Introduction to Bioinformatics

23-27 March 2020

Practical 1: Databases and Tools

Part i) – Aniridia from The NCBI

Thursday 16 April 2020

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

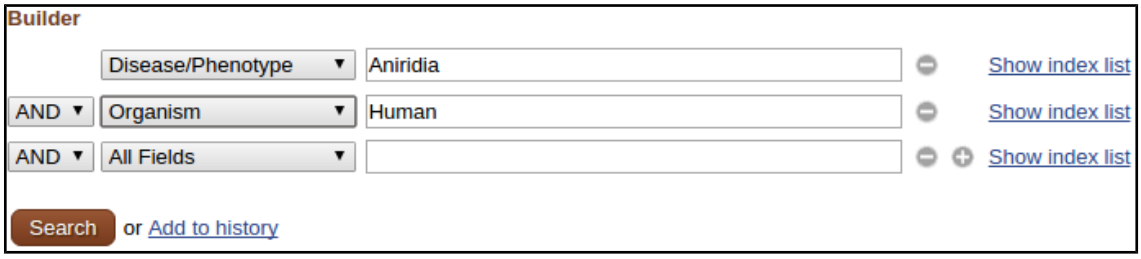
As a starting point for this exercise, imagine you have a vital interest in discovering and investigating the main human gene responsible for the terrible disease of the eye, **Aniridia**. There are many ways (including **google**!) you could discover what this gene might be. I choose to delve into the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

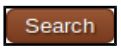
So, begin by going to the **Home Page** of the **The National Center for Biotechnology Information (NCBI)** ("**http://www.ncbi.nlm.nih.gov/**").

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

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

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



Click on the pretty 

button.

Name/Gene ID	Description	Location	Aliases	MIM
WT1 ID: 7490	WT1 transcription factor [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (32387775..32435539, complement)	AWT1, GUD, NPHS4, WAGR, WIT-2, WT33	607102
PAX6 ID: 5080	paired box 6 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31784792..31817961, complement)	AN, AN1, AN2, ASGD5, D11S812E, FVH1, MGDA, WAGR	607108
TRIM44 ID: 54765	tripartite motif containing 44 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (35662692..35818007)	AN3, DIPB, HSA249128, MC7	612298
ELP4 ID: 26610	elongator acetyltransferase complex subunit 4 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31509767..31790324)	AN, AN2, 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. All should really be examined, but this is just an exercise, so trust me … it is **PAX6** that is the most interesting gene[1], in this context. This is the one to follow up by clicking on the link to its details.

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

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

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

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

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

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

Summary

Official Symbol PAX6 provided by HGNC
Official Full Name paired box 6 provided by HGNC

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

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

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

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

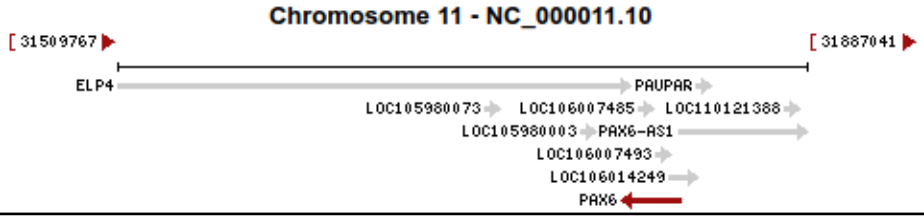
From the **Genomic context** section it can be seen that:

See PAX6 in Genome Data Viewer

Location: 11p13

Exon count: 25

Annotation release	Status	Assembly	Chr	Location
109.20200228	current	GRCh38.p13 (GCF_000001405.39)	11	NC_000011.10 (31784792..31817961, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (31806340..31839509, complement)



- **PAX6** is situated on **Chromosome 11**, band **p13**.

- **PAX6** is on the complementary strand relative to that chosen to represent **Chromosome 11**.

- **ELP4** (another human gene listed as associated with **Aniridia**) is very close, on the opposite strand to **PAX6**.

- There are **25** exons for **PAX6**.

- A number of other features are recorded here. Most are not genes and so we will ignore them for now. However, do note the feature PAX6-AS1. This is a **non-coding RNA** that will play a small part in the dance to follow.

- Note also the feature PAUPAR. This is it only appearance that we come across at the **NCBI**, but it does seem to have a marginally higher profile elsewhere. Both **PAX6-AS1** and **PAUPAR** are reported as **PAX6 antisense RNA**s, which informs one only slightly.

- Note how the location of the **PAX6** gene has moved slightly between the **current assembly** of the **Human Genome** and the **previous assembly**.

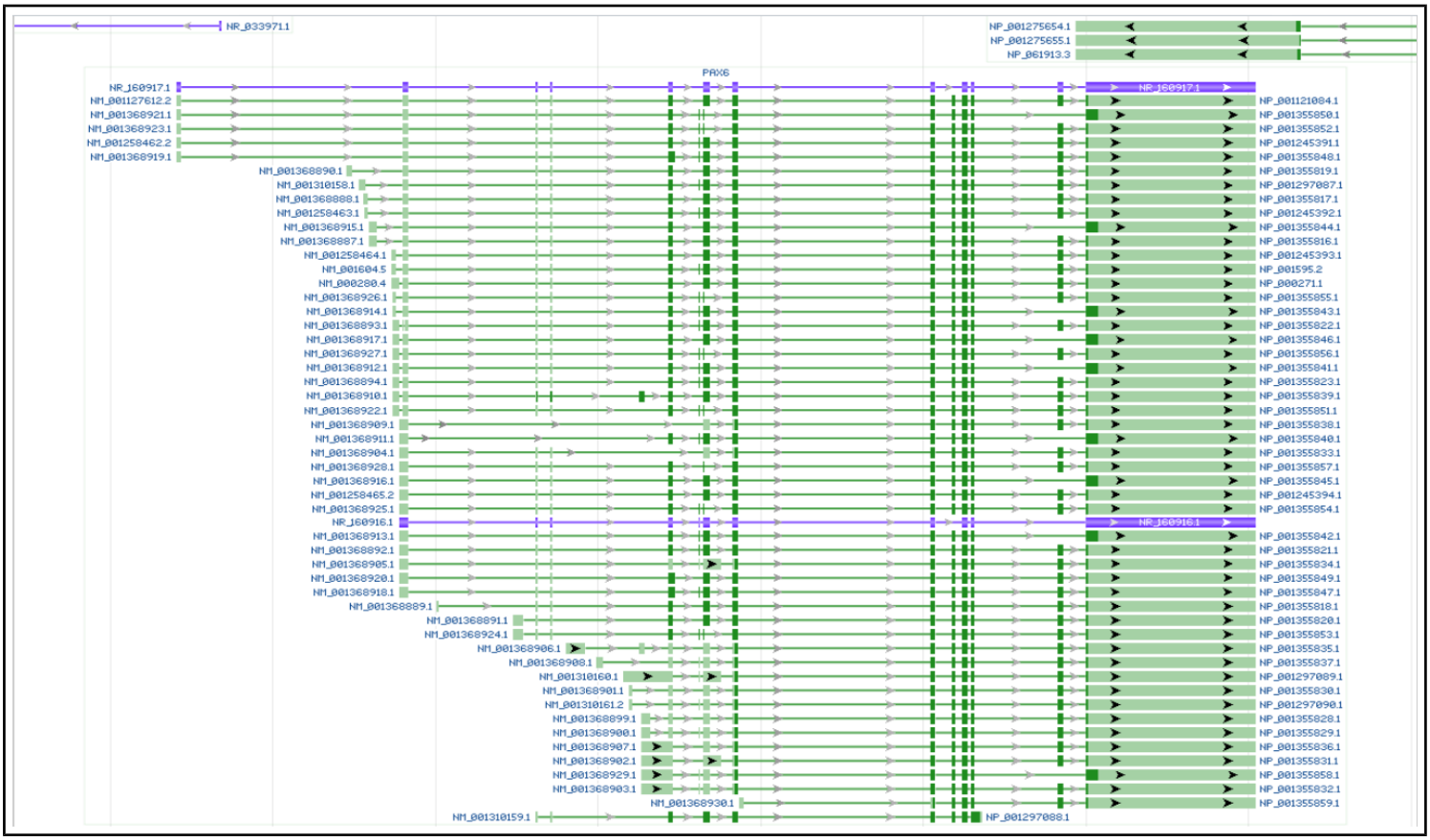
Annotation release	Status	Assembly	Chr	Location
109.20200228	current	GRCh38.p13 (GCF_000001405.39)	11	NC_000011.10 (31784792..31817961, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (31806340..31839509, complement)

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

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

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

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



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

The top **PAX6** line represents one of the two the **non-coding Transcripts** that this database associates with **PAX6**. A **non-coding Transcript** has a name (**Accession Code**) that begins **NR_** (**N**on-coding **R**NA). The **Accession Code** is displayed to the left of the transcript line (in this case, **NR_160917**, the **.1** at the end is the version number). The **Accession Code** for the **Protein Product** of the **Transcript** is displayed at the right hand end of the **Transcript** line (in this case it is blank, of course). The pretty **blue** blobs represent the **Introns**, the equally attractive **blue** lines joining the blobs, represent the **Exons**.

The choice of the first two letters of the **Accession codes** you see here reflect the status of the **Genes**, **Transcripts**, or **Proteins** they represent. Here we see, **NR_**, **NM_** and **NP_** representing **non-coding Transcripts**, **curated coding Transcripts** and **protein products of NM_ Transcripts** respectively. There are, for example, also **Accessions codes** prefixed by **XR_**, **XM_** and **XP_** representing ***predicted*** **non-coding Transcripts**, ***predicted*** **coding Transcripts** and **protein products of XM_ Transcripts** respectively. A full list of **RefSeq Accesson code** prefixes can be found Here.

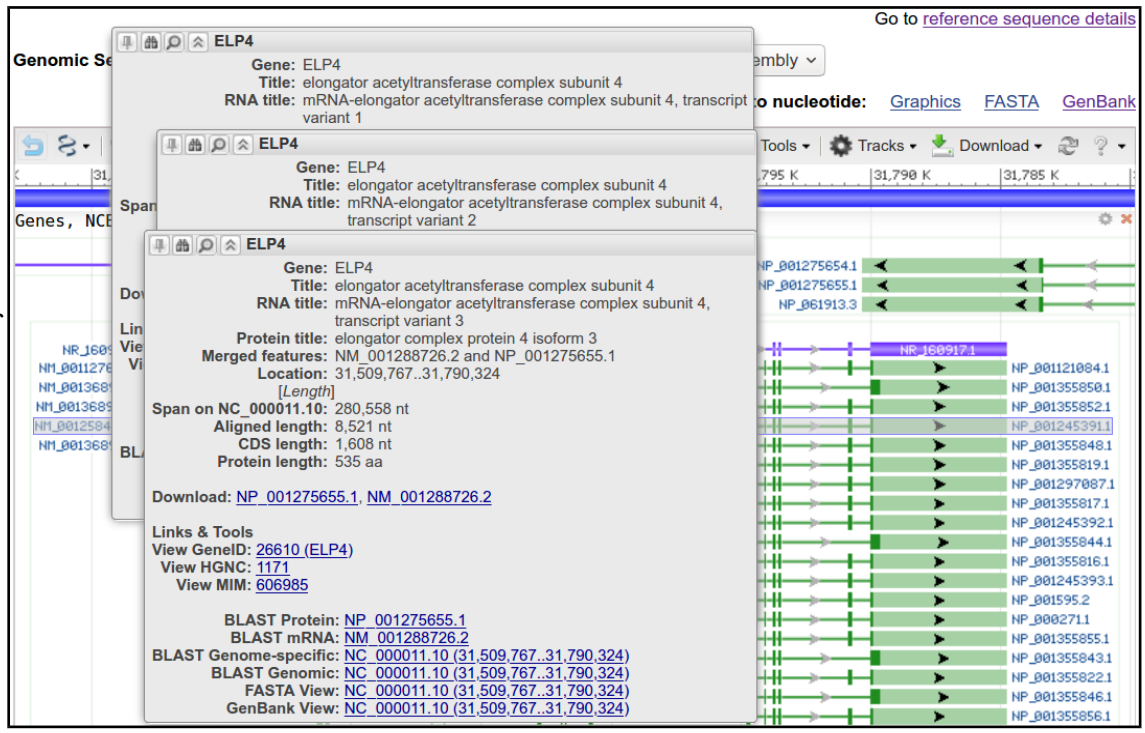
All the **PAX6 Transcripts** shown here, excepting the two **blue NR_ Transcripts**, are **curated coding Transcripts**. Each **coding Transcript** is represented by a **Transcript** line showing **CoDing Sequence** (**CDS**) **Introns** as **dark green** blobs, **Untranslated Regions** (**UTR**s) in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

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

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

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

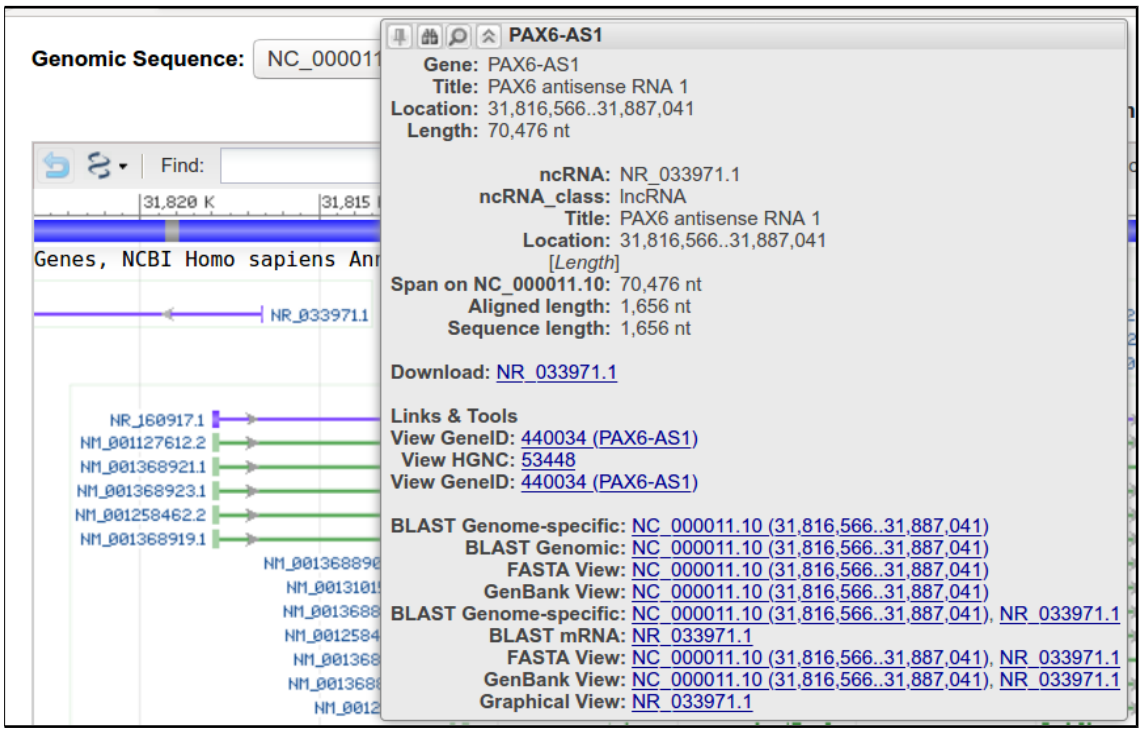
It should be clear that these are three **Transcripts** of the **ELP4** gene that was noted when looking at the **Genomic context** section.



Note also an enigmatic **non-coding Gene** with just one **Transcript** called, endearingly, **NR_0339711**. Hover over the **NR_0339711** transcript line and a new box of tasteful grey will sally forth telling all there is to know about the enigmatic **NR_0339711**!

A swift glance will be sufficient for you to see that **NR_0339711** is simply the **PAX6-AS1** gene we first met in the **Genomic context** section, in rather thin disguise. **NR_0339711** being the **Accession code** for the **non-coding RNA** product of the gene called **PAX6-AS1**.

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



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

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

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

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

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

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

Genomic

1. **NG_008679.1 RefSeqGene**

Range	5001..38170
Download	GenBank, FASTA, Sequence Viewer (Graphics), LRG_720

mRNA and Protein(s)

1. NM_000280.4 → NP_000271.1 **paired box protein Pax-6 isoform a**

See identical proteins and their annotated locations for NP_000271.1

Status: REVIEWED

Description	Transcript Variant: This variant (1) initiates from the B (P1) promoter and encodes isoform a. Variants 1, 3, 6, 7, and 12-16 all encode the same isoform (a).
Source sequence(s)	BP394576, DA078958, M93650, Z83307
Consensus CDS	CCDS31451.1
UniProtKB/Swiss-Prot	P26367
UniProtKB/TrEMBL	Q66SS1
Related	ENSP00000495109.1, ENST00000643871.1
Conserved Domains (2) summary	
smart00351 Location:4 → 128	PAX; Paired Box domain
pfam00046 Location:214 → 267	Homeobox; Homeobox domain

2. NM_001127612.2 → NP_001121084.1 **paired box protein Pax-6 isoform a**

See identical proteins and their annotated locations for NP_001121084.1

Status: REVIEWED

Description	Transcript Variant: This variant (3) differs in the 5' UTR compared to variant 1. It initiates from the A (P0) promoter. Variants 1, 3, 6, 7, and 12-16 all encode the same isoform (a).
Source sequence(s)	AK314470, BE221553, BM557761, BM725029, BP394398, BP394576, BU072567, BX089704, BX114225, CA397536, DA183294, Z83307
Consensus CDS	CCDS31451.1
UniProtKB/Swiss-Prot	P26367

Slide gracefully to the bottom of this list and you will see that the **NCBI** admit to **51 Messenger RNAs** and **2 non-coding RNAs**.

50. NM_001368930.1 → NP_001355859.1 **paired box protein Pax-6 isoform o**

Status: REVIEWED

Source sequence(s)	Z83307
Related	ENSP00000492769.1, ENST00000638965.1
Conserved Domains (1) summary	
pfam00046 Location:13 → 66	Homeobox; Homeobox domain

51. NM_001604.5 → NP_001595.2 **paired box protein Pax-6 isoform b**

See identical proteins and their annotated locations for NP_001595.2

Status: REVIEWED

Description	Transcript Variant: This variant (2) uses an alternate splice site in the 5' UTR, and includes an alternate in-frame exon in the 5' coding region, compared to variant 1. It initiates from the B (P1) promoter. The encoded isoform (b, also known as 5a) is longer than isoform a. Variants 2, 4, 5, 8, and 17-19 all encode the same isoform (b).
Source sequence(s)	BP394576, BX640762, CV569250, DA078958, DA141443, Z83307
Consensus CDS	CCDS31452.1
UniProtKB/Swiss-Prot	P26367
UniProtKB/TrEMBL	F1T0F8
Related	ENSP00000404100.1, ENST00000419022.6
Conserved Domains (2) summary	
smart00351 Location:4 → 142	PAX; Paired Box domain
pfam00046 Location:228 → 281	Homeobox; Homeobox domain

RNA

1. NR_160916.1 **RNA Sequence**

Status: REVIEWED

Source sequence(s)	Z83307, Z95332

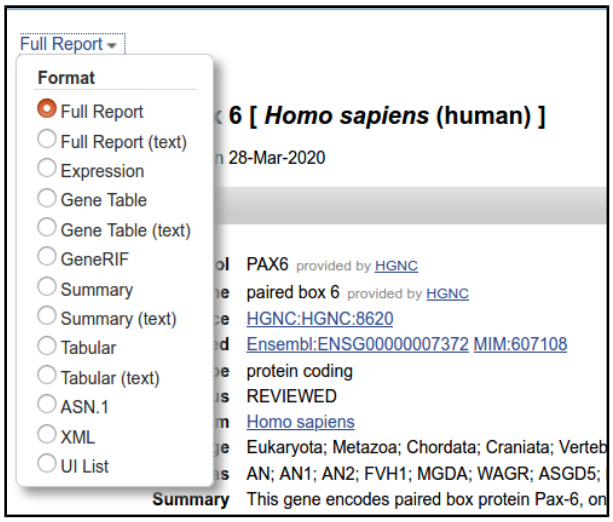
2. NR_160917.1 **RNA Sequence**

Status: REVIEWED

Source sequence(s)	Z83307, Z95332

Fine, now for quest two, which is to discover how many distinct **isoforms** does the **NCBI** assign to **PAX6**?

Largely because of the mendacious insistence that each **mRNA** generates a different protein product (not unique to the **NCBI**), this is not going to be straight forward. However, it can be done in a number of ways, by careful examination of the textual records for **PAX6**. Perhaps the most efficient way is to explore the **Textual Gene Table** for **PAX6**.



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

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

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

```
PAX6 paired box 6[Homo sapiens]
Gene ID: 5080, updated on 28-Mar-2020


Reference GRCh38.p13 Primary Assembly NC_000011.10  (minus strand) from: 31817961 to: 31784792
RNA transcript variant 53 NR_160917.1, 12 exons,  total annotated spliced exon length: 6797

Exon table for  RNA  NR_160917.1
Genomic Interval Exon          Gene Interval Exon             Exon Length     Intron Length
-----------------------------------------------------------------------------------------------
31817961-31817809              1-153          153             6793
31811015-31810828              6947-7134              188             3902
31806925-31806849              11037-11113            77              386
31806462-31806402              11500-11560            61              3567
31802834-31802704              15128-15258            131             927
31801776-31801561              16186-16401            216             704
31800856-31800691              17106-17271            166             5902
31794788-31794630              23174-23332            159             827
31793802-31793652              24160-24310            151             98
31793553-31793438              24409-24524            116             2577
31790860-31790710              27102-27252            151             690
31790019-31784792              27943-33170            5228

RNA transcript variant 52 NR_160916.1, 11 exons,  total annotated spliced exon length: 6641

Exon table for  RNA  NR_160916.1
Genomic Interval Exon          Gene Interval Exon             Exon Length     Intron Length
-----------------------------------------------------------------------------------------------
31811121-31810828              6841-7134              294             3902
31806925-31806849              11037-11113            77              386
31806462-31806402              11500-11560            61              3567
31802834-31802704              15128-15258            131             791
31801912-31801871              16050-16091            42              94
31801776-31801561              16186-16401            216             704
31800856-31800691              17106-17271            166             5902
31794788-31794630              23174-23332            159             827
31793802-31793652              24160-24310            151             98
31793553-31793438              24409-24524            116             3418
31790019-31784792              27943-33170            5228
```

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

```
mRNA transcript variant 24 NM_001368903.1, 10 exons,  total annotated spliced exon length: 7282
protein isoform d NP_001355832.1, 7 coding  exons,  annotated AA length: 286

Exon table for  mRNA  NM_001368903.1 and protein NP_001355832.1
Genomic Interval Exon           Genomic Interval Coding         Gene Interval Exon              Gene Interval Coding            Exon Length     Coding Length   Intron Length
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
31803673-31802704               14289-15258             970             791
31801912-31801871               16050-16091             42              94
31801776-31801561               16186-16401             216             704
31800856-31800691               31800805-31800691               17106-17271             17157-17271             166             115             5902
31794788-31794630               31794788-31794630               23174-23332             23174-23332             159             159             515
31794114-31794032               31794114-31794032               23848-23930             23848-23930             83              83              229
31793802-31793652               31793802-31793652               24160-24310             24160-24310             151             151             98
31793553-31793438               31793553-31793438               24409-24524             24409-24524             116             116             2577
31790860-31790710               31790860-31790710               27102-27252             27102-27252             151             151             690
31790019-31784792               31790019-31789934               27943-33170             27943-28028             5228            86

mRNA transcript variant 51 NM_001368930.1, 7 exons,  total annotated spliced exon length: 6011
protein isoform o NP_001355859.1, 6 coding  exons,  annotated AA length: 221

Exon table for  mRNA  NM_001368930.1 and protein NP_001355859.1
Genomic Interval Exon           Genomic Interval Coding         Gene Interval Exon              Gene Interval Coding            Exon Length     Coding Length   Intron Length
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
31800661-31800539               17301-17423             123             5750
31794788-31794630               31794708-31794630               23174-23332             23254-23332             159             79              515
31794114-31794032               31794114-31794032               23848-23930             23848-23930             83              83              229
31793802-31793652               31793802-31793652               24160-24310             24160-24310             151             151             98
31793553-31793438               31793553-31793438               24409-24524             24409-24524             116             116             2577
31790860-31790710               31790860-31790710               27102-27252             27102-27252             151             151             690
31790019-31784792               31790019-31789934               27943-33170             27943-28028             5228            86

mRNA transcript variant 9 NM_001310159.1, 9 exons,  total annotated spliced exon length: 1393
protein isoform c NP_001297088.1 (CCDS86190.1), 8 coding  exons,  annotated AA length: 401

Exon table for  mRNA  NM_001310159.1 and protein NP_001297088.1
Genomic Interval Exon           Genomic Interval Coding         Gene Interval Exon              Gene Interval Coding            Exon Length     Coding Length   Intron Length
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
31806925-31806849               11037-11113             77              386
31806462-31806402               31806411-31806402               11500-11560             11551-11560             61              10              3567
31802834-31802704               31802834-31802704               15128-15258             15128-15258             131             131             927
31801776-31801561               31801776-31801561               16186-16401             16186-16401             216             216             704
31800856-31800691               31800856-31800691               17106-17271             17106-17271             166             166             5902
31794788-31794630               31794788-31794630               23174-23332             23174-23332             159             159             515
31794114-31794032               31794114-31794032               23848-23930             23848-23930             83              83              229
31793802-31793652               31793802-31793652               24160-24310             24160-24310             151             151             98
31793553-31793205               31793553-31793264               24409-24757             24409-24698             349             290
```

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

```
mRNA transcript variant 24 NM_001368903.1, 10 exons,  total annotated spliced exon length: 7282
protein isoform d NP_001355832.1, 7 coding  exons,  annotated AA length: 286
```

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

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

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

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

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

Here you see the portion of the **RefSeq** entry for the entirety of **Chromosome 11** that covers the **PAX6** gene region As you can see the **Chromose 11 RefSeq** entry is **NC_000011**. '**C**' for **Chromosome**, of course. As previously, the number after the '**.**' is a version number.

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

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

Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly

NCBI Reference Sequence: NC_000011.10

FASTA Graphics

```
LOCUS       NC_000011               33170 bp    DNA     linear   CON 02-MAR-2020
DEFINITION  Homo sapiens chromosome 11, GRCh38.p13 Primary Assembly.
ACCESSION   NC_000011 REGION: complement(31784792..31817961)
VERSION     NC_000011.10
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.39
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 33170)
  AUTHORS   Taylor,T.D., Noguchi,H., Totoki,Y., Toyoda,A., Kuroki,Y., Dewar,K.,
            Lloyd,C., Itoh,T., Takeda,T., Kim,D.W., She,X., Barlow,K.F.,
            Bloom,T., Bruford,E., Chang,J.L., Cuomo,C.A., Eichler,E.,
            FitzGerald,M.G., Jaffe,D.B., LaButti,K., Nicol,R., Park,H.S.,
            Seaman,C., Sougnez,C., Yang,X., Zimmer,A.R., Zody,M.C.,
            Birren,B.W., Nusbaum,C., Fujiyama,A., Hattori,M., Rogers,J.,
            Lander,E.S. and Sakaki,Y.
  TITLE     Human chromosome 11 DNA sequence and analysis including novel gene
            identification
  JOURNAL   Nature 440 (7083), 497-500 (2006)
   PUBMED   16554811
REFERENCE   2  (bases 1 to 33170)
  CONSRTM   International Human Genome Sequencing Consortium
  TITLE     Finishing the euchromatic sequence of the human genome
  JOURNAL   Nature 431 (7011), 931-945 (2004)
   PUBMED   15496913
REFERENCE   3  (bases 1 to 33170)
  AUTHORS   Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C.,
            Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., Funke,R.,
```

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

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

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

```
     mRNA            join(1..153,6947..7134,11037..11113,11500..11560,
                     15128..15258,16186..16401,17106..17271,23174..23332,
                     23848..23930,24160..24310,24409..24524,27102..27252,
                     27943..33170)
                     /gene="PAX6"
                     /gene_synonym="AN; AN1; AN2; ASGD5; D11S812E; FVH1; MGDA;
                     WAGR"
                     /product="paired box 6, transcript variant 3"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NM_001127612.2"
                     /db_xref="GeneID:5080"
                     /db_xref="HGNC:HGNC:8620"
                     /db_xref="MIM:607108"
     mRNA            join(1..153,6947..7134,11037..11113,11500..11560,
                     15128..15258,16050..16091,16186..16200,17106..17271,
                     23174..23332,23848..23930,24160..24310,24409..24524,
                     27943..33170)
                     /gene="PAX6"
                     /gene_synonym="AN; AN1; AN2; ASGD5; D11S812E; FVH1; MGDA;
                     WAGR"
                     /product="paired box 6, transcript variant 42"
mRNA                 Highlight All  Match Case  Match Diacritics  Whole Words  1 of 54 matches
```

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

```
     mRNA            complement(27638..>33170)
                     /gene="ELP4"
                     /gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
                     PAX6NEB; PAXNEB"
                     /product="elongator acetyltransferase complex subunit 4,
                     transcript variant 2"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NM_001288725.2"
                     /db_xref="GeneID:26610"
                     /db_xref="HGNC:HGNC:1171"
                     /db_xref="MIM:606985"
     mRNA            complement(27638..>33170)
                     /gene="ELP4"
                     /gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
                     PAX6NEB; PAXNEB"
                     /product="elongator acetyltransferase complex subunit 4,
                     transcript variant 3"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NM_001288726.2"
                     /db_xref="GeneID:26610"
                     /db_xref="HGNC:HGNC:1171"
                     /db_xref="MIM:606985"
     mRNA            complement(27638..>33170)
                     /gene="ELP4"
                     /gene_synonym="AN; AN2; C11orf19; dJ68P15A.1; hELP4;
                     PAX6NEB; PAXNEB"
                     /product="elongator acetyltransferase complex subunit 4,
                     transcript variant 1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefSeq."
                     /transcript_id="NM_019040.5"
                     /db_xref="Ensembl:ENST00000640961.2"
                     /db_xref="GeneID:26610"
                     /db_xref="HGNC:HGNC:1171"
                     /db_xref="MIM:606985"
     ORIGIN
mRNA                 Highlight All  Match Case  Match Diacritics  Whole Words  52 of 54 matches
```

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

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

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

```
     CDS             join(17157..17271,23174..23332,23848..23930,24160..24310,
                     24409..24524,27943..28332)
                     /gene="PAX6"
                     /gene_synonym="AN; AN1; AN2; ASGD5; D11S812E; FVH1; MGDA;
                     WAGR"
                     /note="isoform n is encoded by transcript variant 50;
                     Derived by automated computational analysis using gene
                     prediction method: BestRefSeq."
                     /codon_start=1
                     /product="paired box protein Pax-6 isoform n"
                     /protein_id="NP_001355858.1"
                     /db_xref="GeneID:5080"
                     /db_xref="HGNC:HGNC:8620"
                     /db_xref="MIM:607108"
                     /translation="MGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
                     EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVF
                     ARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVY
                     QPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQDSFPLVCQFQ
                     FKFPEVNLICLNTGQDYSKKKKKKKKERKYCVNSVSDYGDTTVELSGKKEKWLLEPLQ
                     FYNCVLYCTTGEGMDLKQGPLYTEGTISVGTNLHFGIQTFIHFGVLFVNGHLYVIMKK
                     RTM"
     CDS             join(17157..17271,23174..23332,23848..23930,24160..24310,
                     24409..24524,27102..27252,27943..28028)
                     /gene="PAX6"
                     /gene_synonym="AN; AN1; AN2; ASGD5; D11S812E; FVH1; MGDA;
                     WAGR"
                     /note="isoform d is encoded by transcript variant 22;
                     Derived by automated computational analysis using gene
                     prediction method: BestRefSeq."
                     /codon_start=1
                     /product="paired box protein Pax-6 isoform d"
                     /protein_id="NP_001355830.1"
                     /db_xref="GeneID:5080"
                     /db_xref="HGNC:HGNC:8620"
PAX-6 isoform        Highlight All  Match Case  Match Diacritics  Whole Words  37 of 51 matches
```

Just one question remains. How did I determine the answers to queries two and three? Well … I most certainly did not spend ages reading through the tables or the **GenBank** format!! I spent just long enough to see ***HOW*** the queries could be answered, and then I downloaded the text files to my computer and wrote simple programs to extract the information I wanted from each of the text files.

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

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

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

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

DPJ – 2020.04.16