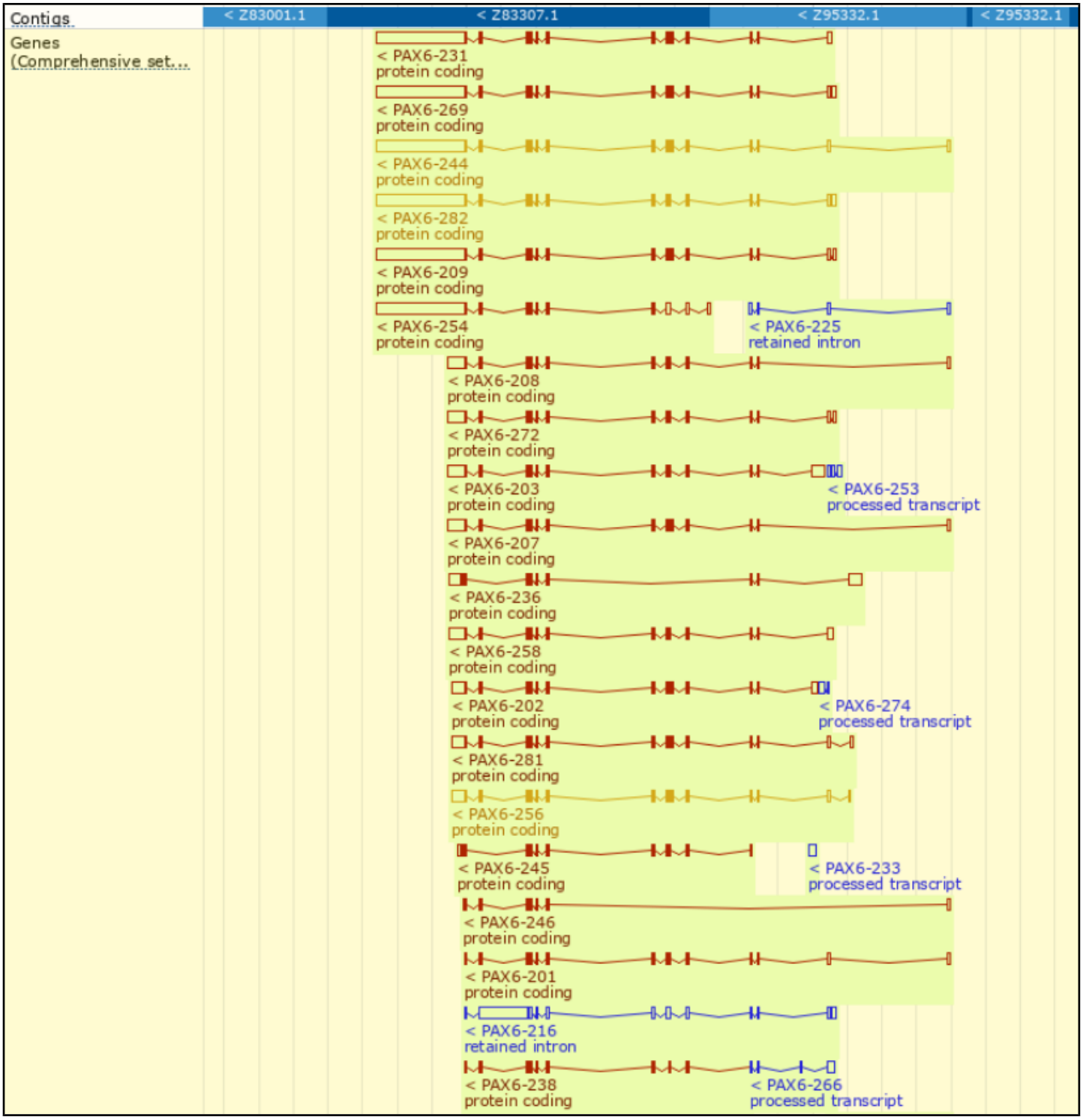
- **ELP4** has many more than the **three Transcripts** claimed by the **NCBI**.
- **PAX6-AS1** has disappeared altogether!
- **PAUPAR** is claimed to overlap the **PAX6** region by **Ensembl** and to be far longer and more complex than claimed by the **NCBI**.
- **AL035078** appears from nowhere!!! This is a "*novel protein*". It has appeared since the last **Assembly** of the **Human Genome**. I upsets my story!! I will ignore it in the hope it will ***go away***!!

Mostly for fun (**???**), I investigated the loss of **PAX6-AS1** and concluded that **Ensembl** does still predict its existence, but as a **Transcript** of **PAUPAR** and not a separate **Gene**.

I strongly suggest you just accept this, as it is not so important. However, if you wish to share the several hours of pain I endured coming to this conclusion, feel free, however, ***I strenuously advise against this diversion***, unless you are really desperate for something to do.

Moving down to the **PAX6** strand. Notice the region on display is the **PAX6** region, plus a bit either side. Note that just **6** of the **82 Transcripts** appear to have the very big terminating **Exon** that the **NCBI** predicts for all its **Transcripts**.

Note the line (**Track** is the official term), labelled **Contigs**, that lies on top of all the **PAX6 Transcript** lines. These **Contigs** are the sequences of random sections of the **Human Genome**, separately sequenced and then assembled to determined the sequence of entirety of **Chromosome 11**. The **Contigs** are coloured alternately **light blue** and **dark blue** in the order of their **Assembly** to form the **Chromosome**. Later, there will be further discussion of these "**Contigs**" and their role in the sequencing large **Genomes** around the era of the **Human Genome Sequencing Project** (**HGMP**). For now, just note that the **PAX6** region seems to involve the overlap between just two **Contigs**.



For all organisms, **Ensembl** employs an entirely automated strategy for predicting **Protein Coding Transcripts**. The Automatic Gene Annotation Pipeline, involves running a series of computer programs (a **pipeline**), completely unsupervised, to detect and annotate interesting features.

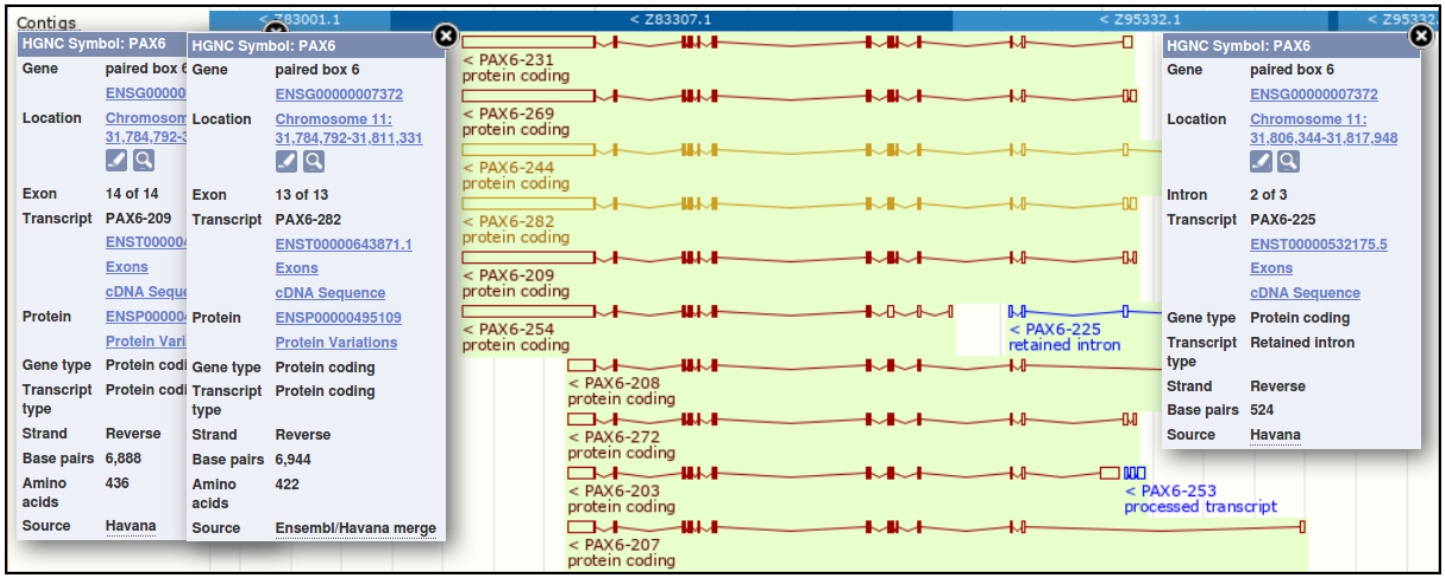
For a restricted number of major organisms (currently **Human**, **Mouse**, **Zebrafish** and **Rat**), a second, manual strategy is available. Manual Annotation by the Havana Team, involves running further program pipelines in conjunction with **Human Examination** of a wealth of other information including data from publications. Only the manual methods can predict **non-coding Transcripts**.

Where manual annotation is available (as it is here), the predictions of the two strategies are merged, and displayed as above.

Where **Protein Coding Transcripts** are predicted by one or the other predictive startegy, ***but not both***, the **Transcript** is coloured **Red**. When both strategies agree on all the important features of a **Protein Coding Transcript**, it is coloured **Gold**.

All the **Blue** **Transcripts** are some variety of **non-coding Transcript**. All **Blue** **Transcripts** must be predicted by **Manual Annotation** methods.

Should it become, for whatever bizarre reason, interesting to discover which annotation strategy predicted a specific **Red** **Transcript**, one might click on the **Transcript** representation. A **pop up** window of enlightenment will sally forth. Right at the bottom there will be a **Source** line which will claim **Ensembl** for a purely **Automatic** prediction or **Havana** for a purely **Manual** prediction. All the **Gold** **Transcripts** will declare **Ensembl/Havana merge**. All the **Blue** **Transcripts** will, of course, be **Havana**.



In this case, I could not find any **Red** **Transcripts** that declared their source as "**Ensembl**". Just **2** **Transcripts** were **Gold** and so "**Ensembl/Havana merge**". So it looks like the **Automated Annotation Pipeline** predicted just **2** **Protein coding Transcripts**, both of which were confirmed by **Havana Manual Annotation** strategies.

The database searches used as the fundamental strategy to identify **Transcripts** take a very long time to execute, even given the immense computing resources available to the **NCBI** and the **Ensembl** teams. Some clever strategies are employed to minimise the time spent on these searches. **Only for those who find such things "*interesting?*"**, I include an **Appendix** to discuss the some of the ways **Ensembl** design their pipelines for the **Human Genome** for maximal efficiency (and a few related issues). Feel free to skip this **Appendix** if you have better things to do.

Finally, for later analysis we need to save the sequence of the **PAX6** genomic region. This is tedious, not instructive and wastes time, so I have done it for you. You will find the required sequence in a file called **pax6_genomic.fasta** in a directory called **Backup_Files**. Should you really want to do this for yourself, I include the instructions.

Time for a break? I think so ….

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 **CCDS Family** correspond to more than one **Protein** from **UniProtKB**?

A very similar statement to that included in the previous exercise in refence to individual **coding Transcripts** corresponding to more that one **UniProtKB Protein**.

All **Transcripts** belonging to a given **CCDS Family**, by definition, share exactly the same **C**o**D**ing **S**equence (**CDS**).

Therefore, all **Transcripts** belonging to a given **CCDS Family** must code for the same **Protein isoform**.

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 **Transcripts** belonging to a given **CCDS Family** should correspond to ***one and only one*** **Uniprot** entry (all **CCDS Protein isoforms** will certainly be included in **UniProtKB**).

As touched upon in one of the introductory videos for this exercise, this is true, but only if one considers solely the fully annotated section of the **Uniprot Protein Sequence Database** (**Swissprot**).

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 the **CCDS Families** be annotated during the time of duplication, some CCDS Families might well appear, **erroneously**, to match ***two*** proteins.

That is what has happened here in the case of **two** of the **four PAX6 CCDS Families**.

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 —

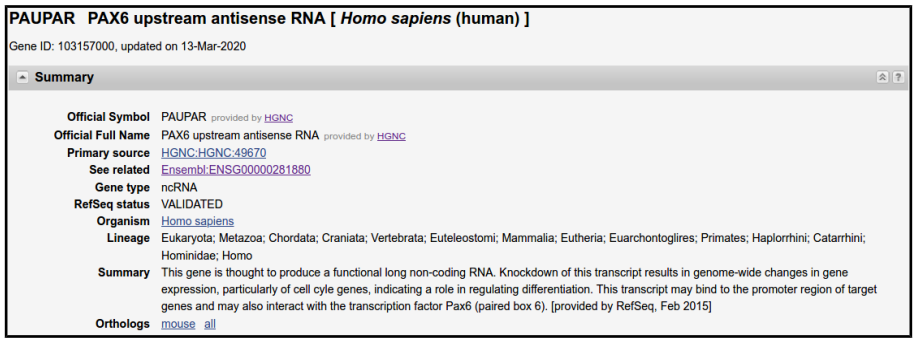
Special Warning! This is long and only of peripheral value. Before you start reading all that follows, please realise it is not really part of the exercise. It is just notes to justify my theory explaining the non-existence of **PAX6-AS1** in the **Ensembl** view of the **PAX6** region of the **Human Genome**.

A record for me? Of interest only to the more pedantic of you. Continue at your own peril, or hurry back to the main exercise instructions (**recommended**!). **Click Here to Return to the Exercise →**

PAUPAR and **PAX6-AS1** – seen differently from the **NCBI** and **Ensembl**.

The **NCBI** Story:

PAUPAR is a **PAX6 upstream antisense RNA gene**, one **Exon**, one **Transcript**, generates a **non-coding RNA**.



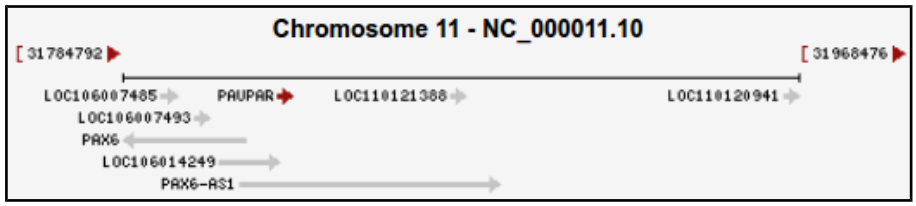
PAUPAR PAX6 upstream antisense RNA [*Homo sapiens* (human)]

Gene ID: 103157000, updated on 13-Mar-2020

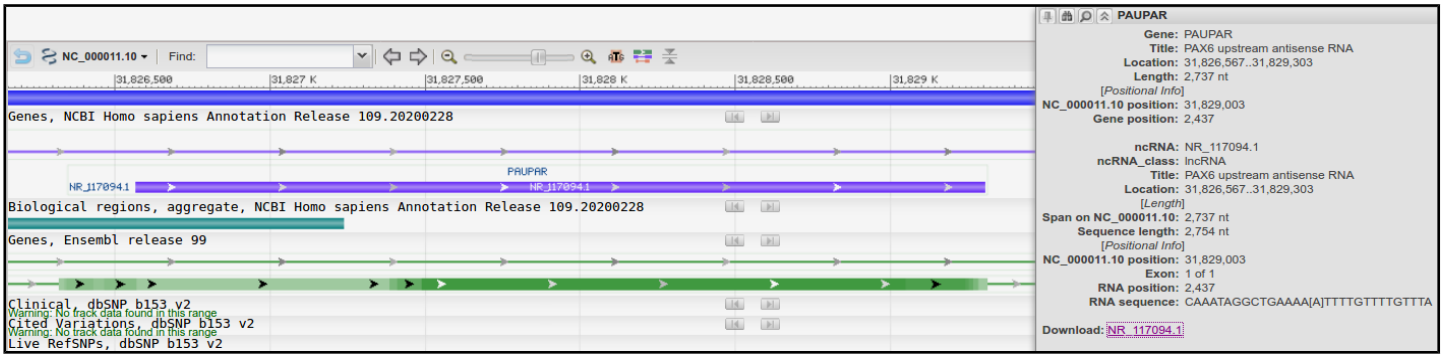
Summary

Official Symbol	PAUPAR provided by HGNC
Official Full Name	PAX6 upstream antisense RNA provided by HGNC
Primary source	HGNC:HGNC:49670
See related	Ensembl:ENSG00000281880
Gene type	ncRNA
RefSeq status	VALIDATED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Summary	This gene is thought to produce a functional long non-coding RNA. Knockdown of this transcript results in genome-wide changes in gene expression, particularly of cell cyle genes, indicating a role in regulating differentiation. This transcript may bind to the promoter region of target genes and may also interact with the transcription factor Pax6 (paired box 6). [provided by RefSeq, Feb 2015]
Orthologs	mouse all

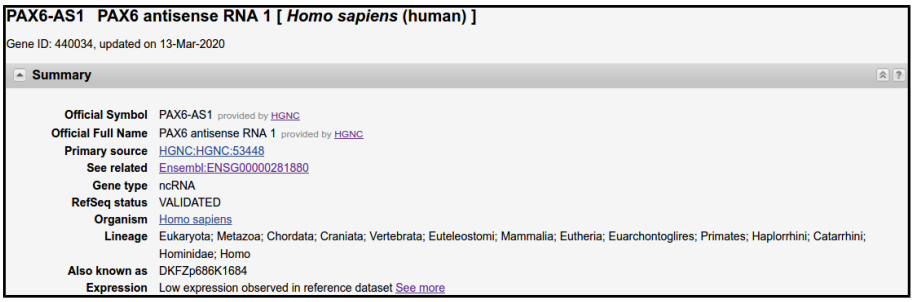
PAUPAR is small relative to **PAX6-AS1** and completely contained within **PAX6-AS1**.



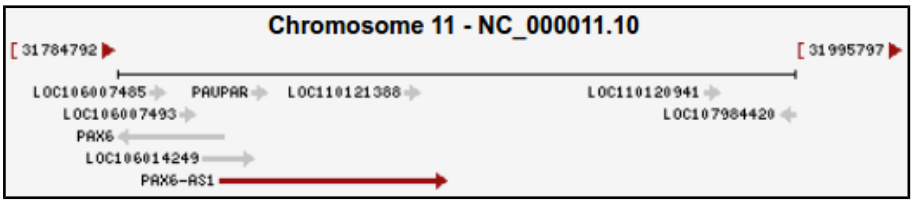
The single exon **Gene** spans **2,737 nt** of the **Genome**. The **RNA** itself is **2,754 nt**, the difference is the **PolyA tail**.



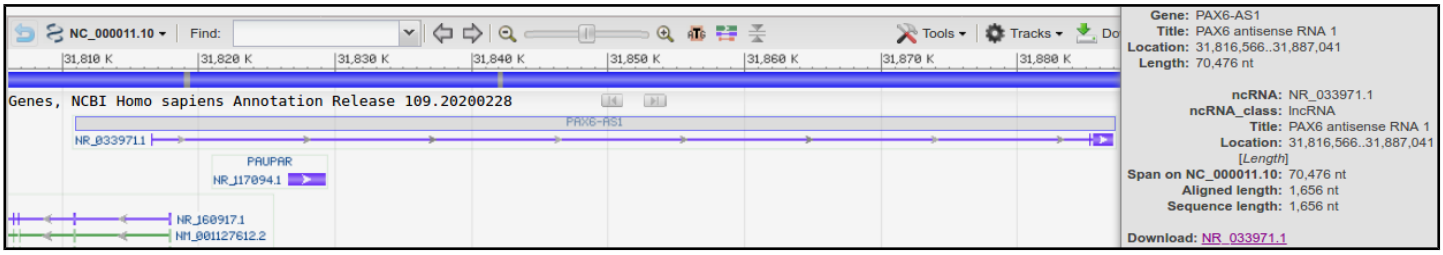
PAX6-AS1 is also a **PAX6 antisense RNA gene**, three **Exons**, one **Transcript**, generates a **non-coding RNA**.



PAX6-AS1 is large relative to **PAUPAR** and completely contains **PAUPAR**.



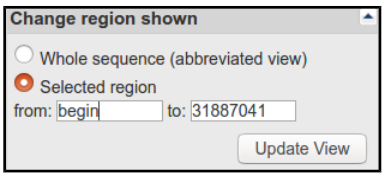
The one **Transcript** for this **Gene** involves all **three Exons**. One **Exon** is at the start of the **Gene**, the other **two** at the extreme end. The **Exons** span **1,656 nt** of the **Genome**. The **RNA** is also **1,656 nt** long (no **PolyA tail** this time).



Viewing the **GenBank Format** version of the sequence for the **PAX6-AS1** single **Transcript**, one can see it is a splicing of three **Exons** whose positions relative to the start of the **Gene** are:

1 to 74, **68,729 to 68,809** and **68,976 to 70,476**.

```
ncRNA           join(1..74,68729..68809,68976..70476)
                /ncRNA_class="lncRNA"
                /gene="PAX6-AS1"
                /gene_synonym="DKFZp686K1684"
                /product="PAX6 antisense RNA 1"
                /note="Derived by automated computational analysis using
                gene prediction method: BestRefSeq."
                /transcript_id="NR_033971.1"
                /db_xref="GeneID:440034"
                /db_xref="HGNC:HGNC:53448"
```



But … I need the **Exon** positions relative to the start of the **Chromosome**! Too lazy to do the readily available arithmetic, I change the **Selected region** (top right) to start at **1** (the system changes this to **begin**).

Then, in the new sequence view, I search (^**F**) for **PAX6-AS1**.

Now it is clear that the **three Exons** for **PAX6-AS1** span the regions of **Chromosome 11**:

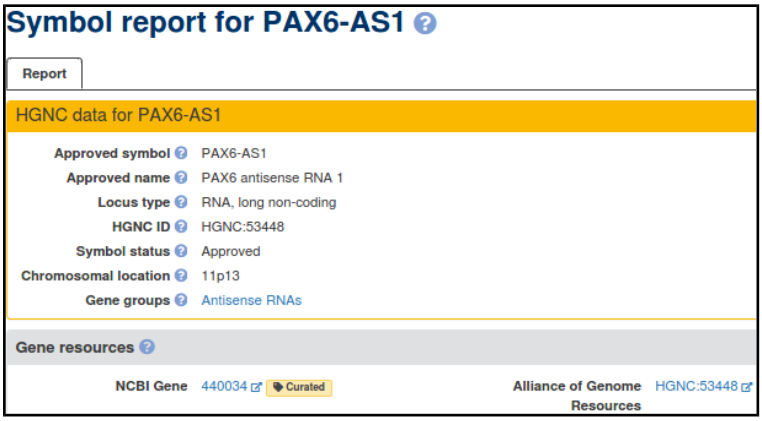
31,816,566 to 31,816,639
31,885,294 to 31,885,374
31,885,541 to 31,887,041

```
gene            31816566..31887041
                /gene="PAX6-AS1"
                /gene_synonym="DKFZp686K1684"
                /note="PAX6 antisense RNA 1; Derived by automated
                computational analysis using gene prediction method:
                BestRefSeq."
                /db_xref="GeneID:440034"
                /db_xref="HGNC:HGNC:53448"
ncRNA           join(31816566..31816639,31885294..31885374,
                31885541..31887041)
                /ncRNA_class="lncRNA"
                /gene="PAX6-AS1"
                /gene_synonym="DKFZp686K1684"
                /product="PAX6 antisense RNA 1"
                /note="Derived by automated computational analysis using
                gene prediction method: BestRefSeq."
                /transcript_id="NR_033971.1"
                /db_xref="GeneID:440034"
                /db_xref="HGNC:HGNC:53448"
```

The **Ensembl** Story:

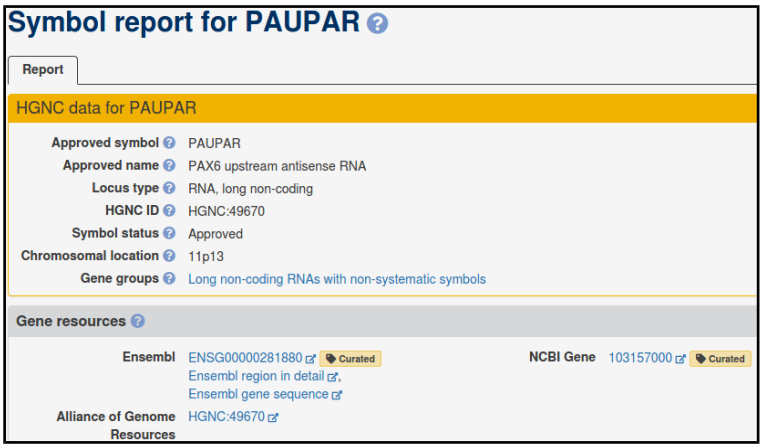
PAX6-AS1 does not exist according to **Ensembl**! Search for **PAX6-AS1** and you will get **PAX6** references, but no mention I could find of a **Gene** called exactly "**PAX6-AS1**".

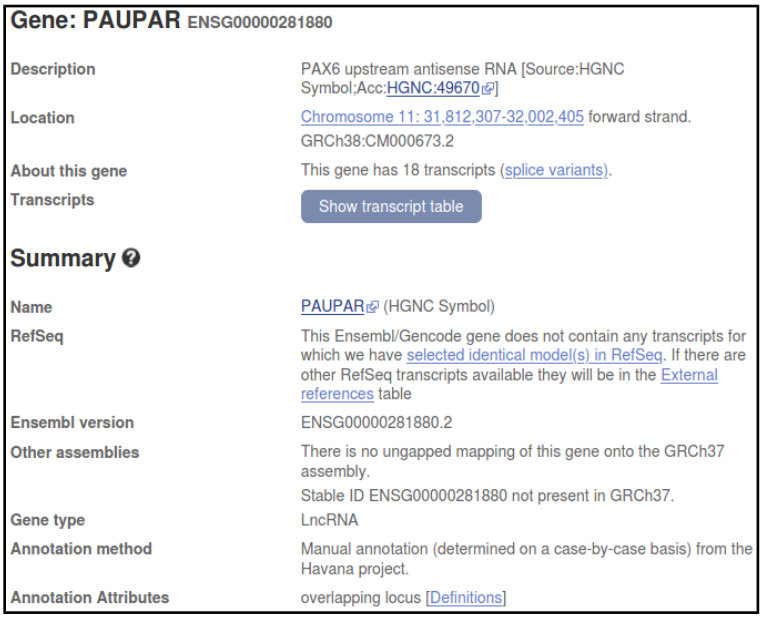
To be extra sure, I look at the **HGNC** entries for both **PAX6-AS1** ...



… and **PAUPAR**.

The former quotes only the **NCBI** as a **Gene resource**, the latter quotes both the **NCBI** and **Ensembl**.





PAUPAR is still a **PAX6 upstream antisense RNA Gene**. But … it now spans **Chromosome 11** from **31,812,307 to 32,002,405** (as opposed to the much smaller region, **31,826,567 to 31,829,303** predicted by the **NCBI**).

It is also considered to be a much more complex entity than reported by the **NCBI**. **eighteen Transcripts** as opposed to **one**!

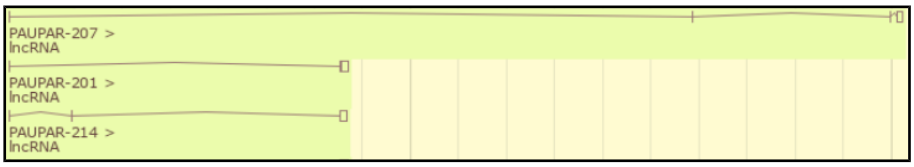
The genomic region assigned by **Ensembl** to **PAUPAR** comfortably includes the entire region assigned to **PAX6-AS1** by **NCBI** (**31,816,566 to31,887,041 - 70,476 bp**).

Note that there are no exact matches between the **NCBI Transcripts** and the **Ensembl PAUPAR Transcripts** for this **Gene**.

Not also that the **Ensembl** view of this **Gene** was radically different in in the analaysis of the previous **Assembly** of the **Human Genome**. THIS, I do not need to be told!!! The story I concocted last year to rationalise all this nonsense was radically different to the current fairy tale!!!

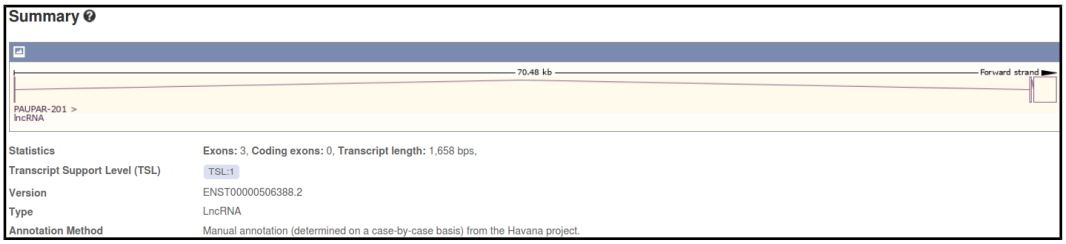
From the **Transcript table**, ranked in ascending order by **Transcript length**, I see that the **Transcript PAUPAR_201** (**1,658 bp**) is almost the same length as the **PAX6-AS1 Transcript** at the **NCBI (1,656 bp**).

Name	Transcript ID	bp ▲	Protein	Biotype
PAUPAR-202	ENST00000530146.2	1033	No protein	lncRNA
PAUPAR-207	ENST00000642818.1	1454	No protein	lncRNA
PAUPAR-214	ENST00000645848.1	1604	No protein	lncRNA
PAUPAR-201	ENST00000506388.2	1658	No protein	lncRNA
PAUPAR-209	ENST00000643671.1	1663	No protein	lncRNA
PAUPAR-217	ENST00000646959.1	1724	No protein	lncRNA
PAUPAR-213	ENST00000645824.1	1746	No protein	lncRNA
PAUPAR-216	ENST00000646221.1	1750	No protein	lncRNA
PAUPAR-211	ENST00000643931.1	1816	No protein	lncRNA
PAUPAR-206	ENST00000642614.1	1877	No protein	lncRNA
PAUPAR-218	ENST00000647350.1	1888	No protein	lncRNA
PAUPAR-205	ENST00000642549.1	1996	No protein	lncRNA
PAUPAR-210	ENST00000643826.1	2071	No protein	lncRNA
PAUPAR-215	ENST00000645942.1	2112	No protein	lncRNA
PAUPAR-204	ENST00000642237.1	2232	No protein	lncRNA
PAUPAR-208	ENST00000643436.1	2239	No protein	lncRNA
PAUPAR-212	ENST00000644607.1	2341	No protein	lncRNA
PAUPAR-203	ENST00000630360.1	2965	No protein	lncRNA



I look at the corresponding graphic view and see that **PAUPAR_201** has a very similar **three Exon** structure to the **NCBI PAX6-AS1 Transcript** and spans a very similar **Genomic Region**.

This structural similarlty is even clearer from observed from the graphic specific to the **PAUPAR-201 Transcript**.

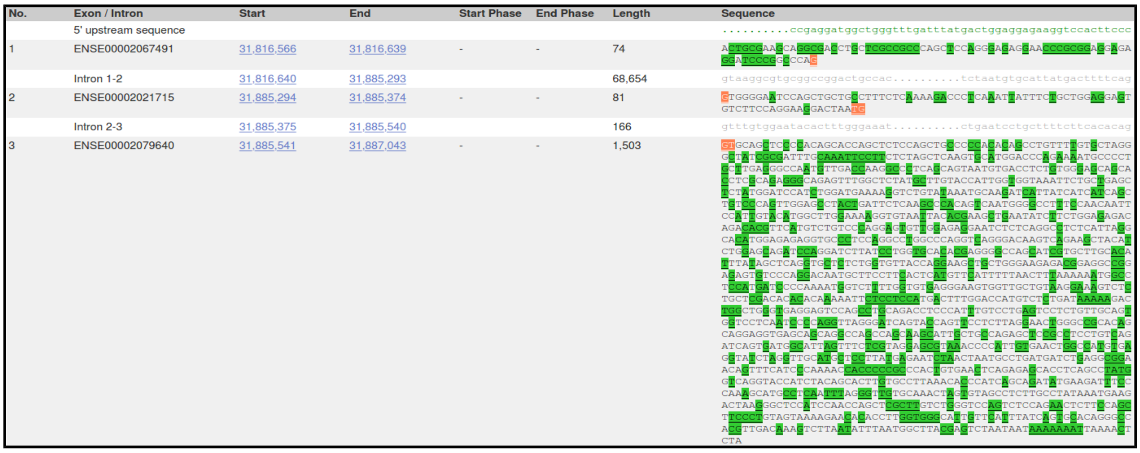


From the **Transcript** tag, click on the **Exons** link (top left hand corner), then click on **Download sequence** and ask to **Preview** the sequence.

This is one way, probably the most stupid, to ascertain the exact locations of the **three Exons** that form **PAUPAR-201**.

```
>PAUPAR-201 ENSE00002067491 exon:lncRNA
ACTGCGAAGCAGGCGACCTGCTCGCCGCCCAGCTCCAGGGAGAGGAACCCGCGGAGGAGA
GGATCCCGGCCCAG
>PAUPAR-201 ENSE00002021715 exon:lncRNA
GTGGGGAATCCAGCTGCTGCCTTTCTCAAAAGACCCTCAAATTATTTCTGCTGGAGGAGT
GTCTTCCAGGAAGGACTAATG
>PAUPAR-201 ENSE00002079640 exon:lncRNA
GTGCAGCTCCCCACAGCACCAGCTCTCCAGCTGCCCCCACACAGCCTGTTTTGTGCTAGG
GCTATCGCGATTTGCAAATTCCTTCTCTAGCTCAAGTGCATGGACCCAGAAAATGCCCCT
GCTTGAGGGCCAATGTTGACCAAGGCCCTCAGCAGTAATGTGACCTCTGTGGGAGCAGCA
CCTCGCAGAGGGCAGAGTTTGGCTCTATGCTTGTACCATTGGTGGTAAATTCTGCTGAGC
TCTATGGATCCATCTGGATGAAAAGGTCTGTATAAATGCAAGATCATTATCATCATCAGC
TGTCCCAGTTGGAGCCTACTGATTCTCAAGCCCACAGTCAATGGGGCCTTTCCAACAATT
CCATTGTACATGGCTTGGAAAAGGTGTAATTACACGAAGCTGAATATCTTCTGGAGAGAC
AGACACGTTCATGTCTGTCCCAGGAGTGTTGGAGAGGAATCTCTCAGGCCTCTCATTAGG
CACATGGAGAGAGGTGCCCTCCAGGCCTGGCCCAGGTCAGGGACAAGTCAGAAGCTACAT
CTGGAGCAGATCCAGGATCTTATCCTGGTGCACACGAGGGGCCAGCATCGTGCTTGCACA
TTTATAGCTCAGGTGCTCTCTGGTGTTACCAGGAAGCTGCTGGGAAGAGACGGAGGCCGG
AGAGTGTCCCAGGACAATGCTTCCTTCACTCATGTTCATTTTTAACTTTAAAAAATGGCC
TCCATGATCCCCAAAATGGTCTTTTGGTGTGAGGGAAGTGGTTGCTGTAAGGAAAGTCTC
TGCTCGACACACACAAAAATTCTCCTCCATGACTTTGGACCATGTCTCTGATAAAAAGAC
TGGCTGGGTGAGGAGTCCAGCCTGCAGACCTCCCATTTGTCCTGAGTCCTCTGTTGCAGT
GGTCCTCAATCCCCAGGTTAGGGATCAGTACCAGTTCCTCTTAGGAACTGGGCCGCACAG
CAGGAGGTGAGCAGCAGGCCAGCCAGCAAGCATTGCTGCCAGAGCTCCGCCTCCTGTCAG
ATCAGTGATGGCATTAGTTTCTCGTAGGAGCGTAAACCCCATTGTGAACTGGCCATGTGA
GGTATCTAGGTTGCATGCTCCTTATGAGAATCTAACTAATGCCTGATGATCTGAGGCGGA
ACAGTTTCATCCCAAAACCACCCCCGCCCACTGTGAACTCAGAGAGCACCTCAGCCTATG
GTCAGGTACCATCTACAGCACTTGTGCCTTAAACACCCATCAGCAGATATGAAGATTTCC
CAAAGCATGCCTCAATTTAGGGTTGTGCAAACTAGTGTAGCCTCTTGCCTATAAATGAAG
ACTAAGGGCTCCATCCAACCAGCTCGCTTGTCTGGGTCCAGTCTCCAGAACTCTTCCAGC
TTCCCTGTAGTAAAAGAACACACCTTGGTGGGCATTGTTCATTTATCAGTGCACAGGGCC
ACGTTGACAAAGTCTTAATATTTAATGGCTTACGAGTCTAATAATAAAAAAATTAAAACT
CTA
```

Far better to notice that, once the **Exon** link is activated, the **Exons** locations (and much more) are laid out for you to admire.



Information provided includes the Exon locations. They are:

31,816,566 to 31,816,639

31,885,294 to 31,885,374

31,885,541 to 31,887,043

Exactly matching the **Exon** locations for **PAX6-AS1** at the **NCBI**, except for a couple of extra base pairs at the end of the third **Exon**.

I conclude that **Ensembl** predicts the non-coding **RNA PAX6-AS1**, but as a part of the **Gene PAUPAR** (specifically the **Transcript PAUPAR-201**) rather than as a separate **Gene**.

Totally unnecessarily, I celebrate this revelation by aligning the sequence of **PAX6-AS1** from the **NCBI** with the sequence of **PAUPAR-201** from **Ensembl**.

```
PAUPAR-201        1  ACTGCGAAGCAGGCGACCTGCTCGCCGCCCAGCTCCAGGGAGAGGAACCC     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1          1  ACTGCGAAGCAGGCGACCTGCTCGCCGCCCAGCTCCAGGGAGAGGAACCC     50

PAUPAR-201       51  GCGGAGGAGAGGATCCCGGCCCAGGTGGGGAATCCAGCTGCTGCCTTTCT    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1         51  GCGGAGGAGAGGATCCCGGCCCAGGTGGGGAATCCAGCTGCTGCCTTTCT    100

PAUPAR-201      101  CAAAAGACCCTCAAATTATTTCTGCTGGAGGAGTGTCTTCCAGGAAGGAC    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1        101  CAAAAGACCCTCAAATTATTTCTGCTGGAGGAGTGTCTTCCAGGAAGGAC    150

PAUPAR-201      151  TAATGGTGCAGCTCCCCACAGCACCAGCTCTCCAGCTGCCCCCACACAGC    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1        151  TAATGGTGCAGCTCCCCACAGCACCAGCTCTCCAGCTGCCCCCACACAGC    200

PAUPAR-201      201  CTGTTTTGTGCTAGGGCTATCGCGATTTGCAAATTCCTTCTCTAGCTCAA    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1        201  CTGTTTTGTGCTAGGGCTATCGCGATTTGCAAATTCCTTCTCTAGCTCAA    250

                                        •••

PAUPAR-201     1501  GCTTGTCTGGGTCCAGTCTCCAGAACTCTTCCAGCTTCCCTGTAGTAAAA   1550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1       1501  GCTTGTCTGGGTCCAGTCTCCAGAACTCTTCCAGCTTCCCTGTAGTAAAA   1550

PAUPAR-201     1551  GAACACACCTTGGTGGGCATTGTTCATTTATCAGTGCACAGGGCCACGTT   1600
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1       1551  GAACACACCTTGGTGGGCATTGTTCATTTATCAGTGCACAGGGCCACGTT   1600

PAUPAR-201     1601  GACAAAGTCTTAATATTTAATGGCTTACGAGTCTAATAATAAAAAAATTA   1650
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6-AS1       1601  GACAAAGTCTTAATATTTAATGGCTTACGAGTCTAATAATAAAAAAATTA   1650

PAUPAR-201     1651  AAACTCTA    1658
                     ||||||
PAX6-AS1       1651  AAACTC--    1656
```

Complete agreement except for the **2 Base Pair** overhang at the very end.

Click Here to Return to the Exercise →

Efficient design for **Ensembl Protein CodingTranscript** prediction for the **Human Genome.**

I tidy this up when I am sure of my facts …

As described already, assuming a suitable comprehensive set of appropriate sequences, the location and structure of all transcripts could be determined by a simple two stage operation:

mapping all quality mRNA sequence onto the genome to discover the exons of the mRNA

mapping all quality proteins onto the geneome to discover the CDSs

or maybe the other way round? Both would work?

To do this efficiently:

first genscan … find most genes

then CCDS (CDS agreed by pipeline, Vega and NCBI … Human/Mouse specific at present) search on genscan hits only …. reveals coding regions accurately

then mRNA (RefSeq and other high quality data/predictions) only on CCDS hits … reveals UTRs accurately

Why it is reasonable to not regard a match of a **RefSeq** mRNA with the **Genome** as, by itself, sufficient evidence to uniquely predict a transcript.

RefSeq mRNA sequences are not determined by careful sequencing of individual mRNA/cDNA. If they were, it would be difficult to argue with the **NCBI** approach of regarding a quality match between a **RefSeq** mRNA and the genome as sufficient evidence to predict the location of a transcript.

However, **RefSeq** mRNA sequences are actually computed from assemblies of many single pass, poor quality, cDNA sequences (**EST**s).

Ensembl regards these sequences as good evidence …. but not conclusive by themselves.

NCBI appears to rely more on the reliability of RefSeq mRNA sequences.

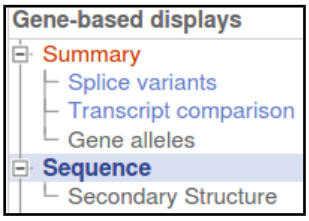
Ensembl uses both the sequences of **RefSeq** mRNAs and those of their protein products (the **RefSeq** entries whose **Accession Code**s commence **NP_**) to predict transcripts, however, **Ensembl** appears to have less blind faith in the accuracy of these data than the **NCBI**.

Note: There is no “one to one” correspondence between **RefSeq** mRNAs and transcript predictions. All **RefSeq** mRNAs are referenced, but ***two*** are used to support the single third transcript in the list. If **Ensembl** regarded **RefSeq** mRNAs as “perfect” (as the **NCBI** appears to do) this would clearly be nonsense!

Above just jumble notes from previous versions … need some Ben input

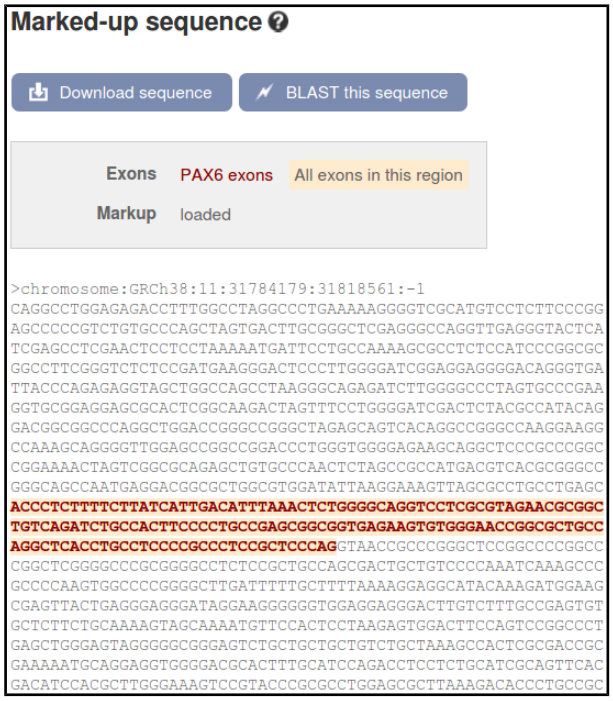
Click Here to Return to the Exercise →

Downloading the **PAX6 Genomic** region for later analysis.

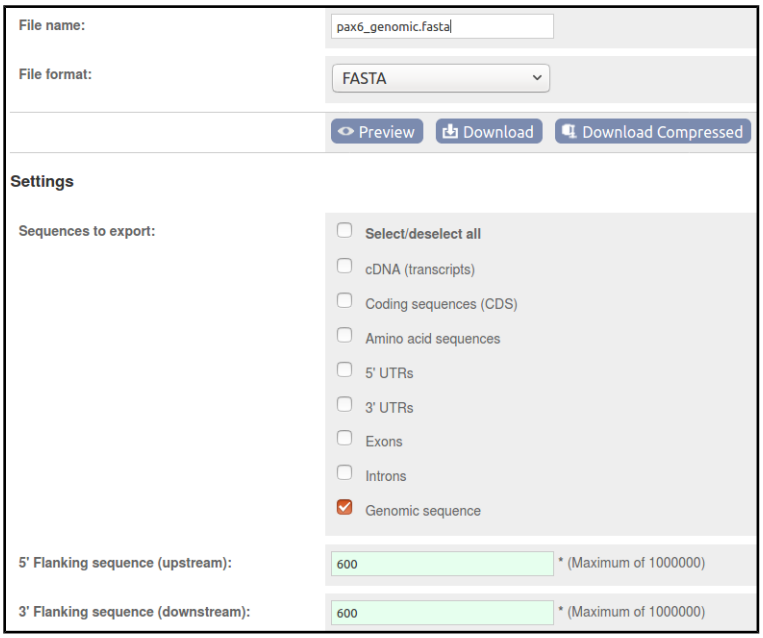


First, click on the **Sequence** link in the top left hand corner of the page.

The **Transcript Graphic** will be replaced by a display of the sequence of the **PAX6** region of the genome.



Exons will be tastefully highlighted for you delectation. The display includes **600** base pairs of **Flanking Sequence** (**3'** and **5')** which are included (by default) when the sequence is downloaded



Now chose to Download sequence. The **Download sequence** form will burst into view.

Set the **File name:** to `pax6_genomic.fasta`

Set the **File format:** to **FASTA**

Accept the default **600** base pairs for both the **5'Flanking sequence (upstream):** and the **3´ Flanking sequence (downstream):**.

Finally, click on the Download button and do whatever it takes to move the file you create to somewhere sensible on your **Desktop**.

Using whatever text editor is most convenient, edit your file to change the first word of the first line of the file to contain information, from **11** to **pax6_genomic**. This first word is defined as the sequence identifier in **FASTA** format. **pax6_genomic** is a far more informative identification than **11** (simply the Chromosome number).

```
>pax6_genomic dna:chromosome chromosome:GRCh38:11:31784179:31818561:-1
CAGGCCTGGAGAGACCTTTGGCCTAGGCCCTGAAAAAGGGGTCGCATGTCCTCTTCCCGG
AGCCCCCGTCTGTGCCCAGCTAGTGACTTGCGGGCTCGAGGGCCAGGTTGAGGGTACTCA
TCGAGCCTCGAACTCCTCCTAAAAATGATTCCTGCCAAAAGCGCCTCTCCATCCCGGCGC
GGCCTTCGGGTCTCTCCGATGAAGGGACTCCCTTGGGGATCGGAGGAGGGGACAGGGTGA
TTACCCAGAGAGGTAGCTGGCCAGCCTAAGGGCAGAGATCTTGGGGCCCTAGTGCCCGAA
GGTGCGGAGGAGCGCACTCGGCAAGACTAGTTTCCTGGGGATCGACTCTACGCCATACAG
GACGGCGGCCCAGGCTGGACCGGGCCGGGCTAGAGCAGTCACAGGCCGGGCCAAGGAAGG
CCAAAGCAGGGGTTGGAGCCGGCCGGACCCTGGGTGGGGAGAAGCAGGCTCCCGCCCGGC
CGGAAAACTAGTCGGCGCAGAGCTGTGCCCAACTCTAGCCGCCATGACGTCACGCGGGCC
GGGCAGCCAATGAGGACGGCGCTGGCGTGGATATTAAGGAAAGTTAGCGCCTGCCTGAGC
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGC
TGTCAGATCTGCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCC
```

Click Here to Return to the Exercise →

DPJ – 2020.04.28

The next investigation might be to discover "How many protein isoforms might there be for **PAX6**?".

Well, whilst the **Ensembl** transcript list is still in view, glance down the **Protein** column which displays the size of the protein products for each transcript. Clearly insufficient evidence for a serious **isoform** count, but enough to set a lower limit, as the same **isoform** cannot be more than one length! If there were not so very many! One might count how many different lengths of proteins were listed. I tried to do this, but I gave up around **twenty-something**. Let us be content to declare that there are **lots**. The most likely looking ones are either **422** or **436** amino acids long. Some of the others might cause a raised eyebrow or two, especially the one that is **3** amino acids long (third from last **Protein coding** entry in the list)? But, who are we to question! **Lots** is the informal **Ensembl** minimum total.

For a more detailed view of the predicted transcripts, click on the Show transcript table link. The transcript predictions are now presented in the form of a table. The protein coding transcripts are all at the top of the table. I counted **56**, but I would not claim to be completely accurate, I wavered half way down the list! Lots more than the **NCBI** anyway.

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
PAX6-245	ENST00000638914.2	7290	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	GENCODE basic APPRIS ALT1
PAX6-270	ENST00000640368.1	6975	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	GENCODE basic APPRIS ALT1
PAX6-283	ENST00000643871.1	6944	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_000280 NM_001258464 NP_000271 NP_001245393	GENCODE basic APPRIS ALT1
PAX6-231	ENST00000606377.6	6901	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:1 GENCODE basic APPRIS ALT1
PAX6-209	ENST00000419022.6	6888	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001604 NP_001595	TSL:1 GENCODE basic APPRIS P4
PAX6-203	ENST00000379109.7	3182	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:2 GENCODE basic APPRIS ALT1
PAX6-273	ENST00000640610.1	2730	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	GENCODE basic APPRIS ALT1
PAX6-284	ENST00000645710.1	2688	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258462 NP_001245391	GENCODE basic APPRIS ALT1
PAX6-259	ENST00000639916.1	2622	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001258465 NP_001245394	GENCODE basic APPRIS ALT1
PAX6-243	ENST00000638903.1	2620	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	GENCODE basic APPRIS P4
PAX6-207	ENST00000379129.7	2614	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS ALT1
PAX6-202	ENST00000379107.7	2579	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS ALT1
PAX6-208	ENST00000379132.8	2576	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:5 GENCODE basic APPRIS ALT1
PAX6-282	ENST00000640975.1	2553	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001310158 NP_001297087	GENCODE basic APPRIS P4
PAX6-257	ENST00000639409.1	2450	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258463 NP_001245392	GENCODE basic APPRIS P4

Ensembl uses both the sequences of **RefSeq** mRNAs and those of their protein products (the **RefSeq** entries whose **Accession Codes** commence **NP_**) to predict transcripts, however, **Ensembl** appears to have less blind faith in the accuracy of these data than the **NCBI**.

Note: There is no "one to one" correspondence between **RefSeq** mRNAs and transcript predictions. All **11 RefSeq** mRNAs are referenced, but ***two*** are used to support the single third transcript in the list. If **Ensembl** regarded **RefSeq** mRNAs as "perfect" (as the **NCBI** appears to do) this would clearly be nonsense! We should discuss why it is reasonable not to not to accept the infallibility of a **RefSeq** mRNA matches with the **Genome**.

PAX6-220	ENST00000525535.2	875	3aa	Protein coding	-	-	-	CDS 3' incomplete TSL:3
PAX6-260	ENST00000639920.1	676	72aa	Protein coding	-	A0A1W2PR58	-	CDS 3' incomplete
PAX6-256	ENST00000639394.1	1988	163aa	Nonsense mediated decay	-	A0A1W2PQW3	-	
PAX6-227	ENST00000533156.2	848	No protein	Processed transcript	-	-	-	TSL:5
PAX6-213	ENST00000464174.6	846	No protein	Processed transcript	-	-	-	TSL:5
PAX6-222	ENST00000530373.6	785	No protein	Processed transcript	-	-	-	TSL:4
PAX6-223	ENST00000530714.6	650	No protein	Processed transcript	-	-	-	TSL:4
PAX6-267	ENST00000640251.1	649	No protein	Processed transcript	-	-	-	
PAX6-229	ENST00000534353.5	540	No protein	Processed transcript	-	-	-	TSL:4
PAX6-254	ENST00000639203.1	532	No protein	Processed transcript	-	-	-	
PAX6-233	ENST00000638278.1	417	No protein	Processed transcript	-	-	-	
PAX6-275	ENST00000640617.1	412	No protein	Processed transcript	-	-	-	
PAX6-279	ENST00000640819.1	368	No protein	Processed transcript	-	-	-	
PAX6-228	ENST00000533333.5	6173	No protein	Retained intron	-	-	-	TSL:2
PAX6-216	ENST00000474783.2	4392	No protein	Retained intron	-	-	-	TSL:2
PAX6-214	ENST00000470027.7	3587	No protein	Retained intron	-	-	-	TSL:2
PAX6-265	ENST00000640172.1	2525	No protein	Retained intron	-	-	-	

Looking further down the list you will see that many **Ensembl** protein coding transcripts are predicted without reference to any **RefSeq** entry.

Hover over the evidence **Flags** associated with the transcript predictions towards the end of the list. How reliable would you judge these predictions to be?

We could go on. Other sources (not necessarily **Genome Databases**) would count the transcripts differently again. Perhaps the best answer to the question "How many transcripts are there for the **PAX6** gene" is "**Several**".