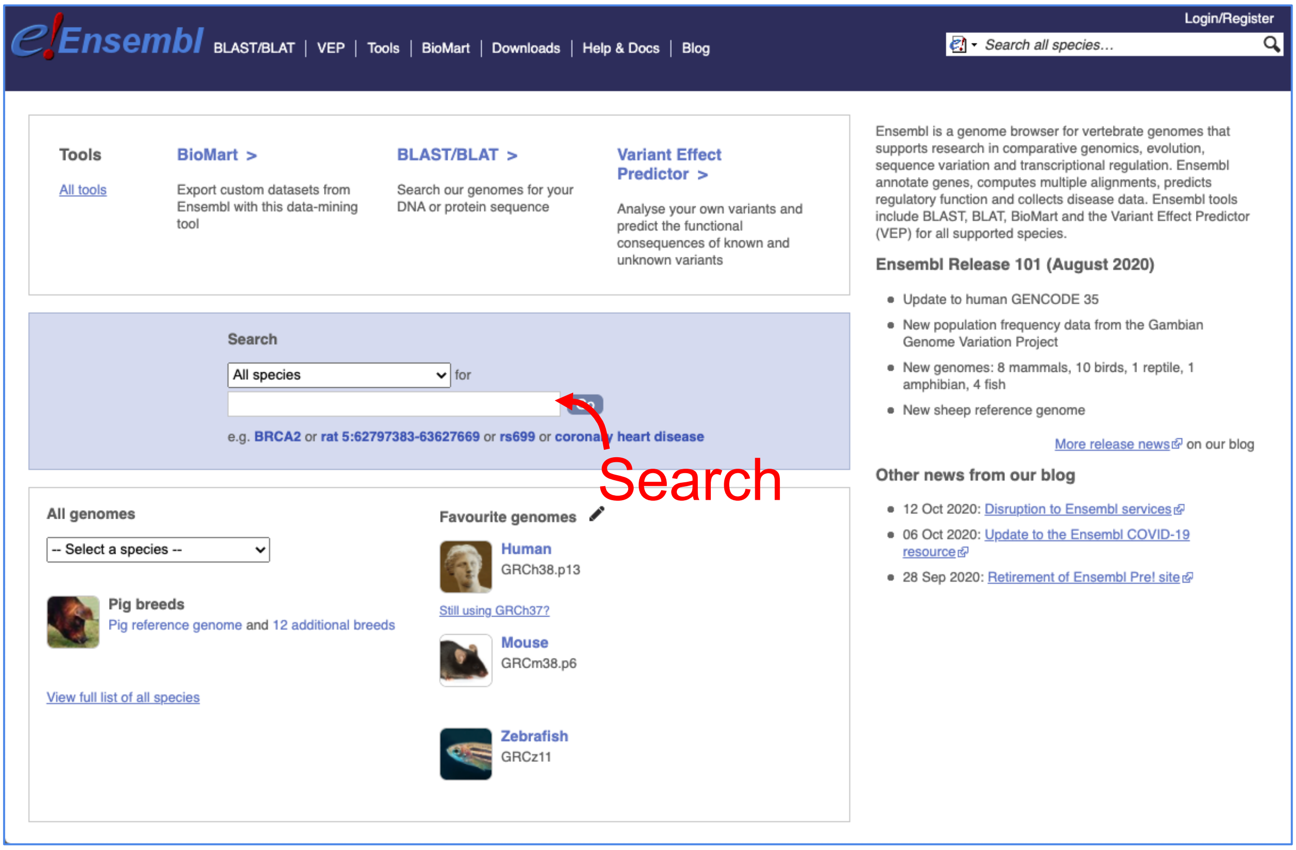
Our Animal DNA:

Comparing genes across the Tree of Life

Practical 1: find a protein sequence and run BLAST

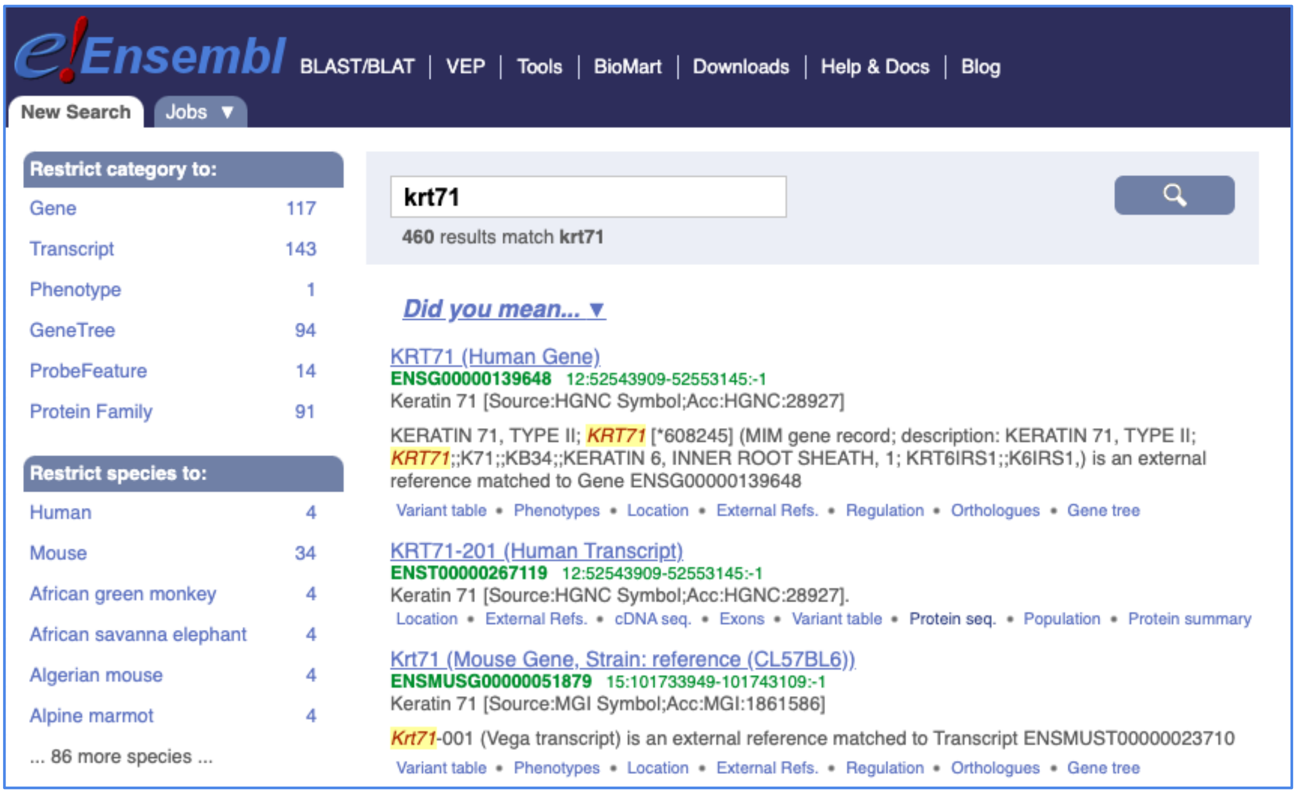
Step 1: Find the gene

We're going to find the protein sequence of the human gene *KRT71*. To find it, we're going to go to ensembl.org.

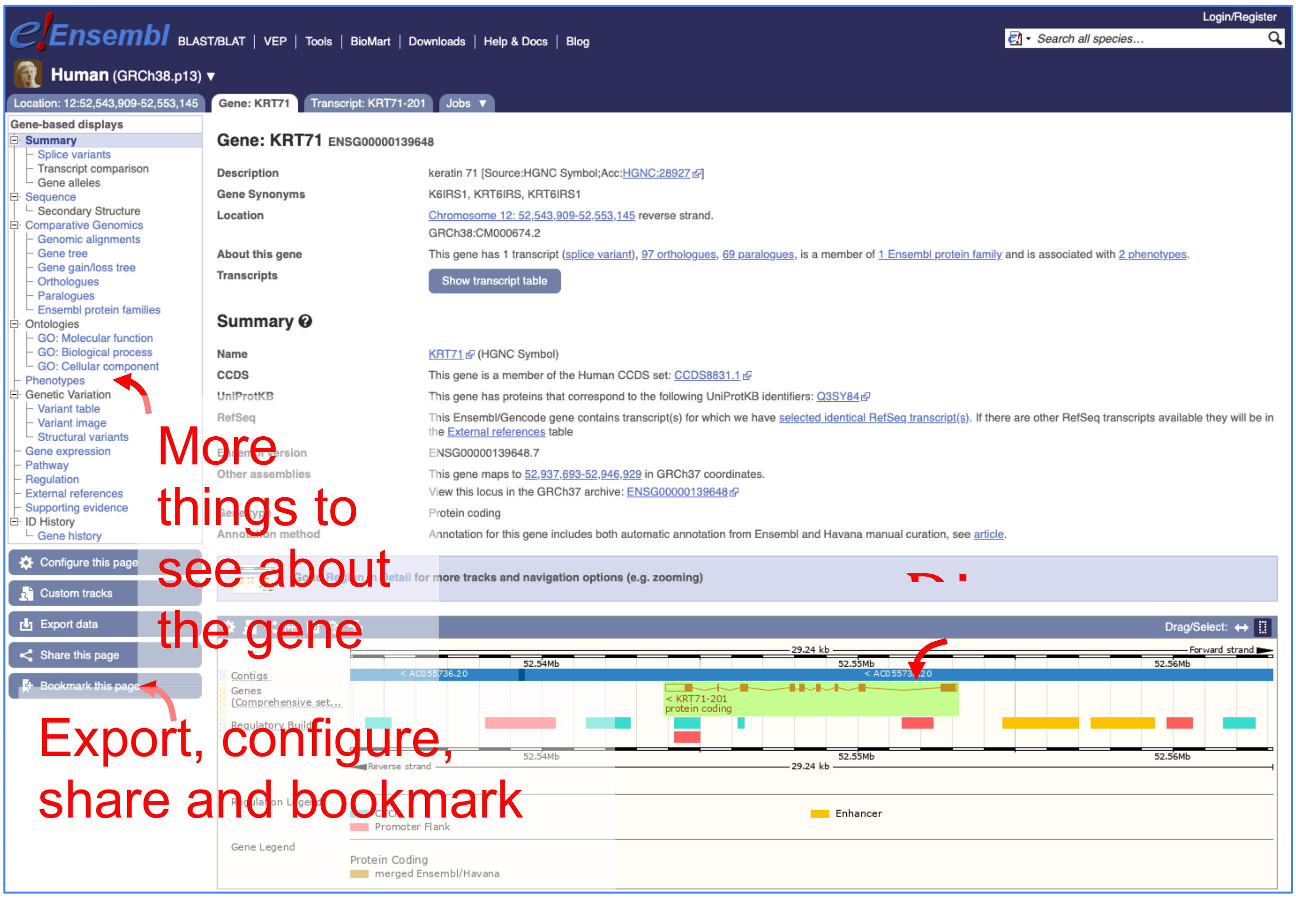


Type the gene name *KRT71* into the search box and hit Go.

Human *KRT71* should be your first search result.



Click on the top result to go the gene page, which looks like this:

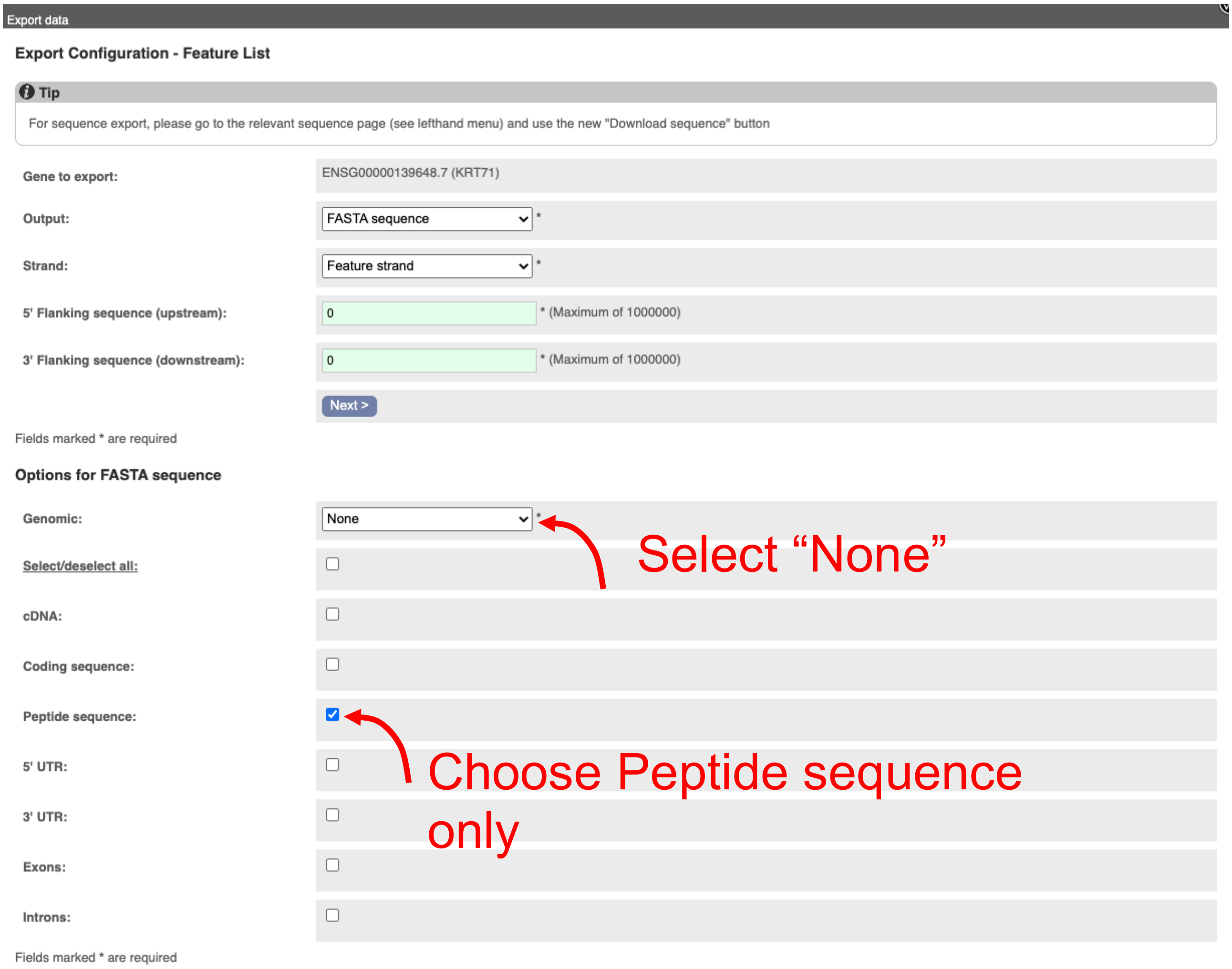


You can find out lots of information about the gene by clicking on some of the links on the left of the page, including its sequence, its orthologues, genetic variation between individuals, where the protein is found and processes the protein is involved in.

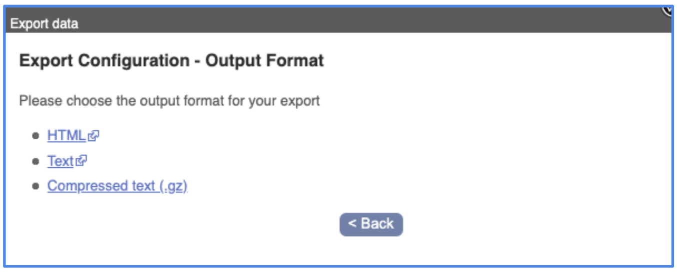
Step 2: Export the sequence

We want to export the protein sequence. Click on the blue button Export data (see above image).

The protein sequence is also known as the **peptide sequence**. Select peptide sequence on the page. You will also need to deselect any other selected options, including the Genomic sequence, where you need to select "None" from the drop-down. Then click Next>.



You can export in different formats, we'll go for Text.



This sequence is in a format called **FASTA**. FASTA is the standard way of writing sequences for both proteins and DNA. It consists of a header line at the top, which is indicated by a > sign at the beginning of the line and contains information about what the sequence is, and lines of sequence of fixed length (usually 60 bases or amino acids) underneath. This is what the sequence should look like:

Header

```
>ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding
MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLYSLGGVRSLNVA
SGSGKSGGYGFGRGRASGFAGSMFGSVALGPVCPTVCPPGGIHQVTVNESLLAPLNVELD
PEIQKVRAQEREQIKALNNKFASFIDKVRFLEQQNQVLETKWELLQQLDLNNCKNNLEPI
LEGYISNLRKQLETLSGDRVRLDSELRNVRDVVEDYKKRYEEEINKRTAAENEFVLLKKD
VDAAYANKVELQAKVESMDQEIKFFRCLFEAEITQIQSHISDMSVILSMDNNRNLDLDSI
IDEVRTQYEEIALKSKAEAEALYQTKFQELQLAAGRHGDDLKNTKNEISELTRLIQRIRS
EIENVKKQASNLETAIADAEQRGDNALKDARAKLDELEGALHQAKEELARMLREYQELMS
LKLALDMEIATYRKLLESEECRMSGEFPSPVSISIISSTSGGSVYGFRPSMVSGGYVANS
SNCISGVCSVRGGEGRSRGSANDYKDTLGKGSSLSAPSKKTSR
```

Amino acid 1-letter codes

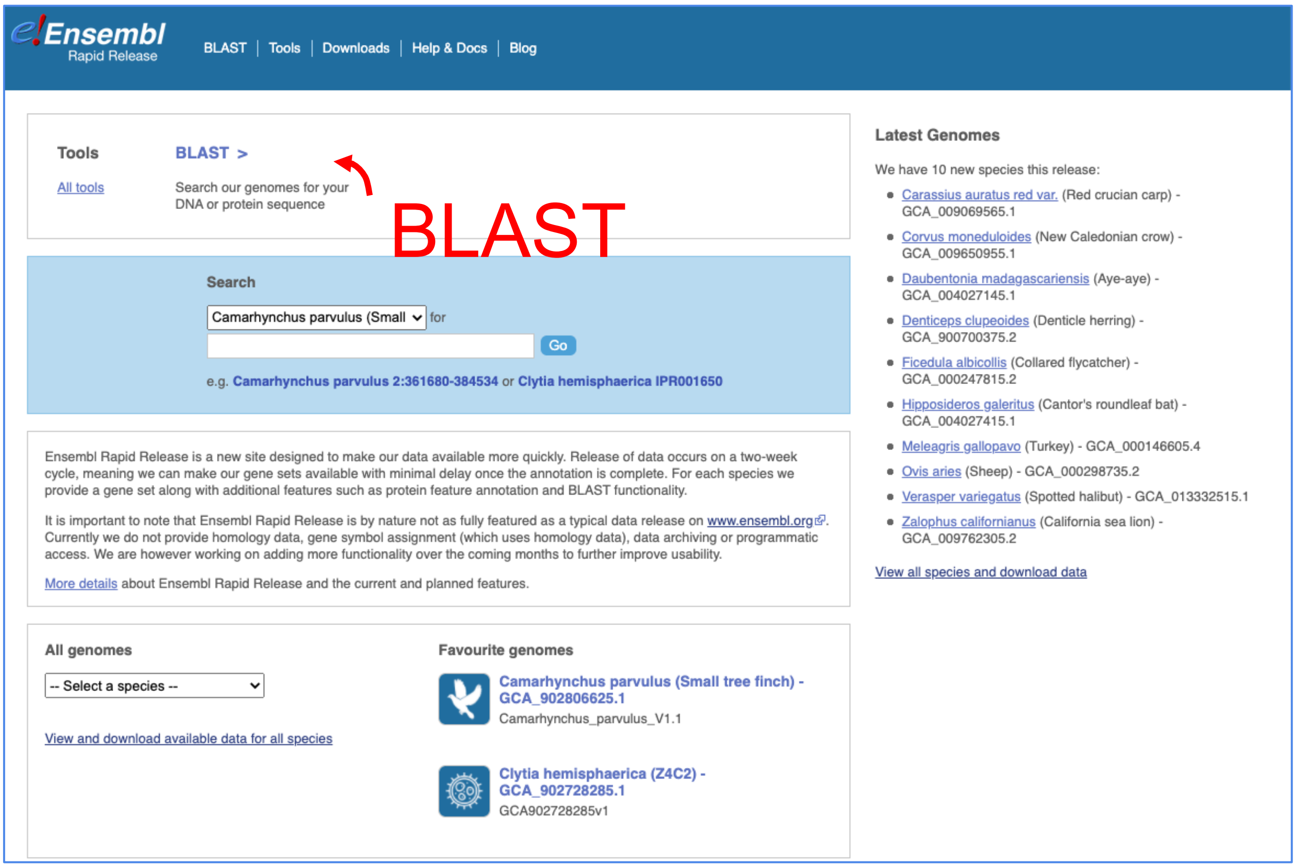
If you have something different, such as extra sequence beyond these 10 lines, go back to the export menu and make sure you've deselected some of the extra options, especially the Genomic option.

We can use this sequence in BLAST.

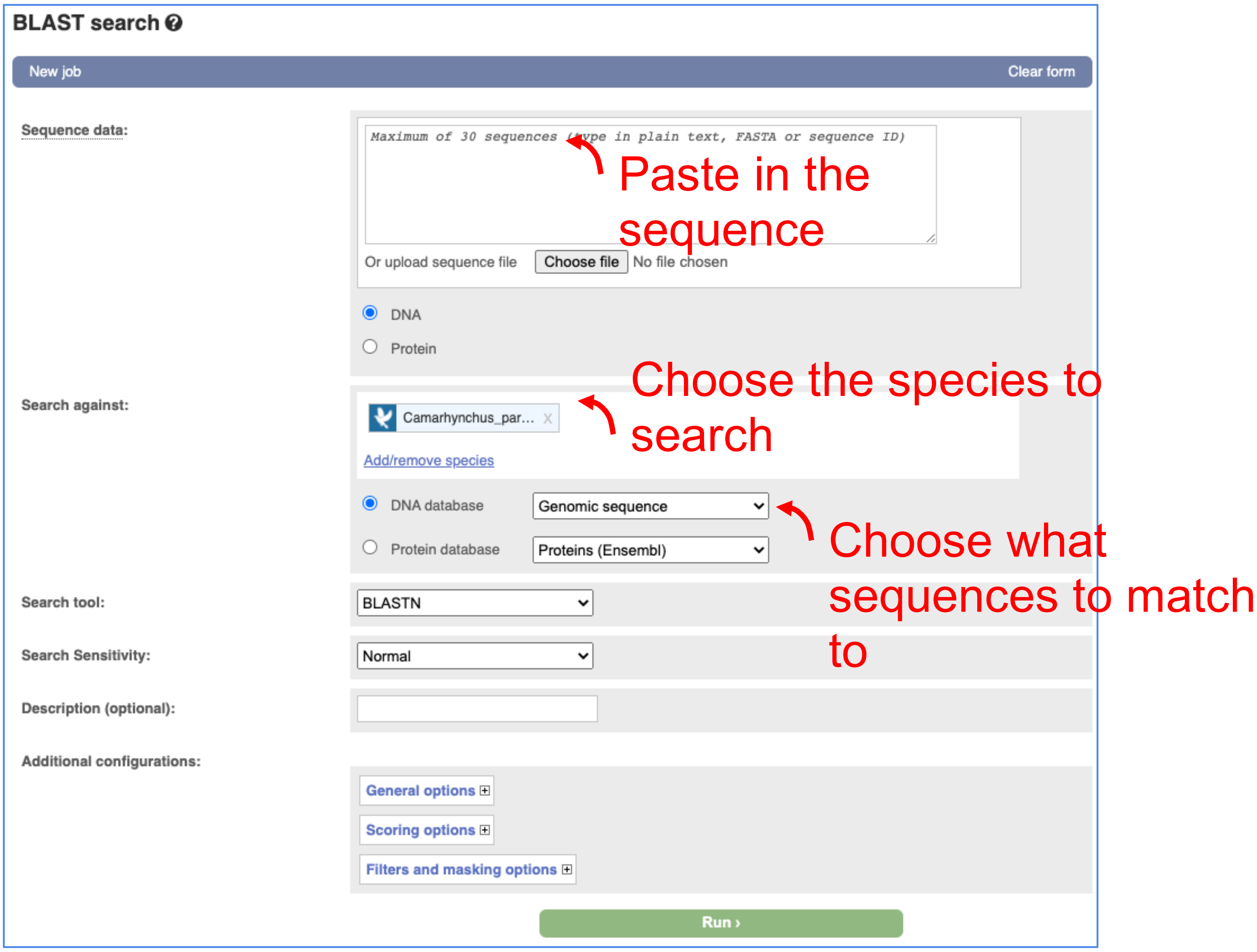
Step 3: Open BLAST

We're going to use this human gene again for Step 9, so keep this tab open.

Open a new tab in your browser and go to rapid.ensembl.org. We're going to use this sequence to find the *KRT71* gene in Darwin Tree of Life species.

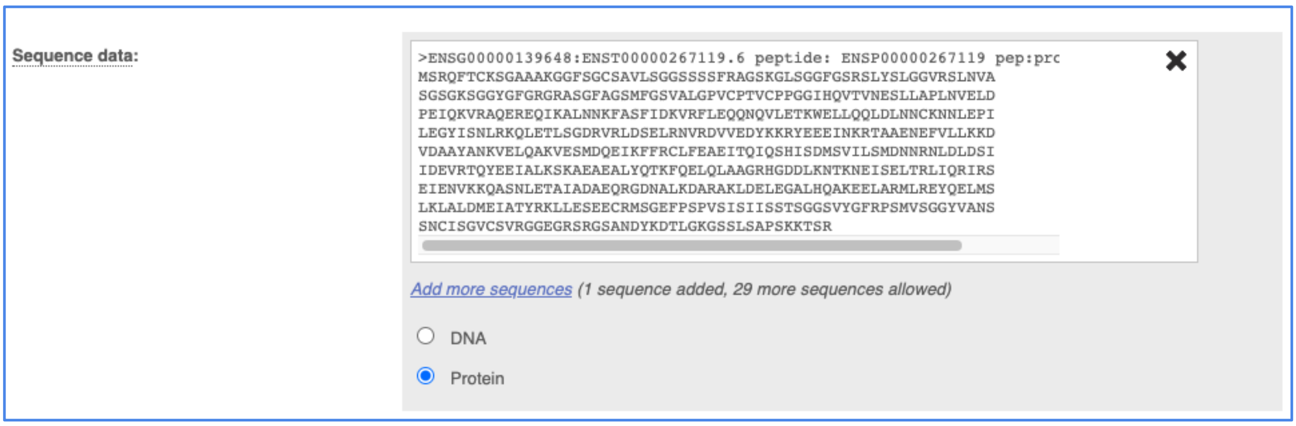


Click on BLAST to open it. This is what the interface looks like:



Step 4: add the sequence

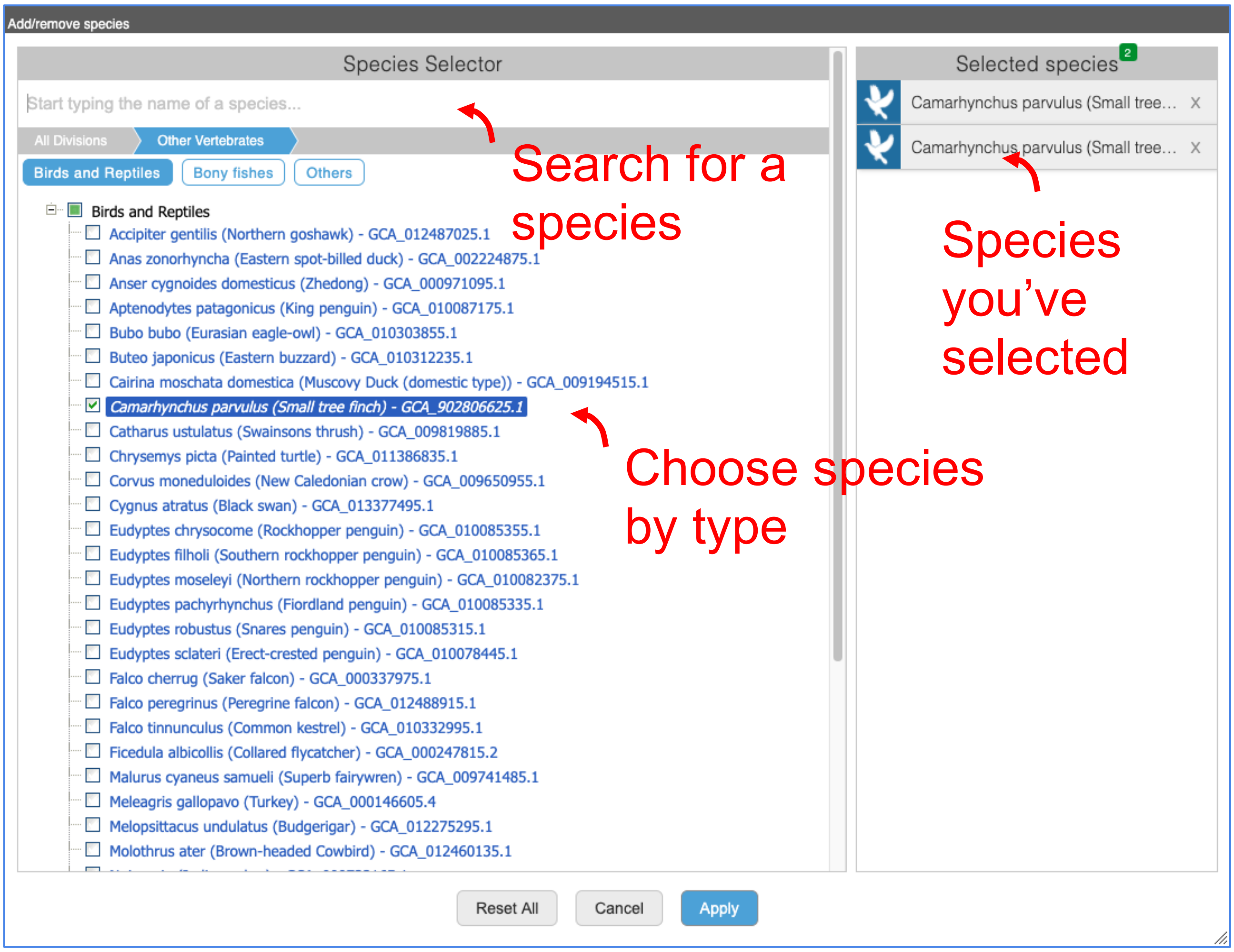
Copy and paste the sequence we found of human *KRT71* into the box. This should include the header line (starts with >) and all the lines of protein one-letter codes.



The tool has automatically detected that the sequence is a protein sequence.

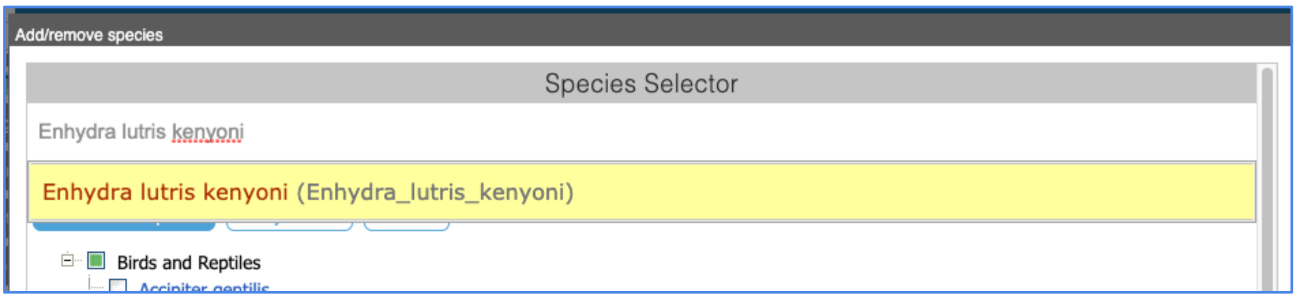
Step 5: Choose species to search against

Click on Add/remove species to choose some species.

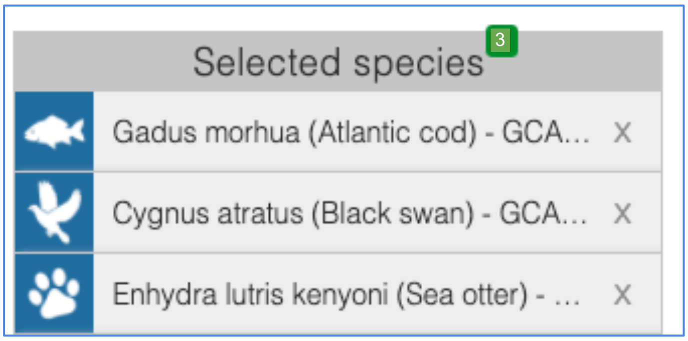


There is a finch species selected by default. You will need to remove this by clicking on the cross next to the species on the right.

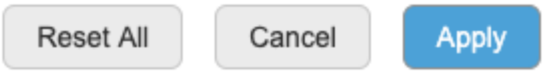
Find species by typing in their name. We're going to add Sea otter first. Its latin name is *Enhydra lutris kenyoni*, so type that into the search box.



Click on the species to add it to your list of species to search. Now find Black swan (*Cygnus atratus*) and Atlantic cod (*Gadus morhua*) in the same way. We now have a mammal, a bird and a fish to compare to.



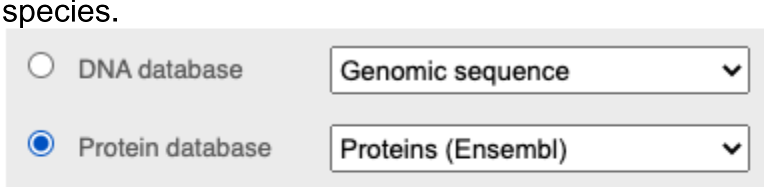
Click on Apply to select these species.



These species now appear in the BLAST interface.

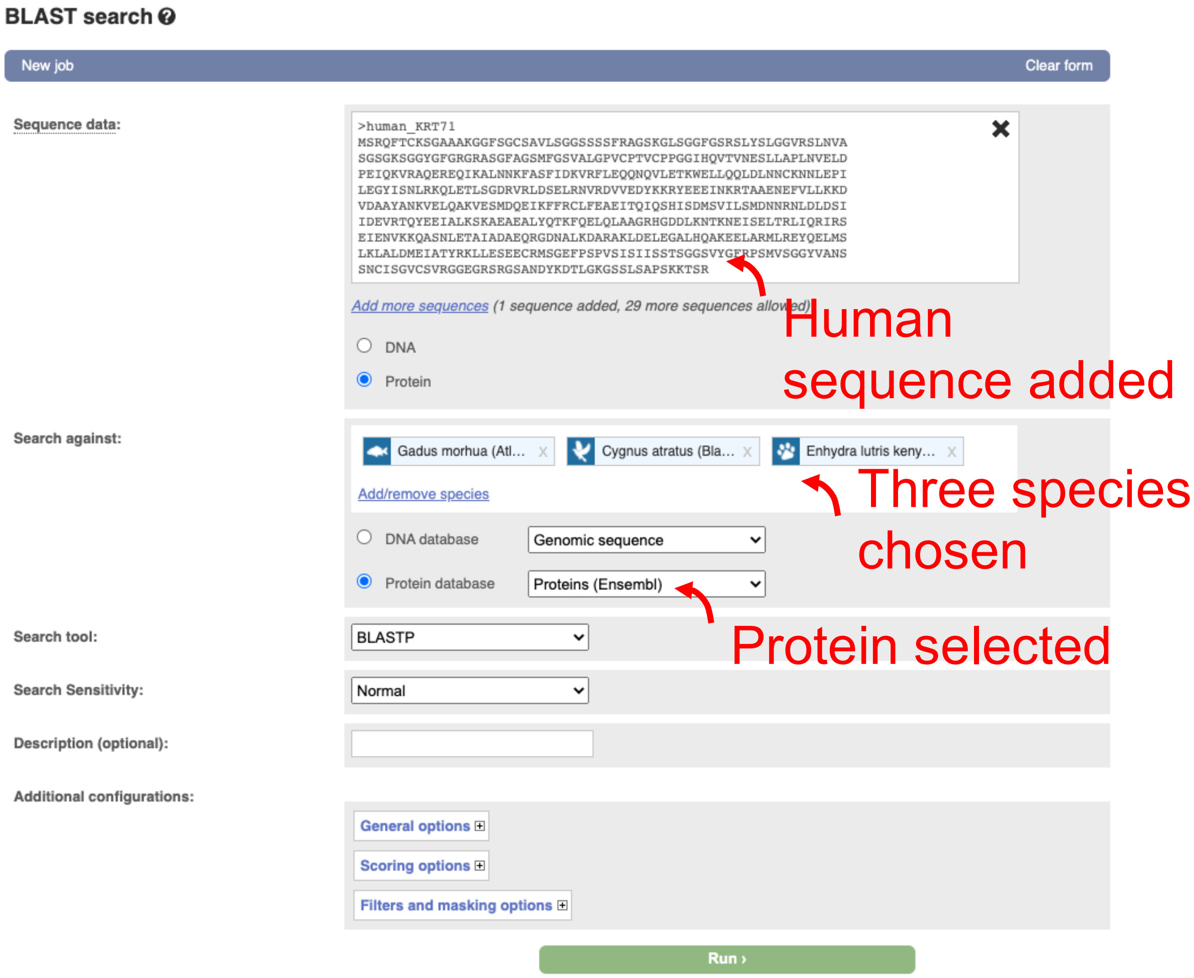
Step 6: choose the database to search

Select Protein database to compare the human protein sequence to proteins from these species.



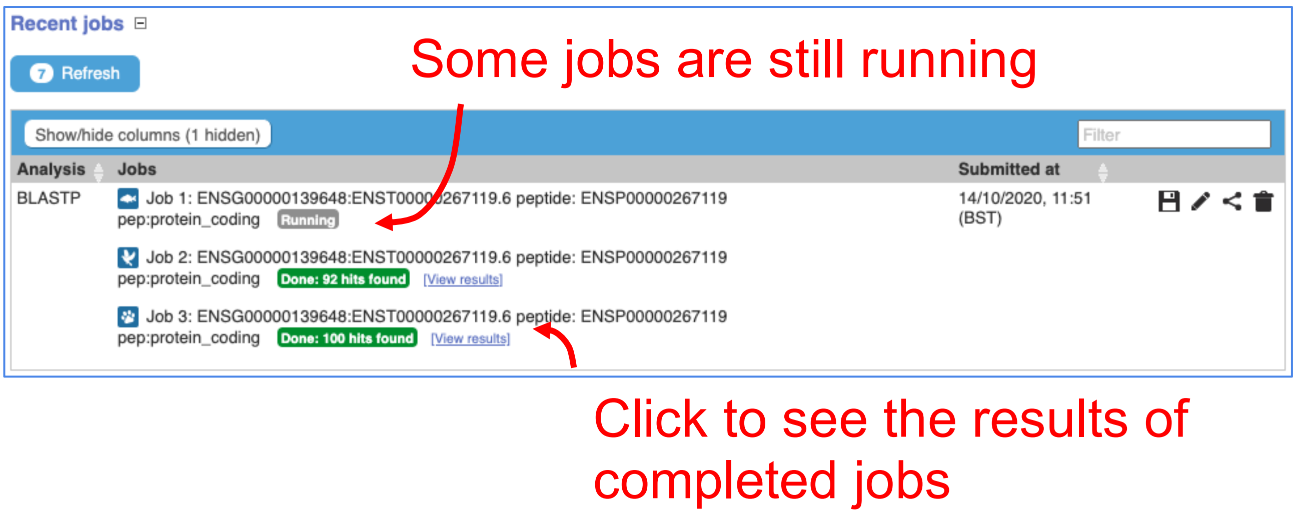
Step 7: Run BLAST

This is what your input form should look like now:



Click on the green Run button at the bottom to run BLAST.

As the job is running, you can see it in a table.



Step 8: view results

Choose the otter job first, indicated by the pawprint (mammal) icon. The bird and fish icons represent the black swan and Atlantic cod jobs respectively. Click on View results.

This will show you the results for this query. You will see a table listing all the otter proteins that match to the human protein. There's a picture of this on the next page.

There are lots of lines to this table. Each line represents one protein in sea otter which has some sequence similarity to the human protein sequence we used as input. We call each one a hit.

The table tells us what that sea otter protein is and where it is found in the sea otter genome. Anything in blue is a link. Feel free to click on the links to see the gene, protein, genomic region and sequence of the hits before you go back to explore the page more. If you click back into BLAST from any page, you'll get to your job table again.

The table tells us where in the human protein the similarity starts and ends (subject start and end) and where in the otter protein (query start and end).

The most important information is in the E-value and %ID (percentage identity) columns. The %ID is how much of the otter protein is the same as the human protein. You'll see that the top hit in the table is ENSELKP00000012757. It is 92.93% identical to the human protein.

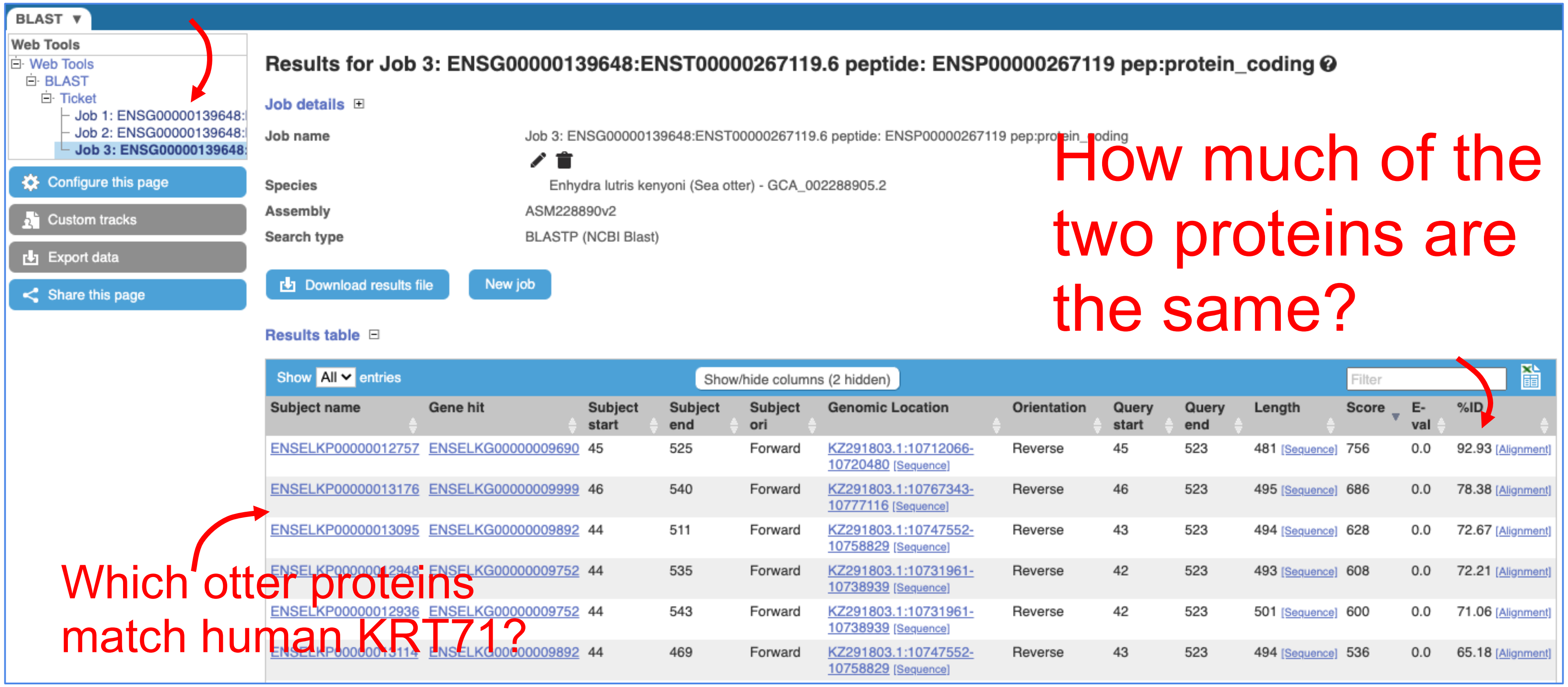
The e-value is a probability that the match is due to chance. This is calculated from the BLAST score, which adds value for matches and introduces penalties for gaps and mismatches. This is then combined with the full length and identity to determine how likely it is that this level of similarity could happen by chance. The smaller the number for the e-value, the better the hit. For the first six hits, the value is zero, but for the seventh it is 2e-177, which means the probability that this is due to chance is two times E (a mathematical constant around 2.718) to the power of -177, an infinitesimally small number. In practice, you need to consider this number that E is to the power of: the higher the number after the minus, the smaller the e-value and the better the hit.

The top hit is most likely to be the orthologue of the human gene. Other hits may be similar genes in the otter genome, perhaps paralogues. Paralogues are when a gene is copied in a genome, then the two versions of the gene change over time in the same species, so they can have similar but distinct functions.

Navigate to the
other jobs

How much of the two proteins are the same?

Which otter proteins match human KRT71?



BLAST ▼

Web Tools

- Web Tools
  - BLAST
    - Ticket
      - Job 1: ENSG00000139648:
      - Job 2: ENSG00000139648:
      - Job 3: ENSG00000139648:

Configure this page

Custom tracks

Export data

Share this page

Results for Job 3: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding

Job details

Job name	Job 3: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding
Species	Enhydra lutris kenyoni (Sea otter) - GCA_002288905.2
Assembly	ASM228890v2
Search type	BLASTP (NCBI Blast)

Download results file | New job

Results table

Show All entries | Show/hide columns (2 hidden) | Filter

Subject name	Gene hit	Subject start	Subject end	Subject ori	Genomic Location	Orientation	Query start	Query end	Length	Score	E-val	%ID
ENSELKP00000012757	ENSELKG00000009690	45	525	Forward	KZ291803.1:10712066-10720480 [Sequence]	Reverse	45	523	481 [Sequence]	756	0.0	92.93 [Alignment]
ENSELKP00000013176	ENSELKG00000009999	46	540	Forward	KZ291803.1:10767343-10777116 [Sequence]	Reverse	46	523	495 [Sequence]	686	0.0	78.38 [Alignment]
ENSELKP00000013095	ENSELKG00000009892	44	511	Forward	KZ291803.1:10747552-10758829 [Sequence]	Reverse	43	523	494 [Sequence]	628	0.0	72.67 [Alignment]
ENSELKP00000012948	ENSELKG00000009752	44	535	Forward	KZ291803.1:10731961-10738939 [Sequence]	Reverse	42	523	493 [Sequence]	608	0.0	72.21 [Alignment]
ENSELKP00000012936	ENSELKG00000009752	44	543	Forward	KZ291803.1:10731961-10738939 [Sequence]	Reverse	42	523	501 [Sequence]	600	0.0	71.06 [Alignment]
ENSELKP00000013114	ENSELKG00000009892	44	469	Forward	KZ291803.1:10747552-10758829 [Sequence]	Reverse	43	523	494 [Sequence]	536	0.0	65.18 [Alignment]

Take a look at black swan and Atlantic cod. Use the navigation on the left to go to the other jobs. Use the information you find to fill in the table below.

Species	Top hit identifier	% identity	E-value
Sea Otter	ENSELKP00000012757	92.93	0.0
Black Swan	ENSACYP00000008445	62.9	4e-170
Atlantic cod	ENSGMOP00000026268	58.31	9e-142

Questions:

Based on what you've done in the practical so far, consider:

1. Why do you think the proteins in sea otter, black swan and Atlantic cod have long codes instead of names?
2. The human protein is more similar to the sea otter protein than it is to the black swan and the Atlantic cod proteins. Why is this?

We'll go back to this BLAST job again in Practical 2, so leave this tab open.

Step 9: Learn about the human protein

Let's go back to the human protein to find its function. Go back to the human gene page you left open in another tab. If you lost this, go back to page one for instructions on how to find it again.

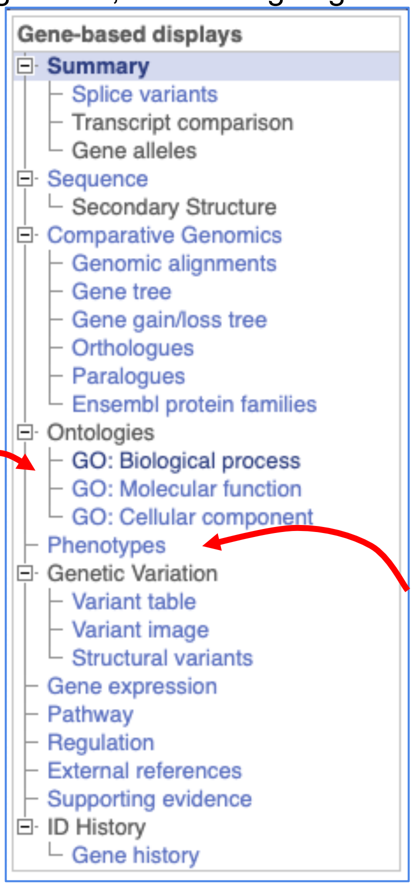
The description at the top of the page gives the full name of the gene: keratin 71. Have you heard of keratin before? Do you know where in the body you would find it?

Gene: KRT71 ENSG00000139648

Description keratin 71 [Source:HGNC Symbol;Acc:HGNC:28927]

The menu on the left-hand side of the page lists different things we can look at for the gene. There are lots of things listed, but we're going to explore GO terms and Phenotypes.

GO describes the function of the gene



Phenotypes describe the effect of altering this gene

Click on Phenotypes to see the effect of altering this gene. Phenotypes are observable differences between individuals. This is what the page will look like:

Effects seen in human

Phenotypes

Phenotypes, diseases and traits associated with this gene ENSG00000139648

Phenotype, disease and trait	Source
Hypotrichosis 13	MIM morbid
Woolly hair	Orphanet

Click on the data source to get descriptions of the effects and how they were discovered

Phenotype, disease and trait annotations associated with variants in this gene

Phenotype, disease and trait	Source(s)	Number of variants	Show/hide details
ALL variants with a phenotype annotation	-	4	Show
Annotated by HGMD	HGMD-PUBLIC	1	Show
Echocardiography	dbGaP	1	Show
Hypotrichosis 13	ClinVar	1	Show
INSULIN	dbGaP	1	Show

Phenotype, disease and trait annotations associated orthologues of this gene in other species

Effects seen in other species

Show All entries

Phenotype, disease and trait	Source	Species	Gene
waved hair	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71
whorled hair	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71
abnormal hair follicle inner root sheath morphology	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71
abnormal vibrissa morphology	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71
abnormal hair follicle morphology	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71
alopecia	MGI	Mouse (*Mus musculus*)	ENSMUSG00000051879 Krt71

There are two phenotypes linked to the gene itself: Woolly hair and Hypotrichosis 13. Can you click around to find out what Hypotrichosis 13 means? What do these phenotypes suggest about what the gene does? Note that some websites may ask for a donation, you can just dismiss this banner.

Go to GO: Biological process in the menu on the left. This will open a table listing the functions the protein has, like the one below:

GO: Biological process

Show/hide columns (1 hidden)

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0031069	hair follicle morphogenesis	IEA	Ensembl	ENST00000267119	• Search BioMart • View on karyotype
GO:0031424	keratinization	TAS	Reactome	ENST00000267119	• Search BioMart • View on karyotype
GO:0045109	intermediate filament organization	IMP	UniProt	ENST00000267119	• Search BioMart • View on karyotype
GO:0070268	cornification	TAS	Reactome	ENST00000267119	• Search BioMart • View on karyotype

We can see that the gene is linked to hair development.

Questions:

Based on what you've learned about the human protein:

3. Knowing *KRT71*'s involvement in hair follicles, now why do you think the fish and bird genes are so different?

4. Orthologues are genes that started out the same and diverged over time between species. Would you class these genes as orthologues?