Earlham Institute summer school on bioinformatics

25-29 July 2016

Basic Bioinformatics Sessions

Practical 1: Databases and Tools

Sunday 7 August 2016

Investigating gene(s) associated with Aniridia

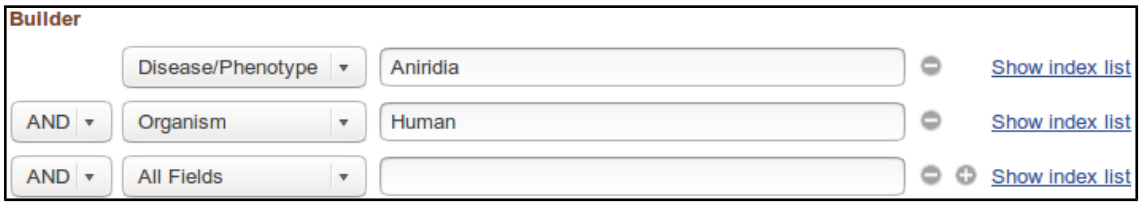
As a starting point for this exercise, imagine you have a vital interest in discovering the main human gene responsible for the terrible disease of the eye, **aniridia**. There are many ways (including **google**!) you could discover what this gene might be. I choose to delve into the vast seas of knowledge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

So, go to the **Home Page** of the **The National Center for Biotechnology Information (NCBI)** ("**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.



Name/Gene ID	Description	Location	Aliases	MIM
PAX6 ID: 5080	paired box 6 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31784792..31817961, complement)	AN, AN2, D11S812E, FVH1, MGDA, WAGR	607108
WT1 ID: 7490	Wilms tumor 1 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (32387775..32435535, complement)	AWT1, EWS-WT1, GUD, NPHS4, WAGR, WIT-2, WT33	607102
ITPR1 ID: 3708	inositol 1,4,5-trisphosphate receptor type 1 [*Homo sapiens* (human)]	Chromosome 3, NC_000003.12 (4493348..4847840)	ACV, CLA4, INSP3R1, IP3R, IP3R1, PPP1R94, SCA15, SCA16, SCA29	147265
ELP4 ID: 26610	elongator acetyltransferase complex subunit 4 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31509729..31784525)	AN, C11orf19, PAX6NEB, PAXNEB, dJ68P15A.1, hELP4	606985
DEL11P13 ID: 100528024	Wilms tumor, aniridia, genitourinary anomalies and mental retardation syndrome [*Homo sapiens* (human)]		C11DELp13, WAGR	194072

Just a few genes survive. OK, this is just an exercise, so trust me … it is **PAX6** that is the most interesting gene, in this context. This is the one to follow up by clicking on the link to its details.

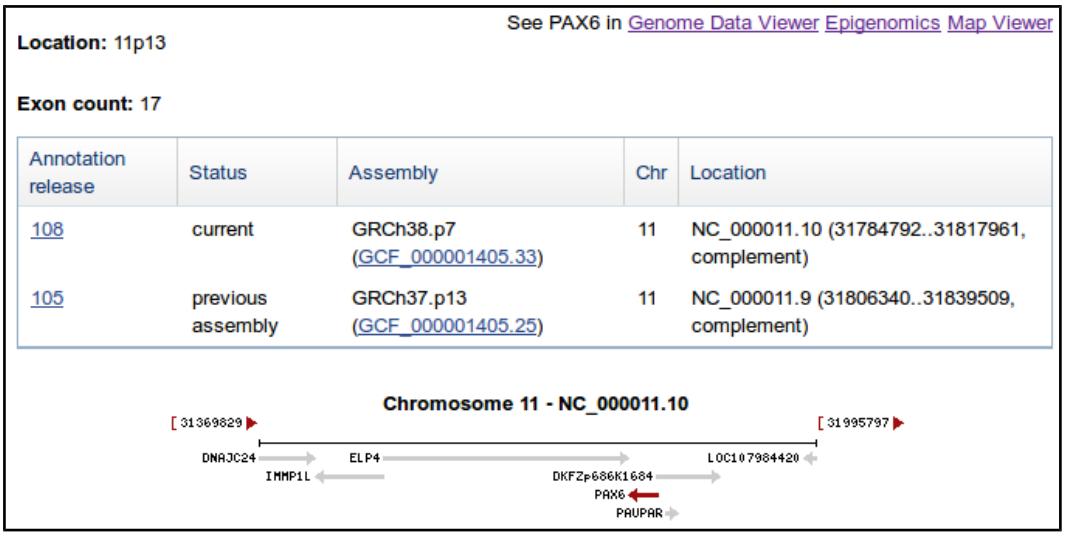
From the Summary section one can conclude (sticking to the features that pertain to this exercise) that:

- there are two major domains, a paired domain and a homeobox, both of which bind DNA
- the gene regulates transcription (is a transcription factor)
- there is more than one isoform, and thus more than one transcript.

Summary This gene encodes a homeobox and paired domain-containing protein that binds DNA and functions as a regulator of 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 result in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2015]

From the Genomic context section it can be seen that:

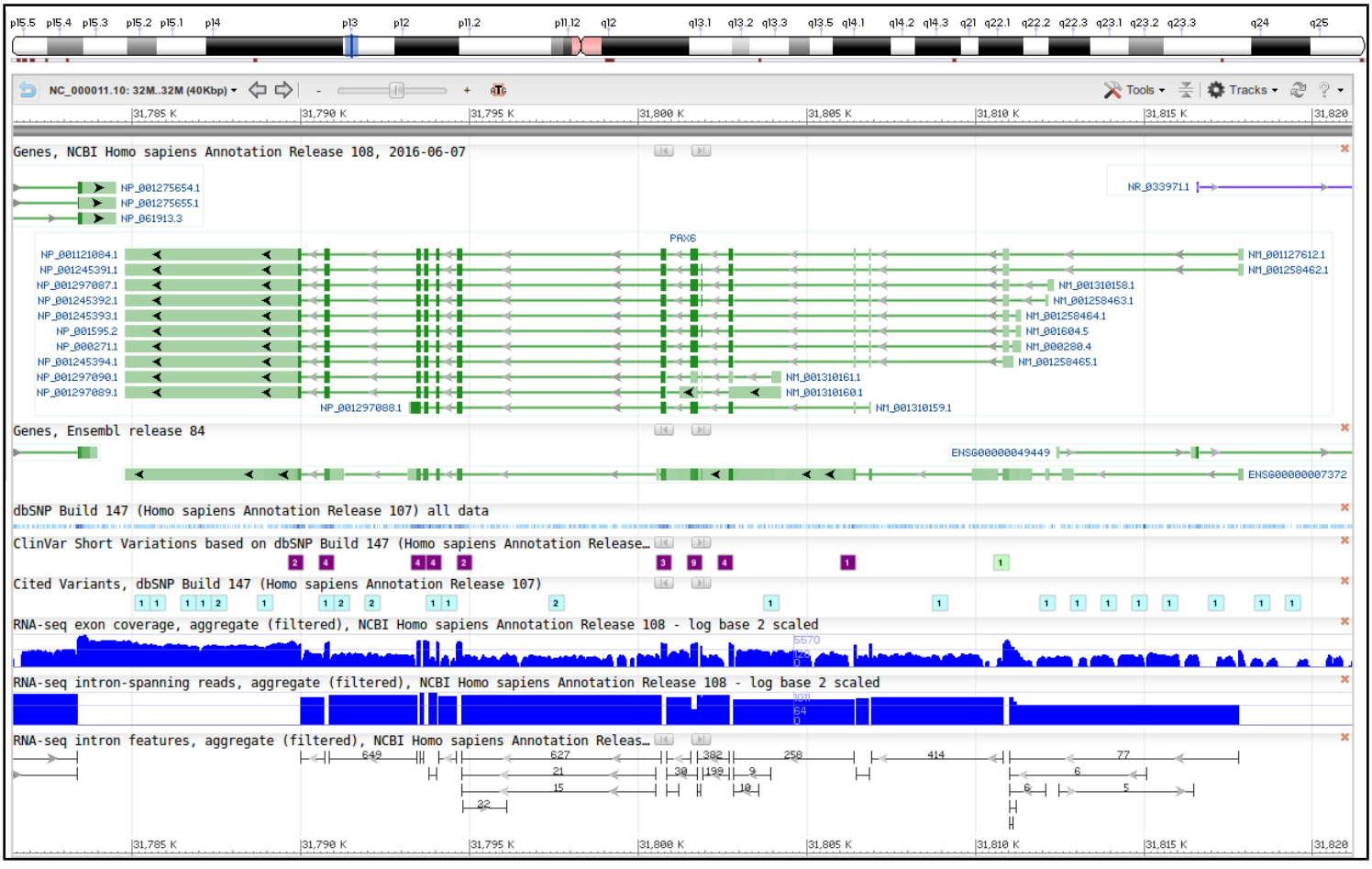
- **PAX6** is situated on **Chromosome 11**, band **p13**
- **PAX6** is on the complementary strand relative to that chosen by **Map Viewer** to represent **Chromosome 11**
- **ELP4** (another gene in the list of human genes associated with **Aniridia**) is exceedingly close, on the opposite strand to **PAX6**. This might be worthy of investigation, at another time?



- There are **17** exons for **PAX6**. 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 **17** 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!

Click either the **Genome Data Viewer** or the **MapViewer** link. Both offer you essentially the same story, the choice really is cosmetic. Do you like your genomes vertical or horizontal. I am a horizontal man myself, so I prefer the **Genome Data Viewer**. The data is from the **Map Viewer Genome Database**, whichever choice you make.

I reproduce both views here. The **Genome Data Viewer** picture is included in the **PAX6** gene page for free, so maybe the **MapViewer link** is the best one for you to choose? Or both, of course! First consider the marginally clearer and simpler **Genome Data Viewer** picture.



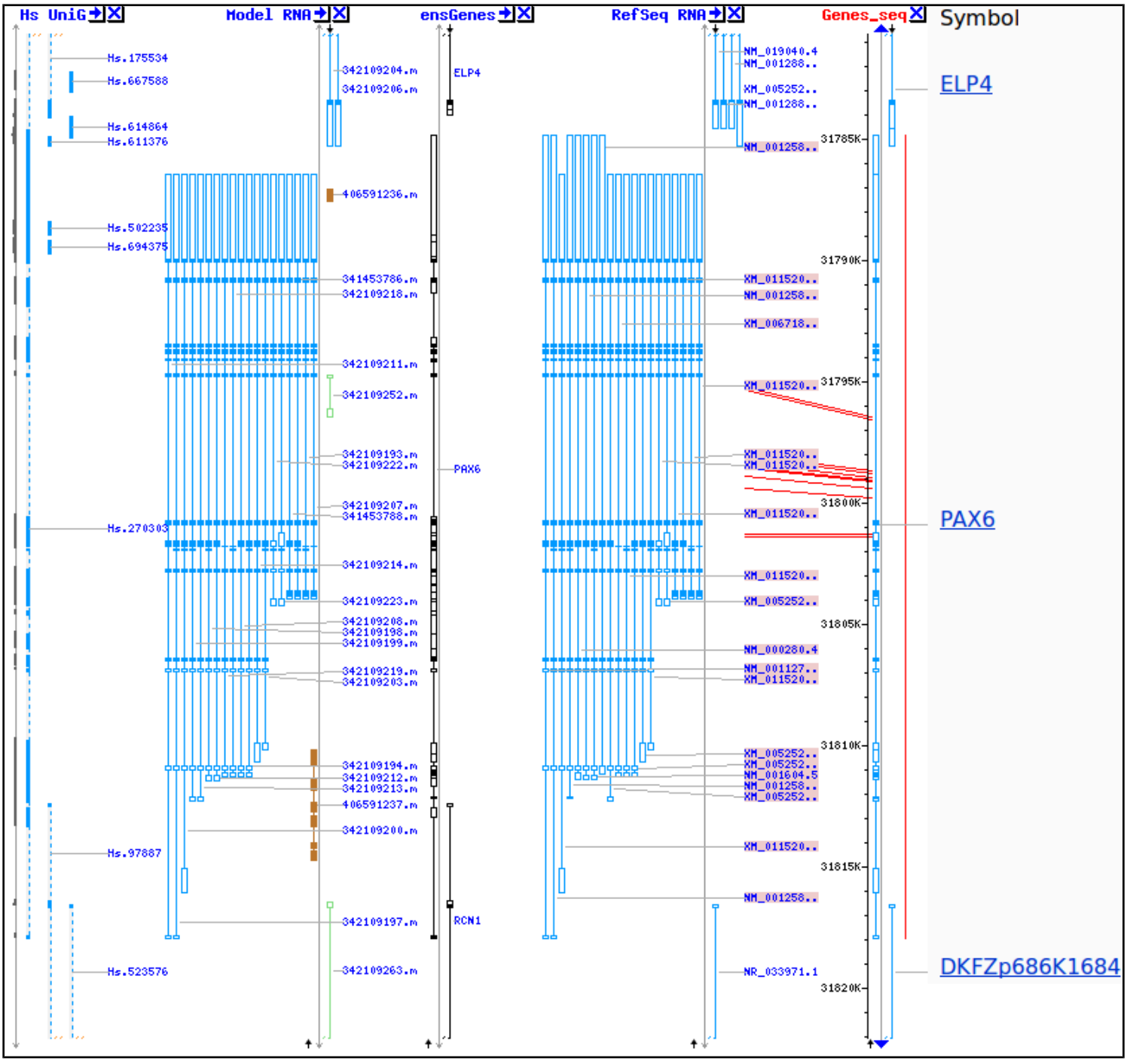
So, if I tell you the region displayed is the entire **PAX6** region of **Chromosome 11** and the green lines labelled on the right as something beginning with **NM_** represent the different transcripts, can you now say how many transcripts

there are according to this view? In passing, the blobs along each line represent the exons. Dark blobs are coding exons. Light blobs represent the exons that form the **3'/5' UTR** regions of each transcript. The Introns are the pale green lines joining the blobs together.

You need also to realise that the prediction of the transcripts shown here are based on database searches of all Human mRNA sequences stored in **RefSeq** against this region of the genome. The theory is that every human mRNA sequence must match (nearly) perfectly somewhere in the human genome. Where it matches, there must be the genomic DNA from which the mRNA was transcribed. How charmingly simple!

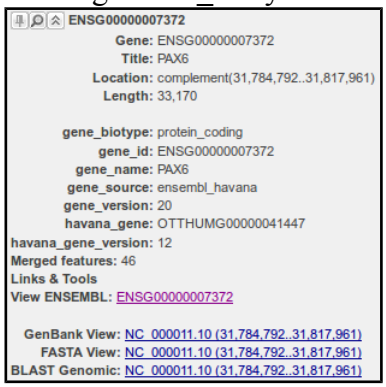
To differentiate between coding and non-coding exons of a transcript, why not compare all human proteins with the genome (after suitable translation to amino acid codes in all six reading frames). They too must match near perfectly somewhere, identifying the coding sequence (CDS) of each transcript. Transcript fully located. Job done! Of course, it does not always work so very neatly, but we need not admit that for the moment at least.

Comparing proteins with the genome is clumsy. compute intensive, slow. For major organisms (currently just Human and Mouse), specially comprehensive databases of extremely reliable **DNA Coding Sequences** have been constructed. Searching with these databases is so much more efficiently searching against proteins, serves exactly the same purpose and is thus very much preferred.



OK, times up, how many transcripts are predicted for **PAX6** then? It would appear the answer depends which viewer you chose to view the **PAX6** region? I count **11** in the **Genome Data Viewer** but **20** in the **MapViewer** picture. There is an explanation, but really!! Sometimes I wonder if there is ever any straight answers out there at all.

The explanation? Well, there will be one transcript predicted for every **PAX6** mRNA sequence in **RefSeq**. There are **20** mRNA sequences in **RefSeq**, however, **9** of these are not as well evidenced as the other **11**. You can tell the difference as the "good" one have names beginning with **NM_** and the "less good" ones begin **XM_**. If you have very good eyesight, you can confirm that there are **11 NM_**s and **9 XM_**s in the **Map Viewer** picture. The **Genome Data Viewer** declines to show the less worthy matches[1].



Of course, deciding how many transcripts there might be is not that "simple". Move back to the page describing the **PAX6** gene. In the familiar graphic at the top of the **Genome regions, transcripts and products** section you will find routes to corresponding information from the **Ensembl Genome Database**. Hover over the **PAX6** (also known as **ESNG00000007372**, by **Ensembl** and other close friends) green line in the bottom half of the picture. You will be rewarded by cheery gray box full of links to **Ensembl** and other exciting places.

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq
PAX6-201	ENST00000419022	6922	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258462 NM_001310158 NM_001310161 NP_001245391 NP_001297087 NP_001297090
PAX6-202	ENST00000606377	6860	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258463 NM_001310161 NP_001245392 NP_001297090
PAX6-009	ENST00000379129	2616	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
PAX6-011	ENST00000379107	2591	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-
PAX6-008	ENST00000379132	2574	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001127612 NP_001121084
PAX6-003	ENST00000379123	2160	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_000280 NM_001258464 NP_000271 NP_001245393
PAX6-001	ENST00000379115	1763	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001604 NP_001595
PAX6-002	ENST00000241001	1631	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-
PAX6-005	ENST00000379111	1627	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001258465 NP_001245394
PAX6-004	ENST00000379109	2157	422aa	Protein coding	-	P26367 Q66SS1	-
PAX6-020	ENST00000525535	677	2aa	Protein coding	-	-	-
PAX6-021	ENST00000524853	574	57aa	Protein coding	-	E9PKM0	-
PAX6-012	ENST00000423822	567	61aa	Protein coding	-	B1B1I9	-
PAX6-016	ENST00000455099	497	124aa	Protein coding	-	B1B1J0	-
PAX6-013	ENST00000438681	455	38aa	Protein coding	-	B1B1I8	-
PAX6-029	ENST00000533156	847	No protein	Processed transcript	-	-	-
PAX6-014	ENST00000471303	782	No protein	Processed transcript	-	-	-
PAX6-027	ENST00000531910	643	No protein	Processed transcript	-	-	-
PAX6-015	ENST00000481563	613	No protein	Processed transcript	-	-	-
PAX6-028	ENST00000530373	572	No protein	Processed transcript	-	-	-
PAX6-025	ENST00000530714	567	No protein	Processed transcript	-	-	-
PAX6-024	ENST00000534353	540	No protein	Processed transcript	-	-	-
PAX6-019	ENST00000533333	6173	No protein	Retained intron	-	-	-
PAX6-006	ENST00000470027	2842	No protein	Retained intron	-	-	-
PAX6-007	ENST00000494377	2460	No protein	Retained intron	-	-	-
PAX6-010	ENST00000464174	979	No protein	Retained intron	-	-	-
PAX6-017	ENST00000474783	702	No protein	Retained intron	-	-	-
PAX6-030	ENST00000532916	627	No protein	Retained intron	-	-	-
PAX6-026	ENST00000534390	578	No protein	Retained intron	-	-	-
PAX6-023	ENST00000532175	524	No protein	Retained intron	-	-	-
PAX6-022	ENST00000527769	487	No protein	Retained intron	-	-	-

Use the link labelled **View ENSEMBL**:. A view of the region of **Chromosome 11** reasonably similar to those you have already considered will leap forth. As before, the exons for each transcript are represented by blobs (filled for coding, empty for **UTR** regions). Introns are represented by wiggly lines joining the blobs. The colour scheme we might consider later. The main thing to notice now is that there are more than **20** transcripts represented here! **31** to be precise, as it says near the top of the page in tiny letters. For a clearer view, click on the Show transcript table link.

Indeed **31**. But see that not all are protein coding. Part of the difference in transcript count is due to the fact that **Ensembl** include non-coding transcripts, whereas the **NCBI** list only transcripts that code for proteins.

Also, and more importantly, **Ensembl** and **MapViewer** use different strategies to predict transcripts (and just about everything else!). Both use database searches in roughly the manner described above, but clearly **Ensembl** have less blind faith is the accuracy of the **RefSeq** mRNAs. As you can see, in the **RefSeq** column of the table, the matches with **RefSeq** mRNAs are taken into consideration. However, they do not appear to be considered individually sufficient to define a transcript. Some transcripts in the **Ensembl** table reference more then one **RefSeq** mRNA. The lesser quality **XM_** mRNAs are ignored. Not all **11** better quality are used (just **1** ignored). Counting just the protein coding transcripts predicted by **Ensembl**, I make it **15**.

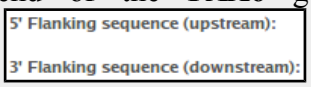
We could go on, other sources (not necessarily **Genome Databases**) would count differently again. Perhaps the best answer to the question "How many transcripts are there for the **PAX6** gene" is "**Several**".

1 The chaps at the **NCBI** have just told me that there is no longer any **XM_** mRNAs for **PAX6** due to an update of **RefSeq**. The difference we see are due to the **MapViewer** view being slightly out of date. I leave things as they are, but I have no idea what you will see by the end of July.

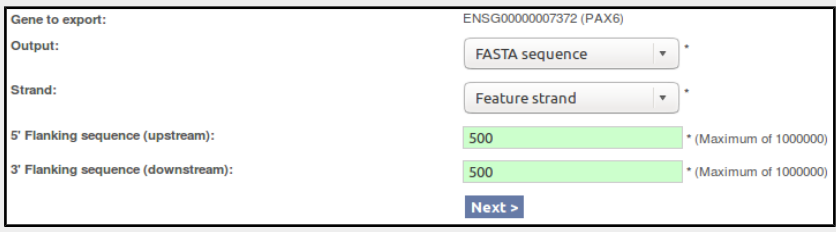
Before leaving **Ensembl**, it would be good to save the genomic sequence of this region for analysis later on.

To do this, first click on the Export data link on the left hand side of the page.

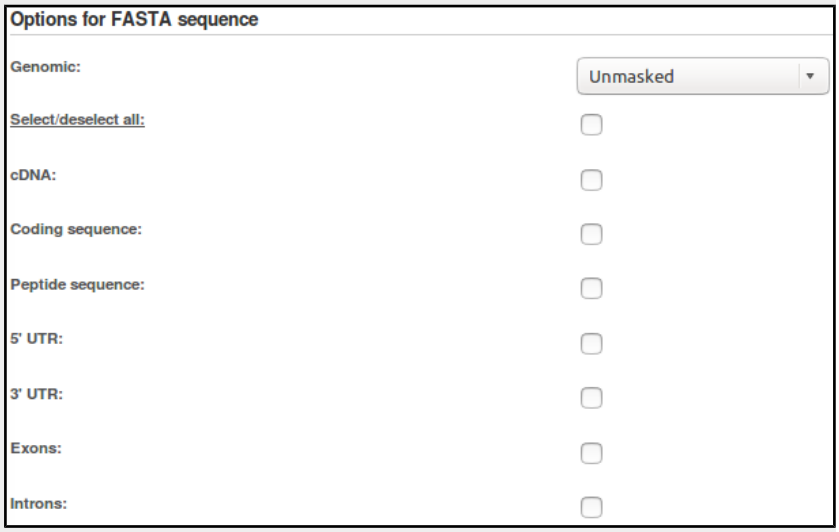
Ask for **500** base pairs of extra sequence at either end of the **PAX6** gene. That is, set both



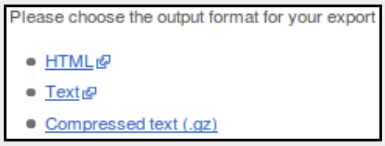
to **500**.



Deselect all the extra **PAX6** related sequences on offer. You just want the one genomic sequence for the entire **PAX6** region.



Click on the Next > button.



Choose **Text** as the **output format** for the sequence to be saved.

In you browser you should now have the genomic region of the **PAX6** gene, with **500** base pairs of flanking sequence on either end, in **FASTA** format.

Do whatever it takes to download this to a file called:

`pax6_genomic.fasta`

on your **Desktop**. If you end up with a big blank bit at the top of your file, as I did, it might be nice (but not essential) to delete it.

The next question might be "How many 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! I conclude the **Ensembl** predicts a minimum of **7 isoforms**. Most are either **422** or **436** amino acids long. Some of the others might cause a raised eyebrow or two, especially the one that is **2** amino acids long? But, who are we to question! **At least 7** is the informal **Ensembl** total.

Click your way back to the **NCBI PAX6** gene entry. Next I would like to discover the number of protein products (**isoforms**) that the **NCBI** predicts. This view makes this simple question clumsy to answer as the protein products of each transcript are reported separately, even when they are identical (as does the **Ensembl** list)???

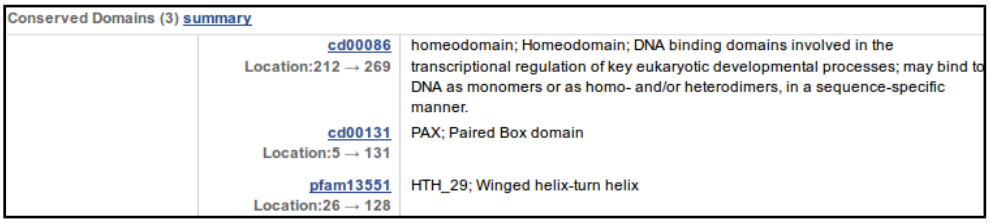
However, it can be done. Go just over half way down the page to the **mRNA and Protein(s)** section. Then skim down the entries for every transcript (just the **11** "good" **NM_** ones here) and check the different isoform names.

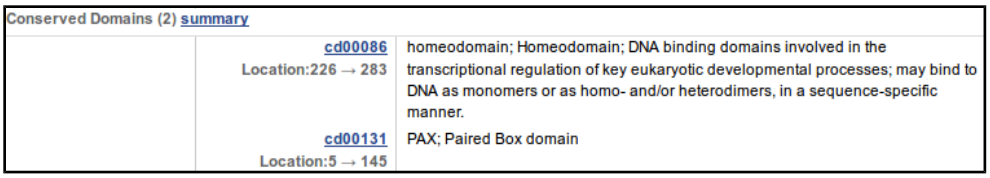
01 - NM_000280.4 → NP_000271.1 paired box protein Pax-6 isoform a
02 - NM_001127612.1 → NP_001121084.1 paired box protein Pax-6 isoform a
03- NM_001258462.1 → NP_001245391.1 paired box protein Pax-6 isoform b
04 - NM_001258463.1 → NP_001245392.1 paired box protein Pax-6 isoform b
05 - NM_001258464.1 → NP_001245393.1 paired box protein Pax-6 isoform a
06 - NM_001258465.1 → NP_001245394.1 paired box protein Pax-6 isoform a
07 - NM_001310158.1 → NP_001297087.1 paired box protein Pax-6 isoform b
08 - NM_001310159.1 → NP_001297088.1 paired box protein Pax-6 isoform c
09 - NM_001310160.1 → NP_001297089.1 paired box protein Pax-6 isoform d
10 - NM_001310161.1 → NP_001297090.1 paired box protein Pax-6 isoform d
11 - NM_001604.5 → NP_001595.2 paired box protein Pax-6 isoform b

I count **4**, imaginatively named **Isoform a**, **Isoform b**, **Isoform c** and **Isoform d**. One associated with each transcript description. Look carefully at the annotations and there is more information. In particular:

Description field: **isoform b** is also known as **isoform 5a**. Why this is important will become apparent in a page or so.

Conserved Domains.



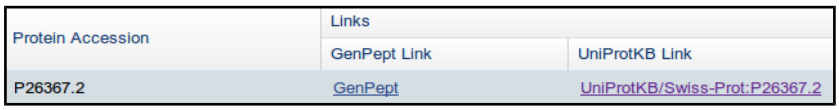


Isoform a has **3**, a **Paired Box Domain** at the beginning, a **Homoebox Domain** further along and a **Winged helix-turn-helix** coincident with most of the **Paired Box**.

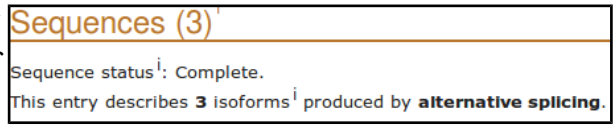
Isoform b lacks the **Winged helix-turn-helix**, which is the major DNA binding element of the **Paired Box**. Its omission might well imply a difference in function between these two isoforms?

UniprotKB offers yet another version of this story. Just for a for a few clicks, let us intrude into the **UniProt** session of your course.

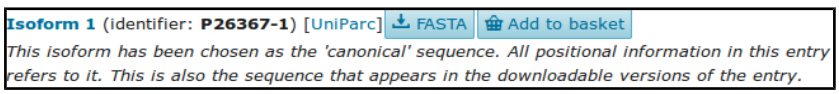
At the very bottom of the current page, you will find a link to **UniprotKB**. Use it[2].



Lo! the **PAX6** human protein as seen and understood by **UniProtKB**. Click on the Sequences (3) button on the left hand side of the page. **UniProtKB** declares **3** isoforms! At least, **3** that it is willing to admit to publicly.



There is **isoform 1**, also known as **isoform a** in America. Note that this is the "*canonical sequence*" for this protein. That is, this is the isoform that is used to represent this protein. The sequence(s) of all other isoform(s) are recorded as elements of the annotation.



2 If things are still as they are as I type, you need now to click another link to move the the latest **UniProtKB** version.

Also we have **Isoform 5a** (or **PAX6-5a**), also known as **isoform b** in America (where it also answers to **isoform 5a** when pressed). Note that the entry declares the sequence difference to be:

Isoform 5a (identifier: **P26367-2**) [UniParc] FASTA Add to basket
Also known as: Pax6-5a
The sequence of this isoform differs from the canonical sequence as follows:
47-47: Q → QTHADAKVQVLDNQN

47-47: Q → QTHADAKVQVLDNQN

Literally:

> "The amino acid at **position 47** in a **Q** in the canonical sequence. In **isoform 5a** this is replaced by the **15** amino acids **QTHADAKVQVLDNQN**".

More coherently this amounts to:

> "**isoform 5a** differs from the canonical **isoform 1** in that it has an insertion of **14** amino acids after the **47th** amino acid of the canonical protein".

It is significant to note that position **47** is right in the middle of the **Paired Box Domain** that occurs in both isoforms and the **Winged helix-turn-helix** that is specific to **isoform 1/a** (see above).

Finally **UniProtKB** proudly presents the somewhat ephemeral **isoform 3** (or **PAX6-5A,6*** for those who enjoy formality). But this one has no known sequence? Not much Bioinformatics can offer here methinks.

Isoform 3 (identifier: **P26367-3**)
Also known as: Pax6-5A,6*
Sequence is not available

So I hope you will agree that the **UniProtKB** count stands at a very modest **2**, plus a ghost.

To visualise the differences between the **2** isoforms with sequence, click on the Align button at the top of the **Sequences** section. After deep thought and much fumbling, **UniProtKB** will multiply align all the isoforms for you. As there are only **2** in this case, this will appear very similar to a **Pairwise** alignment. Highlight the **DNA binding** regions and the **Domain**s

I leave the interpretation of this splendid display to you.

Highlight
Annotation
- Sequence conflict
- Helix
- Beta strand
- Turn
- Chain
- Compositional bias
- DNA binding
- Domain
- Alternative sequence
- Natural variant

The extra **14** amino acids of **isoform 5a** are due to the inclusion of a tiny extra (**42** base pair) exon in some transcripts.

Can you see the evidence for this assertion in the two regional genomic maps of a few pages back?

Alignment

How to print an alignment in color

```
P26367    PAX6_HUMAN          1 MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ------------      47
P26367-2  PAX6_HUMAN          1 MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQ      60
                                ***************************************************

P26367    PAX6_HUMAN         48 -VSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRL     106
P26367-2  PAX6_HUMAN         61 NVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRL     120
                                ************************************************************

P26367    PAX6_HUMAN        107 LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGT     166
P26367-2  PAX6_HUMAN        121 LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGT     180
                                ************************************************************

P26367    PAX6_HUMAN        167 SVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE     226
P26367-2  PAX6_HUMAN        181 SVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE     240
                                ************************************************************

P26367    PAX6_HUMAN        227 KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIP     286
P26367-2  PAX6_HUMAN        241 KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIP     300
                                ************************************************************

P26367    PAX6_HUMAN        287 ISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPP     346
P26367-2  PAX6_HUMAN        301 ISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPP     360
                                ************************************************************

P26367    PAX6_HUMAN        347 VPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQ     406
P26367-2  PAX6_HUMAN        361 VPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQ     420
                                ************************************************************

P26367    PAX6_HUMAN        407 VPGSEPDMSQYWPRLQ                                              422
P26367-2  PAX6_HUMAN        421 VPGSEPDMSQYWPRLQ                                              436
                                ****************
```

We need to save a some protein sequences for future analysis. This is easiest from **UniProtKB** so now is good. To declare your intention to save the entire canonical version of the **PAX6** protein to a file, move back from your alignment. Move to the top of the page where you will find the bizarre invitation to Add to basket ? Just do it.

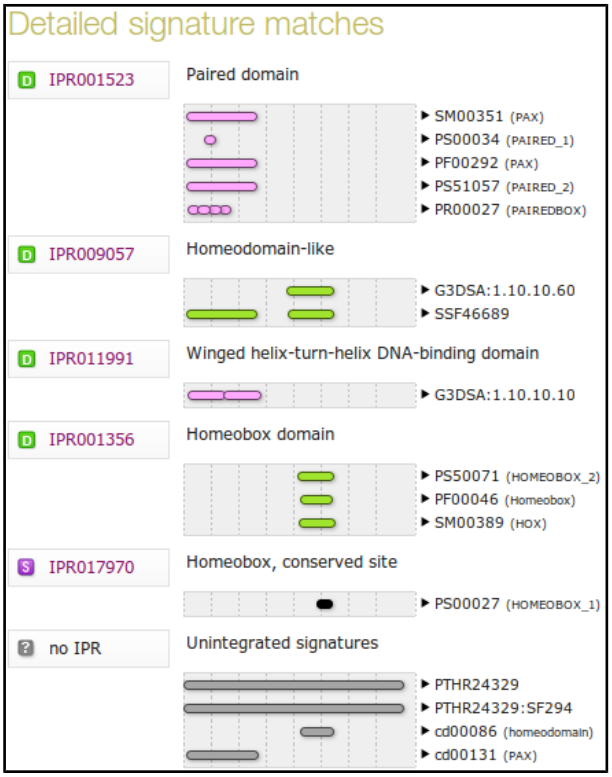
You also need to download the sequences of both domains is separate files, via your basket. First the **Paired Box**. Click the Family & Domains button on the left of the page. Then use the Add button adjacent to the **Paired** entry. Its now in your basket you will be ecstatic to know.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Domain	4 – 130	127	Paired PROSITE-ProRule annotation			Add BLAST

As they are so conveniently in view, take note of the **Compositional bias** features. They will be of interest when we look at database searching.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Compositional bias	131 – 209	79	Gln/Gly-rich			Add BLAST
Compositional bias	279 – 422	144	Pro/Ser/Thr-rich			Add BLAST

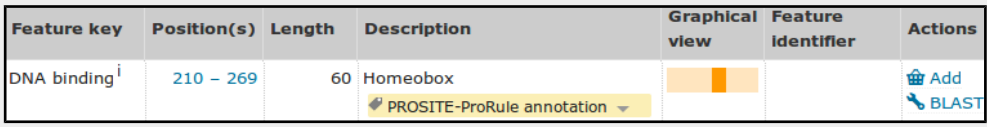
Given we are in the neighbourhood, slide a few inches down to the **Family and domain databases** section. Here is stored the results of comparing the **PAX6** protein with many of available **Domain/Motif Databases**, including those of the **Interpro** Consortium collectively. Are the results broadly as you might expect?



For the best summary, click on the [Graphical view] for the **Interpro** results. If the detail is not entirely transparent, hopefully there will be time to discuss this graphic at some point.

For now, I wish mostly to make the point that there is nothing difficult about this sort of analysis. You can easily produce exactly this result yourself. Maybe you will as part of this exercise?

Back to saving sequences for later! To get to the **Homeobox** domain, you need to click on the Function button on the left hand side of the page.



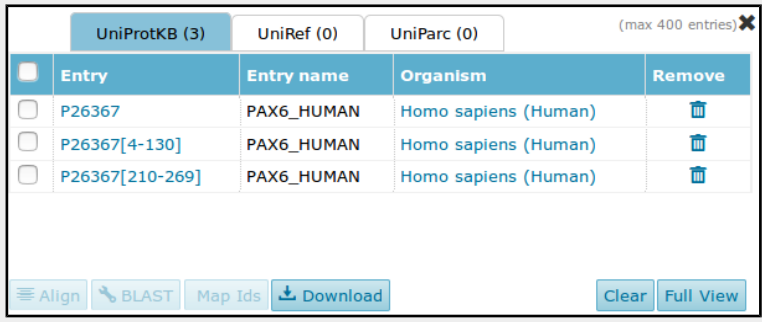
A valid question at this point might be "Why is the **Homeobox** domain a **DNA binding** feature, but the **Paired** domain is a **Domain** feature?" To which the answer is "*History, dear boy, history*" to paraphrase a disputed quote of dear Harold (Macmillan that is).

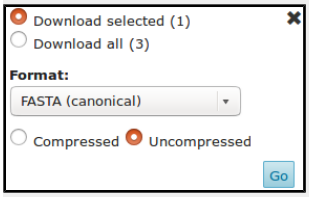
The fact both are **Domains**, and both are **DNA binding**. The illogicality of them being recorded in different places is accepted, however, to fix this early mistake now would not be trivial. So, we live with it. So doing, click on the appropriate Add button and head for the checkout desk (Good Grief! I am beginning to get used to this!).

Shimmy back to the top of the page. You should have Basket 3 things in your basket.

Click on the basket to view your booty.

For each of the 3 items in turn (not all at once or you get all sequences in one file), select and Download .





Each time ensure the download parameters are set to **Uncompressed** and **FASTA (canonical)**. Then click the Go button.

The next few steps, as before, are very browser/OS dependant. Just do whatever it takes to save the three sequences in files called, as appropriate:

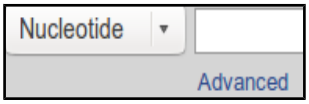
```
pax6_human.fasta
pax_domain.fasta
homeobox_domain.fasta
```

Now move back to America! Back to the **NCBI** view of the **PAX6** gene, before I get into more trouble with Klemens for intruding into your official **Uniprot** session! If you have any problem getting there … click here.

At the bottom of the page, there is a section called **Related sequences**. Click on the last entry, the mRNA called **AB209177.1** . You will be rewarded by a **GenBank** entry in **GenBank** format. Formats are tedious, but we will discuss them briefly at some point. You have already witnessed **FASTA** format. I expect we will bump into **EMBL** format at some point. The other **137** or so formats I suggest be ignored!

Can you see the official gene name **PAX6**, mentioned in this entry? The **Gene Name** field (where **PAX6** should most certainly be mentioned) is entirely missing! If you searched **GenBank** (or **EMBL** come to that) for this sequence using the most obvious search **Keyword**, that is **PAX6**, do you think you would find this **PAX6** mRNA? You clearly should! A case for more consistent annotation, as I feel sure Melanie will agree in the **Gene Ontology** session later.

Next, we search the nucleotide databases, by textual Keyword, for **PAX6** related sequences and down load one or two for investigation. To achieve this worthy goal, change the search space from **Gene** to **Nucleotide** and click on the Advanced search option button[3].



Then in the **Nucleotide Advanced Search Builder**, change All Fields to Title in the pull down menu associated with the first search field and type in the keywords:

chromosome 11

In the second search field, again change All Fields to Title and type in the keyword:

pax6

You are asking **Entrez** to search for all **Nucleotide** database entries that contain the terms **chromosome 11** and **pax6** in the section of their annotation intended to be a succinct brief description (I.e. **Title**) of the entry. Click on the Search button to start the search going.

There is just one matching entry which is

```
LOCUS       NG_008679              40170 bp    DNA     linear   PRI 11-AUG-2014
DEFINITION  Homo sapiens paired box 6 (PAX6), RefSeqGene on chromosome 11.
```

arrayed before you in **Genbank** format, very neat!! It was the **DEFINITION** line that you searched by selecting the **Field** value **Title**. I needed a few tries to get the right search to find just what was needed, and was a bit surprised at the simplicity and accuracy of the final search. You are looking at a **RefSeqGene** (a subset of the **RefSeq** database) entry. As such, it represents a genomic sequence for a "well-characterised gene", in this case **PAX6**.

Take a look at the **FEATURES** for this entry. You will see that there are **two** genes mentioned. **PAX6**, of course, and **ELP4** on the strand that is the complement of that represented here.

```
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:8620"
                /db_xref="MIM:607108"
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
                PAXNEB"
                /note="elongator acetyltransferase complex subunit 4"
                /db_xref="GeneID:26610"
                /db_xref="HGNC:HGNC:1171"
                /db_xref="MIM:606985"
```

```
join(16551..16560,20128..20258,21186..21401,22106..22271,
28174..28332,28848..28930,29160..29310,29409..29524,
32102..32252,32943..33028)
/gene="PAX6"
/gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
/note="isoform a is encoded by transcript variant 1;
paired box homeotic gene-6; oculorhombin; aniridia type II
protein"
/codon_start=1
/product="paired box protein Pax-6 isoform a"
/protein_id=" NP_000271.1 "
/db_xref="GI:4505615"
/db_xref="CCDS: CCDS31451.1 "
/db_xref="GeneID: 5080 "
/db_xref="HGNC: HGNC:8620 "
/db_xref="MIM: 607108 "
/translation="MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISR
ILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEI
RDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRP
GWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFT
QEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRR
QASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPS
FTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ"
```

At the top of your page, **Analyze this sequence** by clicking on the **Highlight Sequence Features** option. The **CoDing Sequence** (**CDS**) feature for **PAX6** is displayed for you by highlighting the relevant parts (the coding **exons**) of the sequence and displaying the **CDS** details including the DNA regions that need to be **join**ed to form the **CDS** and the **translation** of the **CDS**.

Use the controls at the bottom of your page to look at the other features of this entry (select feature **number** and then click on the **Feature** button).

What were the features that you found? _____

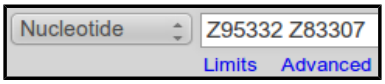
Why might you have expected more features than there were? _____

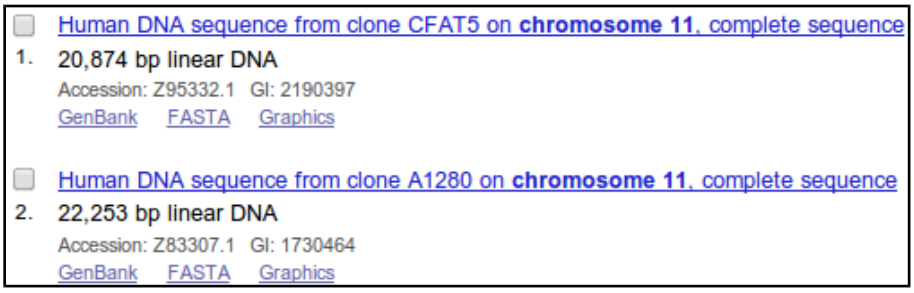
3 Check to see which database you are actually searching at this point. If the URL includes "gene", change it to "nuccore". This is a bug! I have reported it, it may or may not be fixed in time for your course.
Stop Press! Astoundingly, the boys at the **NCBI** claim this to be they intended things to be (although they admit the logic of my complaint)!! I am too shocked to edit my notes to conform to such imbecility! Edit the URL for now, I will implement a work around in the fullness of time.

```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff in
            collaboration with Isabel Hanson, David FitzPatrick. The reference
            sequence was derived from Z95332.1 and Z83307.1.
            This sequence is a reference standard in the RefSeqGene project.

PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
            1-18852             Z95332.1           2023-20874
            18853-40170         Z83307.1           105-21422
```

Take a look at the **COMMENT** and **PRIMARY** sections just above the **FEATURES**. This entry is suggested to be constructed from two sequences from **GenBank**. That is, the products of two sequencing projects.

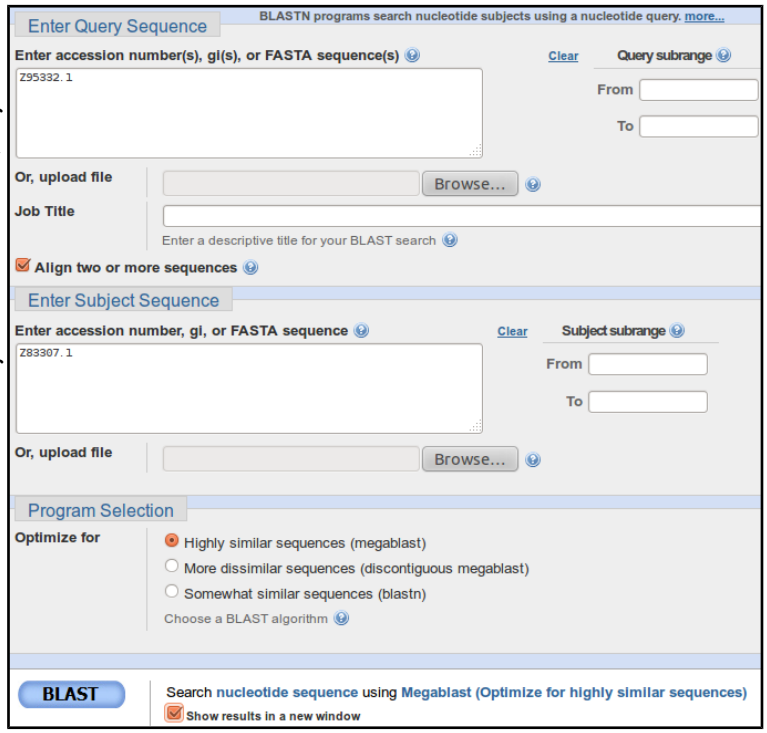
Take a quick look at the **GenBank** entries by entering their **ACCESSION** numbers into the **Search** box at the top of your page. Click on the Search button.





Lo and behold, the two **GenBank** entries are summoned forth. Take a look at one or both. Not particularly illuminating I think[4]. These are clones sequenced as part of the **H**uman **G**enome **P**roject (**HGP**). They served to cover regions of **Chromosome 11** and have little biological significance in themselves.

Move back to the list, as illustrated. Elect to **Analyze these sequences**, selecting from the extensive range of possibilities **Run BLAST**. We will look at **blast** properly later, the idea here is to simple prove that these two sequencing clones really do overlap in the fashion suggested by the evidence so far. So, elect to **Align two or more sequences**[5]. Cut and paste one of the sequencing clone **accession numbers** from the **Enter Query Sequence** box to the **Enter Subject Sequence** section of the form. Elect to **Show results in a new window**[6]. Firmly address the BLAST button.



Just one region of overlap should be identified.

```
Query  20771  GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  20830
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1      GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  60

Query  20831  TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  20874
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct  61     TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  104
```

How does the alignment you generated match up with the annotation of the original **RefSeq** entry you discovered? _

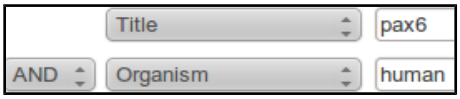
---

4 The annotation is very sparse which makes these entries very hard to find directly. The **EML-Bank** versions include some links to **Ensembl** codes. These would have been helpful but are not part of the official **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration (**INSDC**) annotation that should be consistent between **GenBank**, **E**uropean **N**ucleotide **A**rchive (**ENA**), which includes **EML-Bank**, and **D**NA **D**ata **B**ank of **J**apan (**DDBJ**).

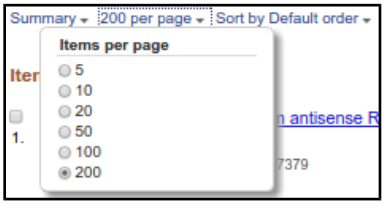
5 As opposed to comparing each of the two clones against an entire sequence database.

6 Just because its neater. In my, significantly less then humble, opinion anyway.

Now for an entirely new search. The easiest way to get a fresh start is to move back to your browser tab displaying the **GenBank** Search results, and then click on the **Advanced** option of the **Search** facility at the top of the page. You should arrive back at the **Nucleotide Advanced Search Builder** offering a fresh start.

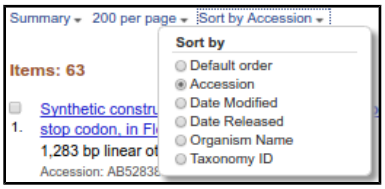


Set up a new search as illustrated and set it going. Ultimately simple this time. You have requested all **Human** sequences that are centrally associated with the gene **PAX6**.



A list of **50** or so sequences, all clearly claiming **PAX6** association and announcing their humanity loudly in Latin, will tumble forth.

You will have more hits than can be displayed in one go. Also, the hits are arranged in a "**Default**" order which has thus far defied all my attempts to associate with any definition of logic!

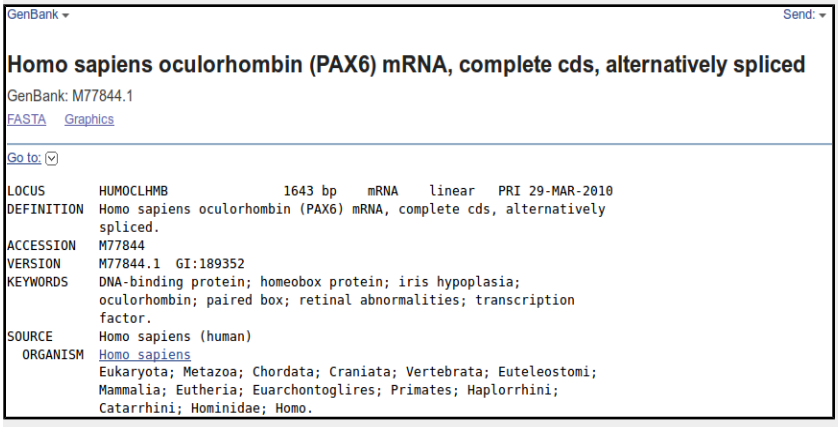


To deal with both of these issues, use the display control pull down menus at the top of your page to set the items **per page** to something big and the **Sort by** option to something sane.

The list shows matches between the terms entered and the ***annotation*** of DNA sequences. Not all relevant sequences will be present. For example, the **mRNA** with accession number **AB209177** was justifiably referenced in the **PAX6 Gene** entry but will not be in this list. **PAX6** appears nowhere in the entire annotation of **AB209177** let alone just its **DESCRIPTION** (or **Title**) field.

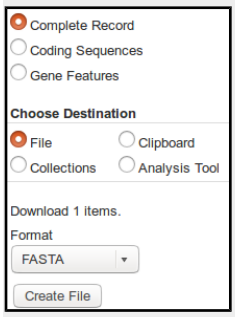
Move far down the list, you will come to the **RefSeq PAX6** mRNAs of a few pages back. Just before these entries is **M77844.1**. Save this one for later analysis. I choose **M77844.1** as it includes a few variations that will add interest. Select the target sequence.

You could now use the diminutive Send to: button which is near the bottom of your page to download all the selected sequences into a single file.



However, as there is only one sequence, and it would be so nice to be introduced properly before such intimacies as "downloading". Click on the link to the database entry to see it in all its **GenBank Format** glory.

The sequence is for analysis rather than decoration, so use the format menu at the top of the page (currently set **GenBank**), and ask for **FASTA** format.



Now click the tiny **Send:** button and **Choose Destination** to be **File**.

Strike the **Create File** button with a firm resolve. With irritating presumption, the choice of file name is made for you. Your sequence will be stored in a file named:

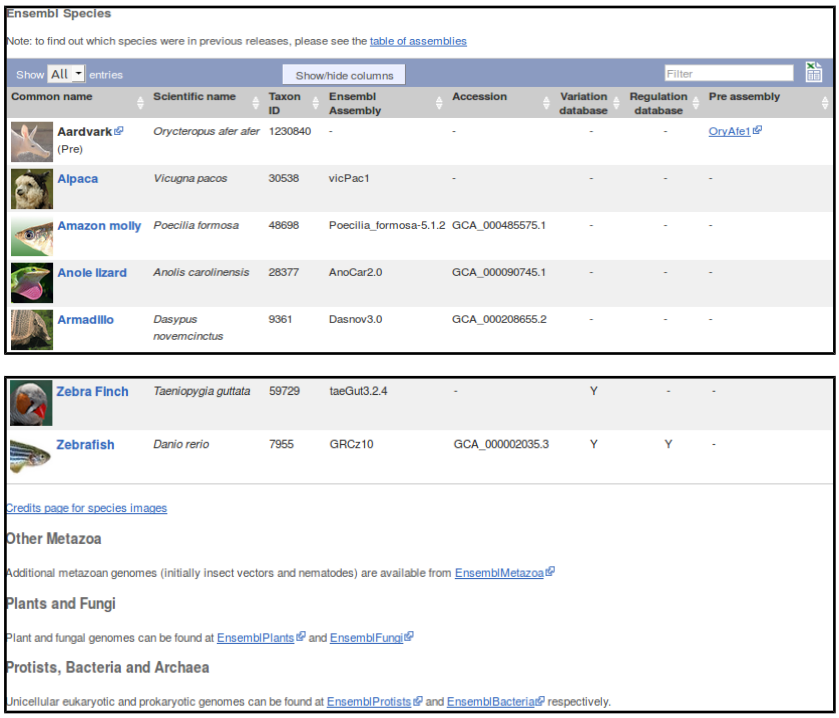
`sequence.fasta`

The **NCBI** is justifiably not famed for its understanding of poetry! Do whatever it takes to rename this file to be called:

`pax6_mrna.fasta`

Back to **Ensembl**. More with the objective of looking at more sources of information via **Ensembl** than becoming expert **Ensembl** users.

Go to the **Ensembl** home page (**`www.ensembl.org`**). Choose to **View full list of all Ensembl species** using the link just under the **Select a species** menu.

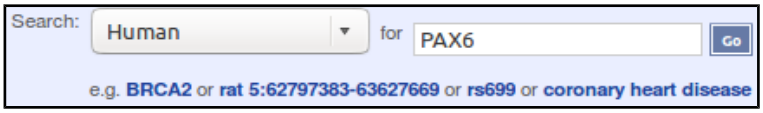


Note that **Ensembl** (and **MapViewer**, of course) offers far more than just the Human Genome.

In particular, note the links to **EnsemblPlants**, **EnsemblFungi**, **EnsemblBacteria etc. Ensembl** databases at the bottom of the list.

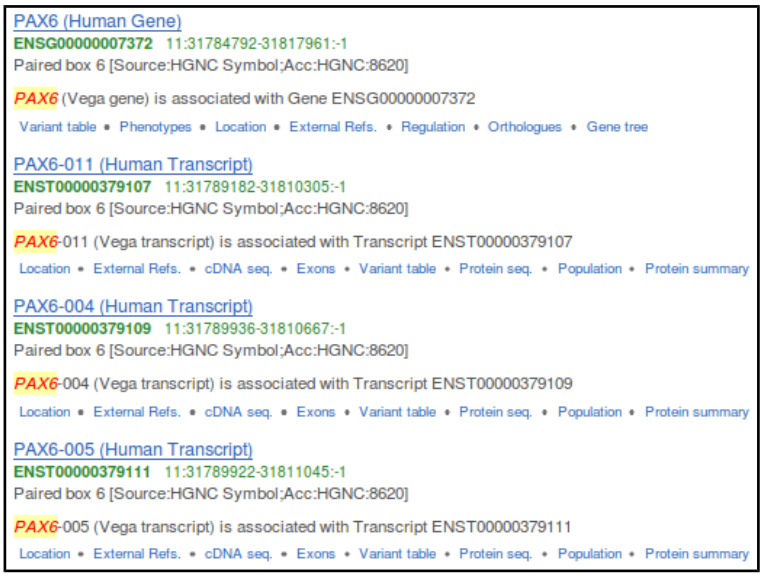
During this exercise, you will only look at the Human genome, by far the most fully developed. However, all the other **Ensembl** genomes are behind the same interface. The techniques required to examine the Human genome are broadly those required to examine any **Ensembl** genome.

Move back to the home page and go straight to the Human **PAX6** gene information by setting up the **Search** fields as shown and clicking the **Go** button boldly.



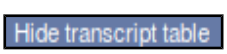
Its the target gene which is top of the hit list.

Click on the link to the **PAX6 (Human Gene)**.



You should recognise the view you now see. The list of transcripts and a view of the genomic region roughly similar to those offered by the **NCBI**.

There is much to investigate here, but maybe that should wait for a specialised **Ensembl** course. They are run regularly in Cambridge and elsewhere.

To make a bit more space, elect to 

.

Click on the **Othologues** link in the left hand side of your browser page. Take a look at some of the alignments providing support for the homologous relations. The protein alignments are the more informative.

Using the evidence of the protein alignments, which **PAX6** isoforms do the fruitfly orthologues most resemble? _____

Once your curiosity is completely sated, click on the **Paralogues** link. **Paralogues** here should match those reported by **GeneCards** as **GeneCards** obtains its **Paralogues** report from **Ensembl**.

Try some • Alignment (protein) links to view an alignments between a **PAX6** isoform and its **paralogues**.

What region of the paralogues seem to be best conserved? Does this surprise you? If not, why not?

How many **PAX** protein paralogues are there for human? Suggest a prettier naming scheme than **PAX1**, **PAX2**, **...**

Next look at some transcript specific features as they are recorded in **Ensembl**. To do this, one must first select a transcript, so Show transcript table once more and select **ENST00000419022**. Again, to make a bit more space, why not Hide transcript table away.

Now click the **Exons** link (from **Transcript-based displays → Sequence**). **Exons**, **Introns** and **Variations** within **Exons** are clearly displayed.

Intron 2-3	31,810,827	31,806,926			3,902	gtgagtccgcttctttcttctcgct.....ttttttctccttctgttttgtcttag
ENSE00001098662	31,806,925	31,806,849	-	-	77	GGGGAAGACTTTAACTAGGGGCGCGCAGATGTGTGAGGCCTTTTATTGTGAGAGTGGACA GACATCCGAGATTTCAG
Intron 3-4	31,806,848	31,806,463			386	gcaagttctgtggtggctgctttgg.....ttaactcctatttttcttgctaacag
ENSE00002523992	31,806,462	31,806,402	-	1	61	AGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAAC A
Intron 4-5	31,806,401	31,802,835			3,567	gtaagtgcctctggtctttctggga.....tttcctctcctctccctttcctcag
ENSE00003602163	31,802,834	31,802,704	1	0	131	GTCACAGCGGAGTGAATGAGCTCGGTGGTGTCTTTGTCAACGGGCGGCCACTGCCGGACT CCAGCCGGCAGAAGATTGTAGAGCTAGCTCACAGCGGGGCCCGGCCGTGCGACATTTCCC GAATTCTGCAG
Intron 5-6	31,802,703	31,801,913			791	gtgatcctcccggcgccgccccact.....ttgaaggtatatttttgtgttatag
ENSE00003512677	31,801,912	31,801,871	0	0	42	ACCCATGCAGATGCAAAAGTCCAAGTGCTGGACAATCAAAAC
Intron 6-7	31,801,870	31,801,777			94	gtaagcttgtcattgtttaatgcat.....ttttctgtccacttcccctatgcag
ENSE00003523920	31,801,776	31,801,561	0	0	216	GTGTGCAACGGATGTCTGAGTAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGA CCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCCAGAAGTTGTAAGCAAAATA GCCCAGTATAAGCGGGAGTGCCCGTCCATCTTTGCTTGGGAAATCCGAGACAGATTACTG TCGAGGGGGTCTGTACCAACGATAACATACCAAGC

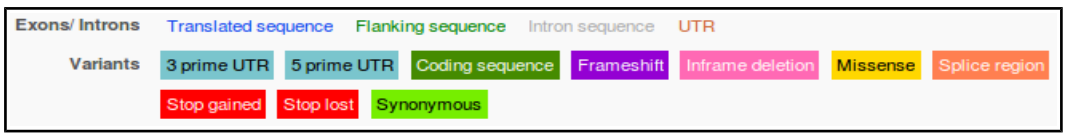
What are the first two bases and what are the last two bases of nearly every intron? _____

How long is the sixth exon and why would this concur with your expectations? _____

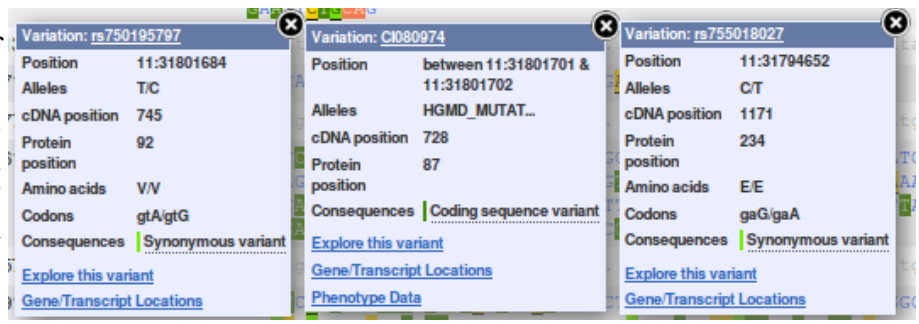
Explain the **Start Phase** and **End Phase** columns? _____

_____

Click on some of the colourful variation locations. The colours are explained in the legend at the top of the display.



The variations come from a number of variation databases, including **dbSNP**. The **dbSNP** entries are those whose names begin with "**rs**". **dbSNP** can be investigated directly at the **NCBI**, of course, but it very handy to have all the variation information built into **Genome Databases** such as **Ensembl**.



Click on the **Domains & features** link (from **Transcript-based displays → Protein Information**).

Domain source	Start	End	Description	Accession	InterPro
PANTHER	1	434	-	PTHR24329	-
PANTHER	1	434	-	PTHR24329:SF294	-
Gene3D	7	86	-	1.10.10.10	-
Gene3D	87	150	-	1.10.10.10	-
Gene3D	201	284	-	1.10.10.60	-
Prosite_profiles	222	282	Homeobox domain	PS50071	IPR001356 [Display all genes with this domain]
Smart	224	286	Homeobox domain	SM00389	IPR001356 [Display all genes with this domain]
Pfam	226	281	Homeobox domain	PF00046	IPR001356 [Display all genes with this domain]
Prosite_patterns	257	280	Homeobox, conserved site	PS00027	IPR017970 [Display all genes with this domain]
Superfamily	6	143	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Superfamily	205	283	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Pfam	4	142	Paired domain	PF00292	IPR001523 [Display all genes with this domain]
Smart	4	142	Paired domain	SM00351	IPR001523 [Display all genes with this domain]
Prosite_profiles	4	144	Paired domain	PS51057	IPR001523 [Display all genes with this domain]
PRINTS	8	23	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	26	44	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	60	77	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	78	95	Paired domain	PR00027	IPR001523 [Display all genes with this domain]

Are you shocked and dismayed that the precise location of the **PAX6** Homeobox domain is not identically predicted by the **SMART** and **Pfam Domain Databases**? If not, why not?

How is that all the predictions, of different domain databases, for a **Paired domain** have the same **Interpro** identifier?

Why does **Prints** appear to predict four **Paired_domains**?_____

Click on the link to the **SMART** entry for the **Paired domain** (**SM00351**).

Here you will find (quoted from **Interpro**) a **Description** of a **Paired domain**.

Where would you expect a **Paired domain** to occur in a protein?

What expectations do you have concerning what typically follows a **Paired domain**?

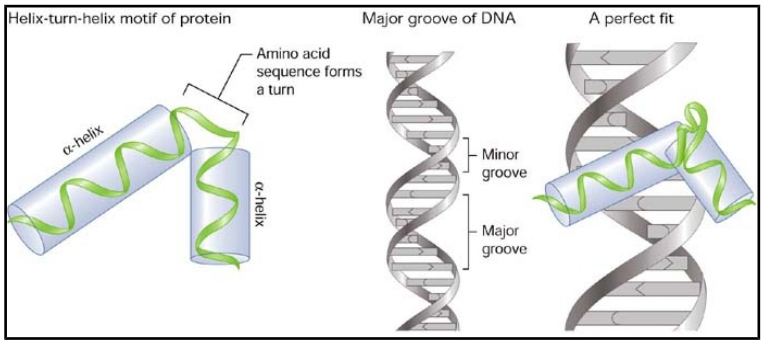
The paired domain is an approximately 126 amino acid DNA-binding domain, which is found in eukaryotic transcription regulatory proteins involved in embryogenesis. The domain was originally described as the 'paired box' in the Drosophila protein paired (prd) [(PUBMED:2877747), (PUBMED:3123319)]. The paired domain is generally located in the N-terminal part. An octapeptide [(PUBMED:10811620)] and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C terminus.

Paired domain proteins can function as transcription repressors or activators. The paired domain contains three subdomains, which show functional differences in DNA-binding. The crystal structures of prd and Pax proteins show that the DNA-bound paired domain is bipartite, consisting of an N-terminal subdomain (PAI or NTD) and a C-terminal subdomain (RED or CTD), connected by a linker. PAI and RED each form a three-helical fold, with the most C-terminal helices comprising a helix-turn-helix (HTH) motif that binds the DNA major groove. In addition, the PAI subdomain encompasses an N-terminal beta-turn and beta-hairpin, also named 'wing', participating in DNA-binding. The linker can bind into the DNA minor groove. Different Pax proteins and their alternatively spliced isoforms use different (sub)domains for DNA-binding to mediate the specificity of sequence recognition [(PUBMED:11103953), (PUBMED:15148315)].

The reason for these two questions will become apparent later.

The second paragraph of the **Description** claims, in gross summary:

- A paired domain is a DNA binding domain that has 2 binding regions each of which involves a helical triplet
- The second and third helices of each helical triplet form **helix-turn-helix** (**HTH**) motifs
- The **HTH** regions bind the DNA major groove[7]



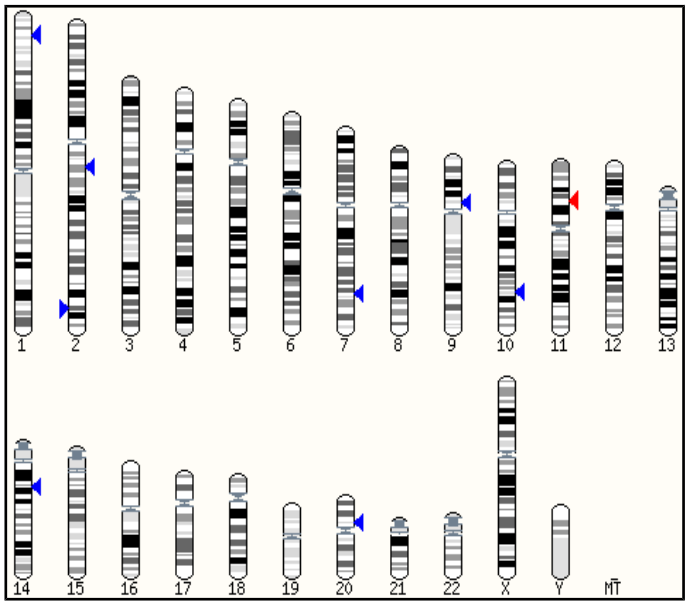
- The first helical triplet is preceded by a **β-turn** and **β-hairpin** ("wing") that participate in the DNA binding

- The linker region between the two helical triplets can bind the **DNA minor groove**

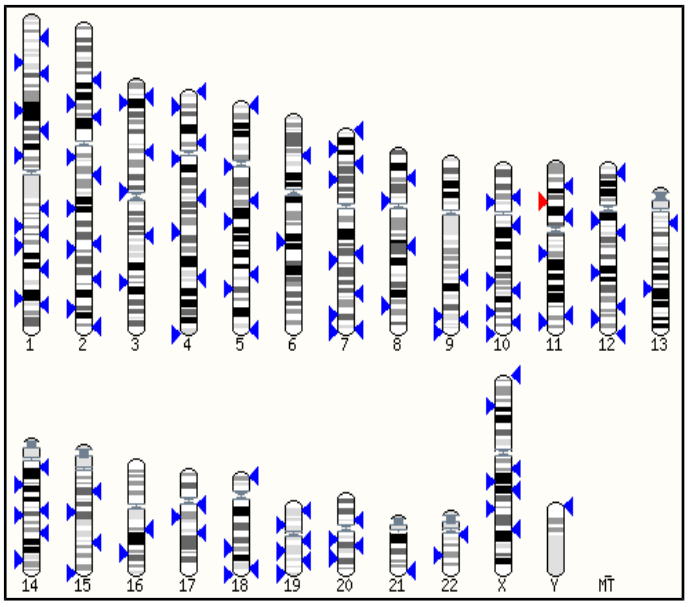
Bear this in mind when looking at the 3D structures a couple of pages on.

Click on **Display all genes with this domain** for the **Paired domaim** and **Homeobox domain InterPro** families. The locations of all genes including each domain will be displayed graphically and textually. **PAX6** is shown in red.

Paired domain - IPR001523



Homeobox domain - IPR001356



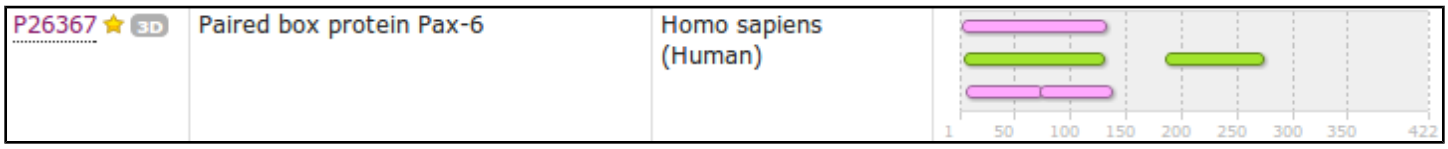
Which domain, **Paired domain** or **Homeobox domain** is more common in humans? _____

How many human **PAX** genes are there? _____

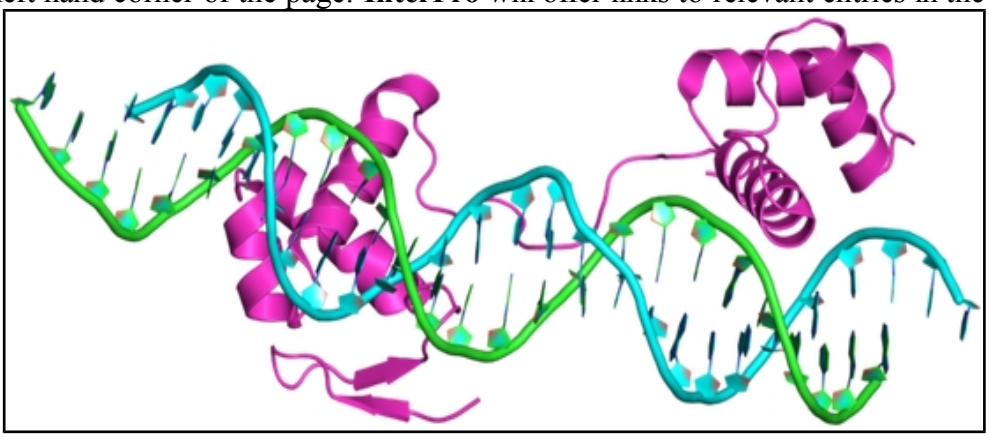
Are all the **PAX** genes on **Chromosome 11**? _____

7 If, like me, you have conceptual problems with major and minor groves. Try this **animated picture**. Helped me at least. As did the image above.

Move back to the **Domains & features** display. Link to the **InterPro** database entry for **Paired domain**, also know as **IPR001523**. Here you will find the origins of the **SMART** documentation. Click on the **Proteins matched** link. You will see listed a number of representations of proteins that, according to **InterPro**, include a **Paired domain**. Amongst these will be the human **PAX6** protein, also known as **P26367**[8]. There are links provided to entries in a number of relevant databases for each listed protein.

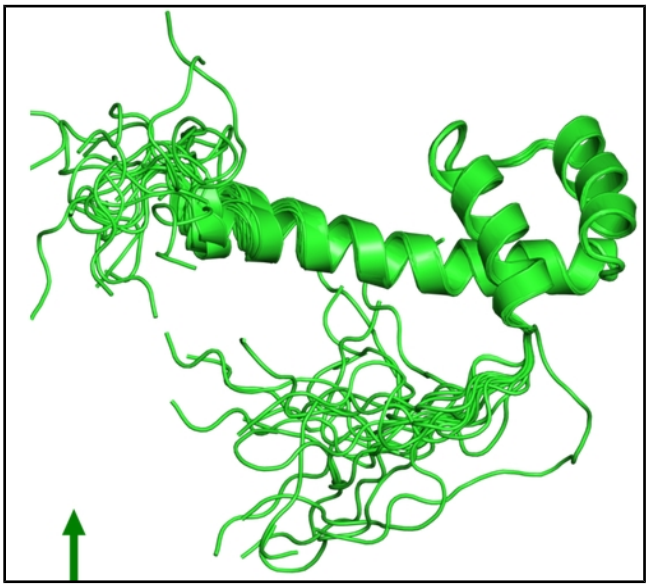


Click on the **Structures** link in the top left hand corner of the page. **InterPro** will offer links to relevant entries in the **PDBe**, **SCOP** and **CATH**[9] databases. Click on the link to the **6pax** entry in the **PDBe** database. You will arrive at the entry for **6pax** in **PDBe**, the European version of **PDB** maintained at the **EBI**. Views of this structure are offered on the right hand side of the page. Click on the largest image which shows the paired box protein domain binding DNA rather beautifully. Once you have admired this image sufficiently, move back to the **6PAX PDBe** entry. From the Quick links on the right of the page, select the **3D Visualisation** option.



The **SMART** documentation you read earlier suggested two paired box subdomains, each of which " ... form a three-helical fold, with the most C-terminal helices comprising a **helix-turn-helix** (**HTH**) motif that binds the **DNA major groove**". Move your image around to confirm this assertion.

The same **SMART** documentation claims the subdomain nearer the N terminal " ... encompasses an N-terminal **beta-turn** and **beta-hairpin**, also named '**wing**', participating in DNA-binding. The linker can bind into the **DNA minor groove**". Manipulate you image to investigate the veracity of these assertions.



Once you have seen all there is to see of **6PAX**, move back to the **Ensembl Domains & features** display. Try the same tricks with the **InterPro Homeobox domain**. This time, it is difficult to find **P26367** in the huge list[10] **Proteins matched**, but you do not need to in order to link to the **Stuctures**. There are many more structures to choose from this time. I suggest you go for **2cue**. You have to imagine the DNA this time.

Can you explain the strangely frayed ends displayed in some of the representations of the **2cue** 3D structure? _____

_____

8 Third from the bottom of the first page, last time I counted.
9 **PDB** is the main database of **3D** protein structures. **SCOP** and **CATH** are also **3D** structure related databases.
10 If you really wanted to, the best approach is to search for **P26367** in the search box at the top of the page and then look for the **Homeobox domain** entry in the **Detailed signature matches** list.

To end, a gesture towards demonstrating that you could quite easily have computed most of the information you have been accessing, ready packed, from various databases. There are many way this objective could be achieved, I choose to search for the features of the **PAX6** protein.

As has been discovered from several information sources, the **PAX6** human protein has two DNA binding domains. A paired box at the **N terminal** and a homeobox a little further along. Both of the domains include **H**elix-**T**urn-**H**elix (**HTH**) motifs. In this exercise, you will investigate how you might discover these domains and motifs using the various freely available domain databases (discussed previously) and other feature prediction programs. Clearly, this is superfluous for this particularly, well documented protein, but a valuable option in other circumstances.

One approach would be to consider each relevant domain database in turn. Each major domain database has its own Home web site and customised software to take **Query** protein sequences, compare those sequences with domain representations (typically based on **H**idden **M**arkov **M**odels) and to report convincing matches. This would work, but would be tedious as there are many viable databases to consider. It would be dangerous to rely on too few of the databases available as none is perfect. You need a consensus prediction to be sure you miss nothing.

Also, you would need to know which databases are particularly appropriate for each domain you considered might be present. All databases cannot be optimised for all types of domain (for example, the **SMART** database specialises in domains that occur in signalling proteins).

So, let us not search individual domain databases in the main part of these exercises. Instead, I offer a supplementary exercise investigating a representative selection of the available searches. I selected the **Prosite**, **Pfam** and **PRINTS** domain databases, If you do this exercises, consider particularly the **PRINTS** section. It illustrates how and why **PRINTS** just fails to see one of the two domains (as you already discovered when looking at **UniProt**).

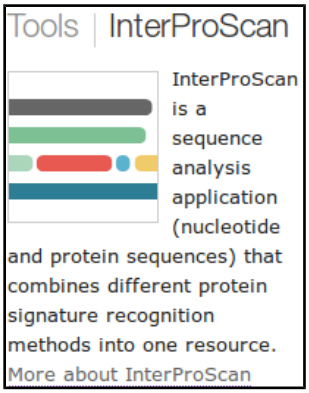
Here, use just **Interpro** to do the whole job. **Interpro** will search for all domains using the appropriate domain databases, thus removing the tedium of interrogating a miscellany of domain searching resources individually.

defines protein families according to the way that proteins match elements of a wide range of protein family databases, including all those we have discussed thus far. **Interpro** provides a search tool that will search all or any of the major protein family databases and assign **Interpro** family associations to the query protein(s) accordingly. To have a look at some of the possibilities offered by **Interpro**, Go to:

`http://www.ebi.ac.uk/interpro/`

If you were to enter the **PAX6** human protein into the obvious place on the **InterPro** home page, you would produce almost exactly the results you saw many pages back, when you were looking at **GeneCards**. Do this if you have the time and inclination.

By implication, **InterPro** offers a fuller experience via the **InterProScan** search tool. Other than the opportunity not to search **ALL** the domain databases, and having the results arranged slightly differently, I am unsure what the extra effort brings? Never mind, there are many things of which I am unsure, so, from the **InterPro** Home page ...

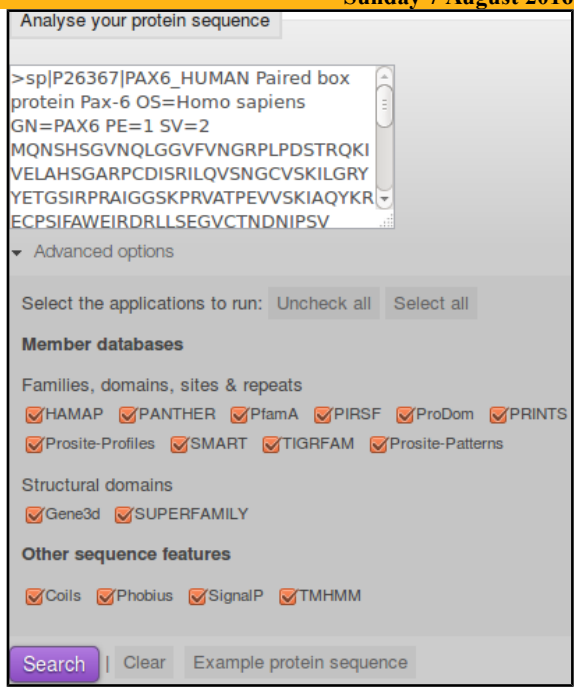


Select the **InterProScan** link. Here you will be offered the opportunity to download the **InterProScan** program.

I am not sure this is too useful an offer for most? But it is there.

For now, chose the online **Sequence search**.

You will arrive at a page that looks very similar to that from which you started, as far as the offer to run a domain search is concerned? Except! We now have **Advanced options**. Click on the **Advanced options**.



The **Advanced options** only allow you to choose which databases you wish to search and which feature prediction programs you wish to run. The default is to use all the databases and to run all the predictor programs. I struggle to imagine an occasion I would want to save the **EBI** servers a few cycles by considering which options to deselect, but it so nice to know I could if I wished to.

In passing, the offer to run the feature predictor programs in the **Other sequence features** section is relatively new. Of course, all these programs could be run individually from their home websites (follow the links behind the program names), in the same way as the domain databases can be searched individually. **Interpro** just aims to make thing easy for the user. The programs currently offered are:

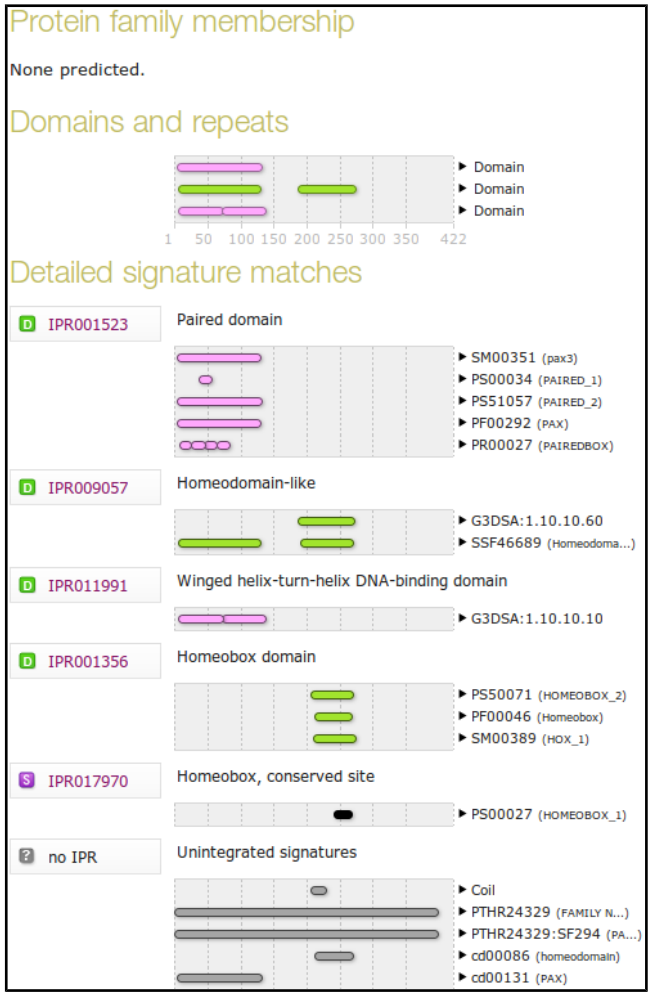
- **Coils** is a program for predicting **coiled coils**.
- **Phobius** & **TMHMM** are programs to predict **Transmembrane regions** (essentially **hydrophobic**, **uncharged** regions). There is no reason to expect any **Transmembrane regions** in this protein.
- **SignalP** predicts the presence and location of **signal peptide cleavage sites** in amino acid sequences from different organisms. I am pretty certain that there is no reason to expect signal peptides in this protein.

Do you think it a good idea for **Interpro** to offer feature prediction programs as well as domain database searches?__

Paste the human **PAX6** sequence into the patiently waiting box (from the file you made earlier called **pax6_human.fasta**). Accept the "***do everything***" default. Click on the **Search** button.

After several moments of deep thought, filtering and validating, you will be presented with a table of results looking very much like the one your saw earlier when looking around **UniProtKB**.

There is, however, one significant difference. In the **Unintegrated signatures** section, you will see that a **coiled coil** has been detected by the program **Coils**. This was not included in the **UniProtKB** information, maybe as **Interpro** has only recently included analysis using **Coils? UniProtKB** might catch up next time it is updated.



Do you think the Coil prediction might be correct?_____

Notice that **Interpro** assigns both the **PAX** domain and the **Homeobox** domain of human **PAX6** to the **Interpro** family **Homeodomain-like**. Both of these associations are based on the hit behind the link **SSF46689**.

SCOP classification

Root:	SCOP hierarchy in SUPERFAMILY [SCOP 0] (11)
Class:	All alpha proteins [SCOP 46456] (284)
Fold:	DNA/RNA-binding 3-helical bundle [SCOP 46688] (14)
Superfamily:	**Homeodomain-like** [SCOP 46689] (19)
Families:	Homeodomain [SCOP 46690] (40)
	Recombinase DNA-binding domain [SCOP 46728] (5)
	Myb/SANT domain [SCOP 46739] (15)
	SLIDE domain [SCOP 100998]
	GARP response regulators [SCOP 81683]
	DNA-binding domain of telomeric protein [SCOP 46745] (2)
	Paired domain [SCOP 46748] (3)

Follow this link and you will see it leads to the **Homeodomain-like superfamily** of the Superfamily 1.75 HMM library and genome assignments server database that specialises in very general (**SCOP**[11] **superfamily** level) protein classifications. One **Superfamily** entry will typically correspond to a number of more specific domain definitions in other domain databases. Here you can see that the **Superfamily** domain **Homeodomain-like** includes both the **Homeodomain** & **Paired domain Families**.

Return to your **Interpro** results page. The **Gene3D** database is similar to **superfamily** but based on the **CATH** database[12]. It suggests the two **HTH** motifs of the paired box are both **Winged helix-turn-helix**. The **HTH** in the **Homeobox domain** is not detected?

Why might you suppose **Interpro** predicts only **2** of the **3 helix-turn-helix domains** that might be expected?_____

_____

Follow the link to the **Interpro** family **Homeodomain-like** (**IPR009057**). Click on the button in the **Domain relationships** section to show the full list of **Homeodomain-like Interpro** domains.

Domain relationships

- **Homeodomain-like (IPR009057)**
  - DNA binding HTH domain, Fis-type (IPR002197)
  - DNA binding HTH domain, AraC-type (IPR018060)
  - DNA binding HTH domain, Psq-type (IPR007889)
  - DNA-binding HTH domain, TetR-type (IPR001647)
  - HTH CenpB-type DNA-binding domain (IPR006600)
  - Homeo-prospero domain (IPR023082)
  - Homeobox domain (IPR001356)
  - Homeodomain, ZF-HD class (IPR006455)
  - Homeodomain, phBC6A51-type (IPR024978)
  - Mor transcription activator (IPR014875)
  - Rap1 Myb domain (IPR015010)
  - Resolvase, HTH domain (IPR006120)
  - SANT/Myb domain (IPR001005)
  - SLIDE domain (IPR015195)
  - SWIRM domain (IPR007526)
  - Transposase IS30-like HTH domain (IPR025246)
  - Transposase, Synechocystis PCC 6803 (IPR002622)
  - TyrR family, helix-turn-helix domain (IPR030828)

Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

- **GENE3D**
  - G3DSA:1.10.10.60 (G3DSA:1.10.10.60)
- **SUPERFAMILY**
  - SSF46689 (SSF46689)

Note also the **Contributing signatures** in the top right hand corner of the page. Here is listed the domain databases that are searched to determine the presence of an **Interpro Homeodomain-like** domain.

Essentially, if **Gene3D** finds a match with its **Demineralisation** domain and/or **Superfamily** finds a match with its **Homeodomain-like** domain, then **Interpro** acknowledges a match with its **Homeodomain-like** domain (**IPR009057**).

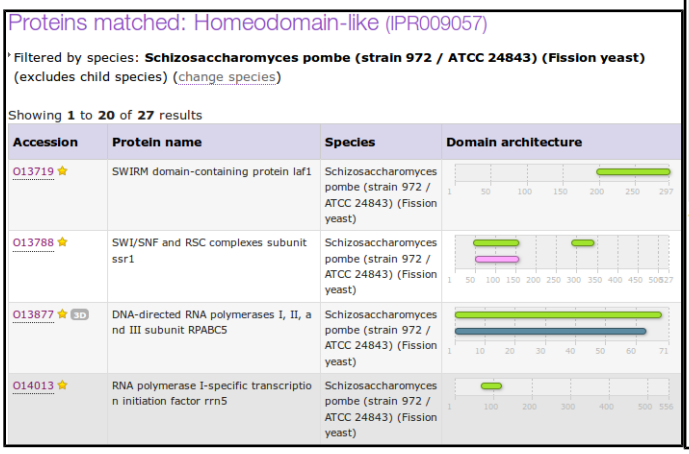
None of the other domain databases **Interpro** searches are used to determine membership of (**IPR009057**).

11 **S**tructural **C**lassification **O**f **P**roteins.

12 **CATH** is similar to **SCOP** in that it is another Structural classification database.

To obtain an impression of how widely spread throughout nature is this domain. Click on the **Species** button on the left hand side of the page.

As you can see, this is a very popular domain. You can make this list enormous by injudicious employment of the expansion buttons. Why not? It amused me for a few moments anyway.

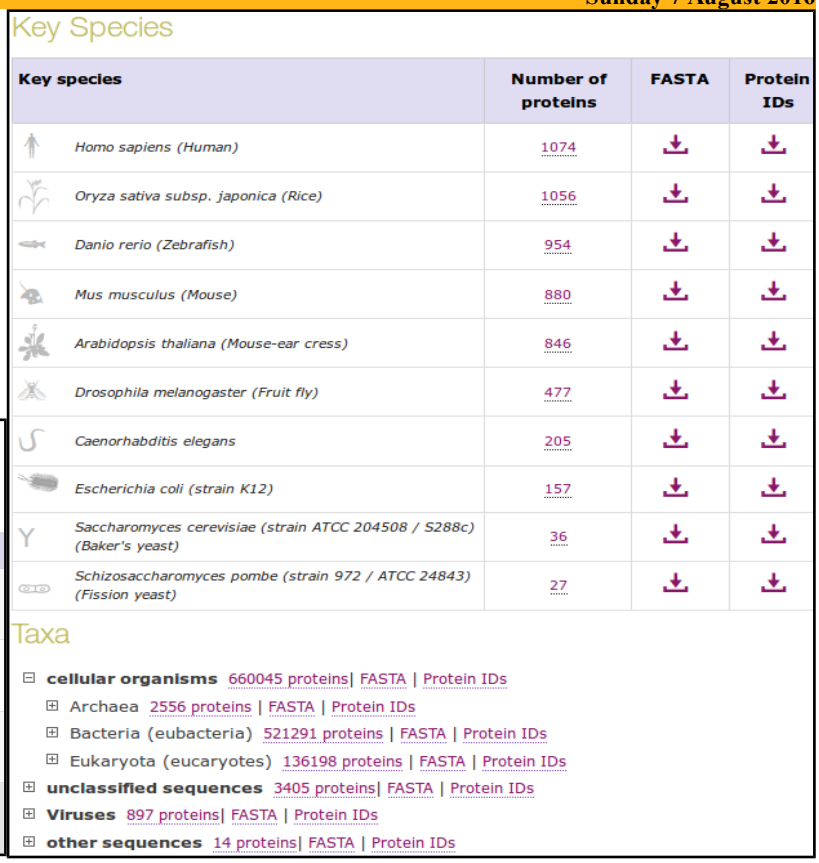


Proteins matched: Homeodomain-like (IPR009057)

Filtered by species: **Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)** (excludes child species) (change species)

Showing **1** to **20** of **27** results

Accession	Protein name	Species	Domain architecture
O13719	SWIRM domain-containing protein laf1	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	
O13788	SWI/SNF and RSC complexes subunit ssr1	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	
O13877 3D	DNA-directed RNA polymerases I, II, and III subunit RPABC5	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	
O14013	RNA polymerase I-specific transcription initiation factor rrn5	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	



Key Species

Key species	Number of proteins	FASTA	Protein IDs
Homo sapiens (Human)	1074		
Oryza sativa subsp. japonica (Rice)	1056		
Danio rerio (Zebrafish)	954		
Mus musculus (Mouse)	880		
Arabidopsis thaliana (Mouse-ear cress)	846		
Drosophila melanogaster (Fruit fly)	477		
Caenorhabditis elegans	205		
Escherichia coli (strain K12)	157		
Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	36		
Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	27		

Taxa

- **cellular organisms** 660045 proteins| FASTA | Protein IDs
  - Archaea 2556 proteins | FASTA | Protein IDs
  - Bacteria (eubacteria) 521291 proteins | FASTA | Protein IDs
  - Eukaryota (eucaryotes) 136198 proteins | FASTA | Protein IDs
- **unclassified sequences** 3405 proteins| FASTA | Protein IDs
- **Viruses** 897 proteins| FASTA | Protein IDs
- **other sequences** 14 proteins| FASTA | Protein IDs

By clicking on the appropriate button, you can get to either the protein sequences in **Fasta** format or list their accessions codes. Try a few, but be careful! It really does get you **ALL** the sequences, and that is often quite a lot, which can take time.

THE END

DPJ – 2016.07.17

Model Answers to Questions in the Instructions Text.

Notes:

For the most part, these "**Model Answers**" just provide the reactions/solutions I hoped you would work out for yourselves. However, sometime I have tried to offer a bit moer back ground and material for thought? Occasionally, I have rambled off into some rather self indulgent investigations that even I would not want to try and justify as pertenent to the objective of these exercises. I like to keep these meanders, as they help and entertain me, but I wish to warn you to only take regard of them if you are feeling particularly strong and have time to burn. Certainly not a good idea to indulge here during a time constrained course event!

Where things have got extreme, I am going to make two versions of the answer. One starting:

Summary:

Which has the answer with only a reasonably digestible volume of deep thought. Read this one.

The other will start:

Full Answer:

Beware of entering here! I do not hold back. Nothing complicated, but it will be long and full of pedantry.

This makes the Model answers section very big. **BUT**, it is not intended for printing or for reading serially, so I submit, being long and wordy does not matter. Feel free to disagree.

From your investigations using **Entrez**:

What were the features that you found?

Summary:

The first feature was the **Co**ding **S**equence (**CDS**) for a **PAX6** isoform.

The other three features were the coding sequences for three **ELP4** isoforms.

```
complement(39424..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_001275654.1"
/exception="annotated by transcript or proteomic data"
/note="isoform 2 is encoded by transcript variant 2;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=3
/product="elongator complex protein 4 isoform 2"
/protein_id=" NP_001275654.1 "
/db_xref="GI:570359562"
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEQIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEP
CSLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENG
GNSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGS
ERETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIEWVQDNYLRQE
RNIYPPGFSYLLKQKDSAWGEGSLQHSTFLMSFLAKATAFASRLVRHSEPLKQNGSGR
IRQAAGPRLWHDGRRQEAPGLLGIPP"
```

```
complement(39438..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_061913.3"
/exception="annotated by transcript or proteomic data"
/note="isoform 1 is encoded by transcript variant 1;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=1
/product="elongator complex protein 4 isoform 1"
/protein_id=" NP_061913.3 "
/db_xref="GI:91208435"
/db_xref="CCDS: CCDS7875.2 "
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEPC
SLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENGG
NSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGSE
RETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIERLHLPPDLSDT
VSRSSKMDLAESAKRLGPGCGMMAGGKKHLDF"
```

```
complement(39533..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_001275655.1"
/exception="annotated by transcript or proteomic data"
/note="isoform 3 is encoded by transcript variant 3;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=2
/product="elongator complex protein 4 isoform 3"
/protein_id=" NP_001275655.1 "
/db_xref="GI:570359564"
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEPC
SLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENGG
NSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGSE
RETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIEAGVQWHDLGSR
QPRLLGSSNSPASASLVAGITGAHHHTQLIFVFLVEMGFHHVGQAGLELLTSGDSSAS
ASQSAGITGMSYRARPRALYFKENKSKVGARQLLETREEHLSSRLLILTQAERLCMGR
RFFTAFHIFNELPCKGDCICLQTCQTQ"
```

Full Answer:

Note that only the final coding exon of **ELP4** is within this **RefSeq** sequence, which is defined as the genomic region for **PAX6**. This is clear from the length of the **translations** offered. The exon referenced is only long enough to code for just over **40** amino acids which is far short of any of the three isoform sequences offered here.

Note also that this final coding exon of **ELP4** (stretching from **39438** to **39569** of this **RefSeq** entry) does **not** overlap the coding region of the **PAX6** gene itself (stretching from **16551** to **33028** of this **RefSeq** entry)[13].

In fact, the two entire genes do not overlap according to the evidence here. The entire **PAX6** gene extends from **5001** to **38170**. The portion of the **ELP4** gene that is included in this entry extends from **40170** (the end) to **38437** (in the opposite direction). This give a gap between the two genes stretching from **38171** to **38436**.

```
join(16551..16560,20128..20258,21186..21401,22106..22271,
28174..28332,28848..28930,29160..29310,29409..29524,
32102..32252,32943..33028)
/gene="PAX6"
/gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
/note="isoform a is encoded by transcript variant 1;
paired box homeotic gene-6; oculorhombin; aniridia type II
protein"
/codon_start=1
/product="paired box protein Pax-6 isoform a"
/protein_id=" NP_000271.1 "
/db_xref="GI:4505615"
/db_xref="CCDS: CCDS31451.1 "
/db_xref="GeneID: 5080 "
/db_xref="HGNC: HGNC:8620 "
/db_xref="MIM: 607108 "
/translation="MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISR
ILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEI
RDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRP
GWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFT
QEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRR
QASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPS
FTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ"
```

```
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:8620"
                /db_xref="MIM:607108"
```

```
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
                PAXNEB"
                /note="elongator acetyltransferase complex subunit 4"
                /db_xref="GeneID:26610"
                /db_xref="HGNC:HGNC:1171"
                /db_xref="MIM:606985"
```

As you will see later, **Ensembl** will confirm the lack of overlap between these two genes graphically as well as their relative positions.

Note that **ELP4** was associated with **aniridia** by **GeneCards**. However, I believe only because of its proximity to **PAX6**.

13 The features here only represent the **CDS** regions of the genes. I wonder why not the entire transcript length?

Why might you have expected more features than there were?

Summary:

All the evidence has suggested that **PAX6** has at least **2** isoforms. This would lead me to expect at least **2** CDS features here related to **PAX6**?

Full Answer:

The explanation from the **NCBI** is that this sort of **RefSeq** entry is intended to be used as a template against which sequences from an individual can be mapped to seek variations. Only a token **CDS** feature is included to indicate the position of the gene. For such an entry, recording every isoform is not essential.

This sounded convincing to me, Until I began to wonder why there were three **CDS** features for **ELP4** which is not even the gene primarily represented by this entry? Maybe I will ask more questions if and when I ever have the strength. In the meantime, mostly for my information, I record their exact explanation here.

" ... note that **RefSeqGene** defines genomic sequences to be used as reference standards for well-characterized genes. These sequences serve as a stable foundation for reporting mutations, for numbering exons and introns, and for defining the coordinates of other variations. We normally select one **RefSeq** transcript to serve as a reference standard. The goal is not to record all introns and exons of all isoforms, but just to choose one representative to help define the locus. Therefore, most of our **RSG** records have only a single **RefSeq** as reference standard. If an **LSDB** manager or other stakeholder requests that other **RefSeqs** be added as alternate standards, this can easily be done (with the complication that, if a public **LRG** exists, the **RefSeqGene** record is fixed). We receive requests from stakeholders to include **RefSeqs** that represent all known exons, or **RefSeqs** that have become community standards. Often, after creating an **RSG** using our own internal criteria, we receive stakeholder requests to change or add transcripts. Many of these requests come from the **LRG** project regarding transcripts to be included on the **LRG** records.

Generally, **RefSeq** accessions can be added or removed without reversioning, unless a transcript is upgraded or a new one defined that extends beyond the bounds of the **RSG**, or matches a new build of the genome, in which case the **RSG** will be extended and reversioned as needed.

Regarding the chromosomal locus, our standard range is 5 kb upstream from the 5' end and 2 kb downstream from the 3' end of the mRNAs with the greatest extent. For this calculation, we do indeed use all available **RefSeq** (**NM_**) accessions. If the database manager or stakeholder has information on promoters or other upstream or downstream regulatory regions, we can certainly extend the **RefSeqGene** locus to accommodate these.

Regarding mismatches, the goal is to exactly match the current build of the genome, unless there is overwhelming transcript and EST evidence that a mismatch should be retained.

Regarding the confusing subject of exon numbering, exon numbers are currently provided only on **RSG** genomic records based on a subset of available transcript **RefSeqs** for the gene. These are often those selected by locus-specific databases as reference sequence reporting standards. You can find an explanation of how exons are numbered here:

http://www.ncbi.nlm.nih.gov/refseq/rsg/faq/#exon

You will find links to more information on **RefSeqGenes** on the home page for the **RefSeqGene** project:

http://www.ncbi.nlm.nih.gov/refseq/rsg/

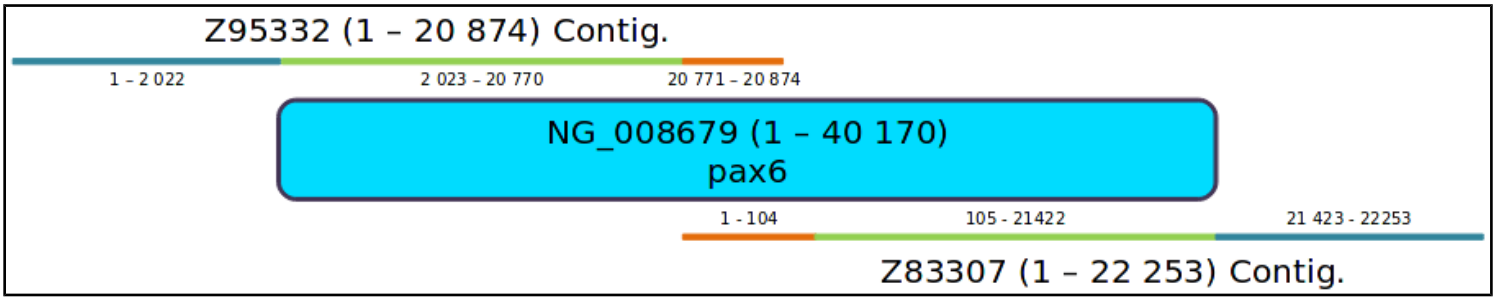
Regarding the **PAX6 RSG** sequence, only difference I see between **NG_008679.1** and the current build of the genome (**GRCh38**) is an extra **'G'** beyond the 3'-UTR of the **PAX6** transcripts (at **NC_000011.10:g.31,819,125**). ... "

Yes, well I think I followed most of that? and that my interpretation is broadly correct? In summary, there are no fixed rules.

How does the alignment you generated match up with the annotation of the original **RefSeq** entry you discovered?

Summary:

The most intuitive way of encapsulating graphically the way these two sequencing clones overlap was donated by **Cecilia Pinto** (**Oeiras**, **2013.12.09-12**). Much better than my rambling attempts, that I keep for sentimental reasons in the "Full Answer". Thank you Cecilia.



Full Answer:

Do not spend to much time working this one out, the picture above should be more than sufficient. I just needed to see it all balanced ... then I can sleep soundly?

If you do want to read on, I strongly suggest you look at the picture contributed by Cecilia (now promoted to the "**Summary** Answer") first. So simple! I have to admit I cannot follow my own wonderful table at all now ... at least, not without bleeding! Although, it did feel good at the time?

So …

Human DNA sequence from clone CFAT5 on **chromosome 11**, complete sequence
1. 20,874 bp linear DNA
Accession: Z95332.1 GI: 2190397
GenBank FASTA Graphics

Human DNA sequence from clone A1280 on **chromosome 11**, complete sequence
2. 22,253 bp linear DNA
Accession: Z83307.1 GI: 1730464
GenBank FASTA Graphics

```
Query  20771  GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  20830
Sbjct  1      GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  60
Query  20831  TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  20874
Sbjct  61     TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  104
```

The Query sequence is **Z95332** (Length **20,874**)

The Subject sequence is **Z83307** (Length **22,253**)

PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-18852	Z95332.1	2023-20874	
	18853-40170	Z83307.1	105-21422	

NG_008679 Range Start	NG_008679 Range End	NG_008679 Range	Z95332 Range Start	Z95332 Range Start	Z95332 Range	Z83307 Range Start	Z83307 Range End	Z83307 Range
-	-	-	1	2022	2022	-	-	-
1	18748	18748	2023	20770	18748	-	-	-
18749	18852	104	20771	20874 (end)	104	1	104	104
18853	40170 (end)	21319	-	-	-	105	21422	21318
-	-	-	-	-	-	21423	22253 (end)	831
		40171			20874			22253

Legend:

Not used in construction of **RefSeq** entry **NG_008679**

Non-overlapping **GenBank** entry used in construction of **RefSeq** entry **NG_008679**

Overlapping **GenBank** entry used in construction of **RefSeq** entry **NG_008679**

Total entry lengths

The **RefSeq** entry was thus constructed by overlapping the two **Genbank** entries and then manually trimming away the edges to form a biologically meaning region. If I was a bit brighter, I think I might have come to that conclusion without the fuss above? Oh well, one has to use what one has.

I refer you again to the far more intuitive way of encapsulating the same message graphically, donated by **Cecilia Pinto** that is now the "**Summary** Answer" above. Much better! Thank you once more Cecilia.

From your investigations using **Ensembl**:

Using the evidence of the protein alignments, which **PAX6** isoforms do the fruitfly orthologues most resemble?

The protein used to represent **PAX6** human is consistently **ENSP00000404100**. This can most easily be confirmed by clicking on the • Alignment (protein) link for each of the **2 Fruitfly** orthologues in turn to view the relevant orthologous protein alignments. This is the protein sequence of **isoform 5a**, probably chosen as it is the longer option (**436** amino acids as opposed to **422**) and so (from the crude informatics viewpoint) represents more information.

There are two **Fruitfly** orthologues recorded here, with the gene names **ey** and **toy**. Looking at the first few lines of the protein alignments for these genes, it is clear that that **14** amino acid insert that defines **isoform 5a** (**THADAKVQVLDNQN**) is not present in either. It is therefore reasonable to conclude that the representative fly proteins are both closest to the canonical protein sequence of **PAX6** human (**isoform 1**).

Protein alignment for **ey**

```
ENSP00000404100/1-436   ------MQN-------------------------SHSGVNQLGGVFVNGRPLPDSTRQ
FBpp0099810/1-898       GKPSPTMEAVEASTASHPHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGGRPLPDSTRQ
                              *:                          .***********.**********


ENSP00000404100/1-436   KIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGG
FBpp0099810/1-898       KIVELAHSGARPCDISRILQ-------------VSNGCVSKILGRYYETGSIRPRAIGG
                        ******************              ************************


ENSP00000404100/1-436   SKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASE
FBpp0099810/1-898       SKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQ
                        *******.******:******************.*.*********************::
```

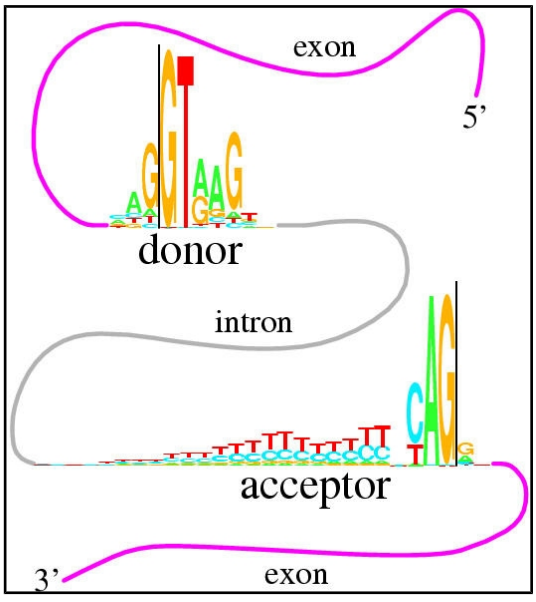
Protein alignment for **toy**

```
ENSP00000404100/1-436   -MQN-----------------------SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHS
FBpp0088249/1-543       MMLTTEHIMHGHPHSSVGQSTLFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHS
                         * .                      .***:******:********************


ENSP00000404100/1-436   GARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATP
FBpp0088249/1-543       GARPCDISRILQ-------------VSNGCVSKILGRYYETGSIKPRAIGGSKPRVATT
                        ***********              ******************:*************.


ENSP00000404100/1-436   EVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADG
FBpp0088249/1-543       PVVQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQ
                         **.***:******************* **..*****************:*:* . :
```

Well, maybe also it is not that simple? I would not be surprised If there were isoforms for **ey** and/or **toy** that were roughly equivalent to human **isoform 5a**. The alignment displayed could well reflect the relatively arbitrary choice of **Ensembl** as to which isoform it decides to use for the alignments, rather than any deep and meaningful biological truth. Already you can see that **Ensembl** prefers the (presumably) less important human isoform, merely because it is longer (more letters to match). Again, useful though these displays are, caution is required before reading too much "biology" into them.

Ensembl does not pick up all fruitfly homologues of **PAX6**? Again, I wonder why. Mind you, **Ensembl** does only claim "**Selected orthologues**"? Still **prd**, in particular, is a pretty important one to pass over!

What are the first two bases and what are the last two bases of nearly every intron?



As you are probably well aware, introns are highly conserved at each end. They typically begin with **GT** and end with **AG**. This rule is obeyed by all but one of the introns of this transcript (**intron 3-4** starts **GC** rather than **GT**).

As the cartoon suggests, the conservation does not apply just to the first and last two bases, but that is where the conservation is most strict. So strict that when exceptions from this rule were sought in the databases, it was thought most of the deviations were due to annotation error!

The cartoon also suggests that introns have **C/T rich regions** towards their ends (the **Polypyrimidine tract**). This too is clearly evident in most of the introns of this transcript, even though only small parts of the introns are displayed.

How long is the sixth exon and why would this concur with your expectations?

It is **42** base pairs long, so it codes for **14** amino acids. Specifically, it codes for the **14** extra amino acids that define **isoform 5a**.

Explain the **Start Phase** and **End Phase** columns?

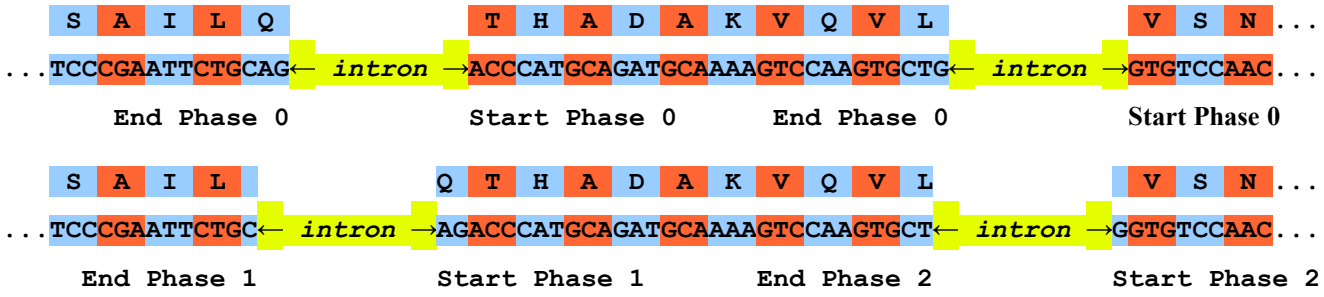
An exon/intron boundary can occur anywhere in a codon. The **Start** and **End Phases** record how an intron has been inserted into a coding region with respect to the coding reading frame.

If an exon ends at the end of a codon, then its **End Phase** is **0**.
Clearly, the next exon must begin at the start of a codon. Its Start Phase is also **0**.

If an exon ends after the first base of a codon, then its **End Phase** is **1**.
Clearly, the next exon must begin after the first base of a codon. Its **End Phase** is also **1**.

If an exon ends after the second base of a codon, then its **End Phase** is **2**.
Clearly, the next exon must begin after the second base of a codon. Its **End Phase** is also **2**.

I attempt a picture, though I am sure that is clear? I just like pictures, and lots of colours.



Why does **Prints** appear to predict four **Paired_domains**?

Prints does not find the **Homeobox_domain** at all. If you were to investigate by using the PRINTS search carefully, you will find it nearly does, but the evidence is not quite strong enough. As has been discussed, none of these systems are perfect. They all occassionally fail. That is why it is always best to use Interpro to consult them all and deliver a consensus answer.

Prints appears to find **FOUR Paired_domains**. This is only because of the way **Prints** works. **Prints** finds **FOUR** signatures (or **motifs**) that together indicate **ONE Paired domain**. **Prints** searches for ordered series of **motifs** that together indicate **domains**. Here it reports each of four motifs separately, but it is only claiming one **Paired domain.**

Which domain, **Paired domain** or **Homeobox domain** is more common in humans?
How many human **PAX** genes are there?

As you will have expected, there are but **9 Paired domain**s in the Human genome. There are many more **Homeobox domain**s.

Are all the **PAX** genes on **Chromosome 11**?

Of course not? What a stupid question!

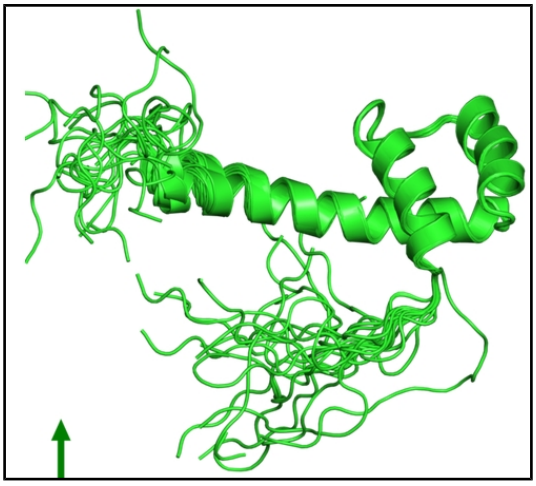
Well, I suppose they could all be on **Chromosome 11**? By chance … or maybe design … who knows, the lack of predictable pattern in all this business never ceases to astound me.

But, philosophy aside, the answer is **NO**.

How does **Interpro** match with the **PAX6 Paralogues** reported by **Ensembl/GeneCards** earlier?

The evidence from both **GeneCards** and **Ensembl** is that there are **9 PAX** paralogues in Human. Yep, we all agree and … these questions are becoming a trifle repetitive one feels!

Can you explain the strangely frayed ends displayed in some of the representations of the **2cue** 3D structure?



2cue is a 3D structure determined by **N**uclear **M**agnetic **R**esonance (**NMR**).This is a process that does not involve immobilizing the target as a crystal (as is the case with structures determined by **X-ray crystallography**). Parts of the protein will still be moving around whilst its structure is being determined.

I think of **NMR** as analogous to taking a long exposure photograph of a group of children. Each child will appear in many different places! The frayed ends represent various positions in which the ends of the **homeobox** were detected during the **NMR** process.

In some views, including the one you were offered to move around, all the possible positions are averaged out before the structure is stored. I prefer the fuzzy view ... much more fun.

I broadly believe that which I have just typed, however, I must stress that my understanding of **NMR** is tragically incomplete. If anyone would like to offer a better explanation, I am very willing to hear it.

From your investigations of **Domain & Motif identification** using **Interpro**

Do you think it a good idea for **Interpro** to offer feature prediction programs as well as domain database searches?

Well … why not? The purpose of **InterProScan** is to associate regions of query proteins with **Interpro** domains. This was originally achieved, exclusively, by simply comparing a query sequence with all entries of relevant individual domain databases. These entries being representations of alignments of examples of specific domains constructed by homology searching (i.e. **blast** and similar).

I would suggest including a few predictor programs would provide extra evidence gathered from more general, more theoretical definitions of domains. I would imagine the inclusion of these programs has improved and widened the picture provided by **InterProScan**.

Searching domain databases, typically composed of **HMM profiles**, such as **Pfam**, **Prosite** and **PRINTS** is quite different to running the predictor programs. As I cannot improve on the justification of this claim offered to me by Geoff Barton (Head of the group responsible for **Jalview**, **Jpred**, **Jnet** and much more), I will just reproduce his explanation here:

> " … The main difference is that with an **HMM profile** you have a "specific" example of a domain or motif whereas with something like **COILS**, you have something trained across all examples.
>
> For example, for secondary structure prediction, you could (a) do predictions of alpha-helix and beta-strand just by aligning a sequence to a protein of known structure, or an **HMM** from a family of aligned proteins of known structure. This is a specific case of secondary structure in the context of one protein family. Or (b) you can train a predictor from **ALL** protein families and then apply this. The advantage of (a) is it is very specific to the individual protein family and so should be more accurate for that family. The disadvantage is that it does not generalise to proteins that are not very like the specific example. The advantage of (b) is that it will work with any protein but will likely be less accurate than (a) for proteins that fit into the (a) category. … "

Do you think the Coil prediction might be correct?

I do not recall anything in what we have discovered thus far that would directly suggest there should be a **coiled coil** here, in the middle of the **HTH**. However, wikipedia does suggest **coiled coils** are associated with **transcription factors** (which **pax6_human** is).

" ... Many **coiled coil**-type proteins are involved in important biological functions such as the regulation of **gene expression**, e.g. **transcription factors**. ... "

I think I would not be overly convinced by this prediction, but I would not make that judgement with any great confidence. The all knowing **wikipedia** says:

" … **Coiled coils** usually contain a repeated pattern, ***hxxhcxc***, of **h**ydrophobic (***h***) and **c**harged (***c***) amino-acid residues, referred to as a **heptad repeat**. … "

Geoff Barton comments:

> " … Sometimes the pattern that is particular to **coiled-coils** also turns up in other helices that pack against each other. You would need to look at some examples of coiled-coil structures to see if the example you are using fits structurally. … "

Which seems very reasonable. The **heptad repeat** pattern could easily occur just by chance. **COILS** surely cannot predict the structure of the helices well enough to make an assured judgement? **COILS** offers a suggestion the user must follow up with other resources.

There is also the evidence that **Jpred** (a system for secondary structure prediction), possibly using the **COILS** program disguised as **LUPAS**, does not detect any coiled coils. This could be for a number of reasons. Possibly **LUPAS** is not the same program as **COILS**, or it is a different version, or **Jpred** runs **COILS**, but with different parameters.

Not many clear and confident answers in Bioinformatics are there!

Why might you suppose **Interpro** predicts only **2** of the **3 helix-turn-helix domain**s that might be expected?

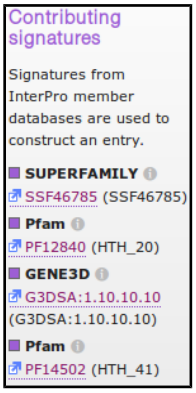
2 Winged helix-turn-helix (wHTH) DNA-binding domains are predicted coincident with the helical triplets of the **Paired domain**. This should broadly match your expectations.

No **helix-turn-helix (HTH) domain** is detected coincident with the **Homeobox domain**, where one might also have been expected?

I am not entirely certain why this might be, so I speculate.

Pfam attempts to classify a variety of types of **HTH**, and offers a range of **HTH** domain models (**HTH_17**, **HTH_38**, **HTH_39** and **HTH_40** to name but a few) and a number of **wHTH** domain models (including **HTH_33** and **HTH_24**).

Interpro also has a considerable number of **HTH** entries (**IPR017895**, **IPR032877**, **IPR007394**, **IPR013197** and more) and **wHTH** entries (**IPR005104**, **IPR023120** to name but **2**).



Interpro does use **Pfam** models to detect its various flavours of **HTH/wHTH** domain, but it does so selectively. For example, to detect the **wHTH** domains discovered here, only two **Pfam** families were used **HTH_20** and **HTH_41**, see illustration). These appear not to have matched in this instant as only a **G3DSA** entry is quoted.

All the above suggests that no one model exists to pick up all **HTH** domains? Possibly also, the fact that **HTH** domains come in such a variety of forms makes them difficult to detect reliably?

There is a simple **EMBOSS** program to detect **HTH**s. It easily detected the **Homeobox domain HTH** but essentially failed to detect the **wHTH**s recorded here. This must be because the, very simple, model (based on a **Weight Matrix** built from about **100** examples) used by the program only reliably applies to a specific range of **HTH** domains/motifs that includes the one in the **Homeobox domain** of the human **PAX6** protein?

I am very open to better explanations. I am not completely convinced by the discussion above.

DPJ – 2016.07.02