

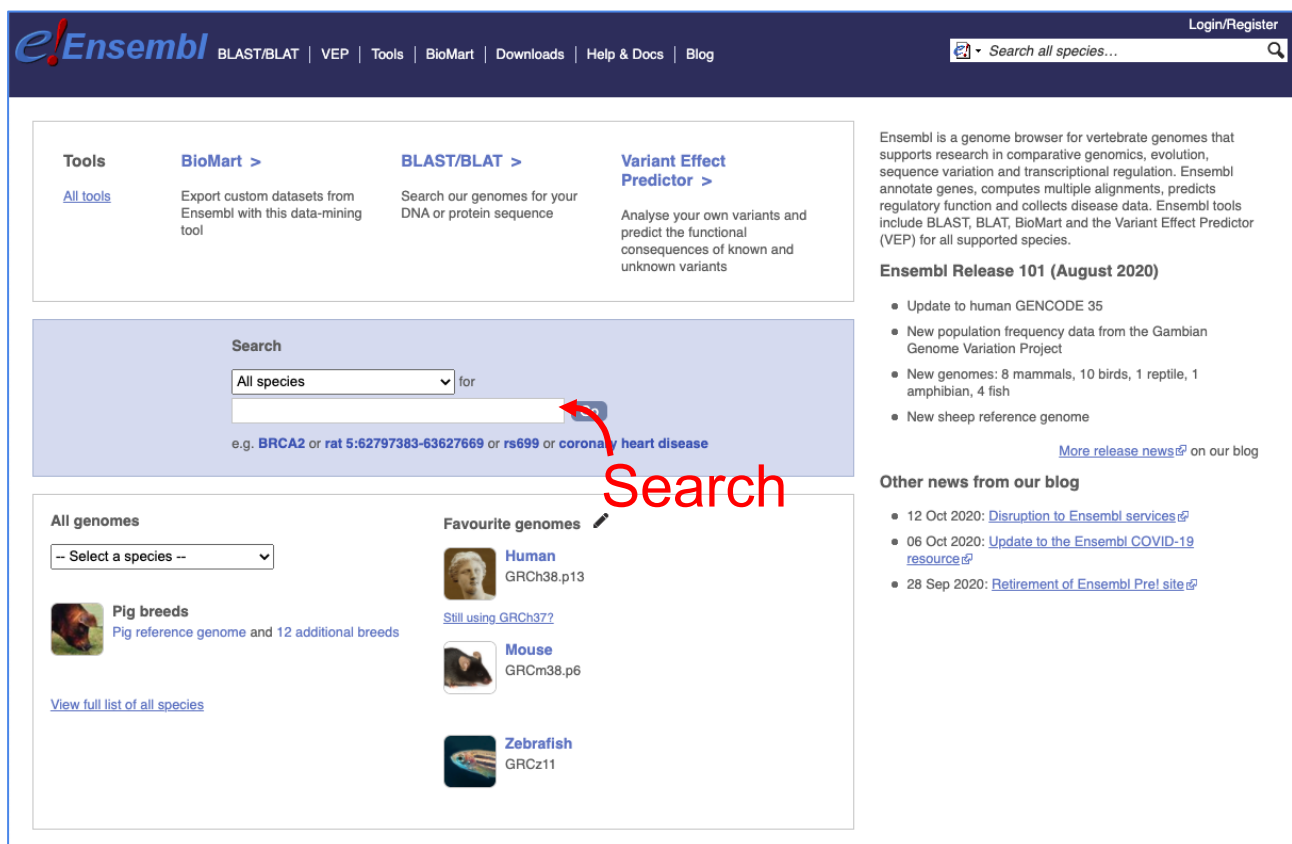
Our Animal DNA:

Comparing genes across the Tree of Life

Practical 1: find a protein sequence and run BLAST

Step 1: Find the gene

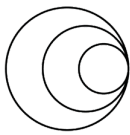
We're going to find the protein sequence of the human gene *KRT71*. To find it, we're going to go to ensembl.org.



The screenshot shows the Ensembl genome browser homepage. The search bar is highlighted with a red arrow and the word "Search" in red text. The search bar contains the text "All species" and "for". Below the search bar, there are examples of search terms: "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease". The page also features navigation links like "Tools", "BioMart", "BLAST/BLAT", and "Variant Effect Predictor". On the right, there is a section for "Ensembl Release 101 (August 2020)" with a list of updates and a "More release news" link. At the bottom, there is a section for "Other news from our blog" with a list of recent news items.

Type the gene name *KRT71* into the search box and hit [Go](#).

Human *KRT71* should be your first search result.



wellcome
connecting
science

EMBL-EBI



e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

New Search Jobs

Restrict category to:

- Gene 117
- Transcript 143
- Phenotype 1
- GeneTree 94
- ProbeFeature 14
- Protein Family 91

Restrict species to:

- Human 4
- Mouse 34
- African green monkey 4
- African savanna elephant 4
- Algerian mouse 4
- Alpine marmot 4
- ... 86 more species ...

Search: **krt71** 460 results match krt71

[Did you mean... ▼](#)

KRT71 (Human Gene)
ENSG00000139648 12:52543909-52553145:-1
Keratin 71 [Source:HGNC Symbol;Acc:HGNC:28927]
KERATIN 71, TYPE II; **KRT71** [*608245] (MIM gene record; description: KERATIN 71, TYPE II; **KRT71**::K71::KB34::KERATIN 6, INNER ROOT SHEATH, 1; KRT6IRS1::K6IRS1,) is an external reference matched to Gene ENSG00000139648
[Variant table](#) • [Phenotypes](#) • [Location](#) • [External Refs.](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

KRT71-201 (Human Transcript)
ENST00000267119 12:52543909-52553145:-1
Keratin 71 [Source:HGNC Symbol;Acc:HGNC:28927].
[Location](#) • [External Refs.](#) • [cDNA seq.](#) • [Exons](#) • [Variant table](#) • [Protein seq.](#) • [Population](#) • [Protein summary](#)

Krt71 (Mouse Gene, Strain: reference (CL57BL6))
ENSMUSG00000051879 15:101733949-101743109:-1
Keratin 71 [Source:MGI Symbol;Acc:MGI:1861586]
Krt71-001 (Vega transcript) is an external reference matched to Transcript ENSMUST00000023710
[Variant table](#) • [Phenotypes](#) • [Location](#) • [External Refs.](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

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

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▼

Location: 12:52,543,909-52,553,145 Gene: KRT71 Transcript: KRT71-201 Jobs

Gene: KRT71 ENSG00000139648

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

Gene Synonyms: K6IRS1, KRT6IRS, KRT6IRS1

Location: [Chromosome 12: 52,543,909-52,553,145 reverse strand.](#)
GRCh38:CM000674.2

About this gene: This gene has 1 transcript ([splice variant](#)), [97 orthologues](#), [69 paralogues](#), is a member of [1 Ensembl protein family](#) and is associated with [2 phenotypes](#).

Transcripts: [Show transcript table](#)

Summary

Name: [KRT71](#) (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: [CCDS8831.1](#)

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: [Q3SY84](#)

RefSeq: This Ensembl/Gencode gene contains transcript(s) for which we have [selected identical RefSeq transcript\(s\)](#). If there are other RefSeq transcripts available they will be in the [External references](#) table

ENSG00000139648.7

This gene maps to [52,937,693-52,946,929](#) in GRCh37 coordinates.

View this locus in the GRCh37 archive: [ENSG00000139648](#)

Protein coding: Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see [article](#).

More things to see about the gene

Export, configure, share and bookmark

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Contigs: < AC055736.20 52.54Mb 29.24 kb 52.55Mb 52.56Mb

Genes (Comprehensive set...): < KRT71-201 protein coding

Regulatory Build: < KRT71-201 protein coding

Regulation Legend: Promoter Flank Enhancer

Gene Legend: Protein Coding merged Ensembl/Havana

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>](#).

Export data

Export Configuration - Feature List

Tip

For sequence export, please go to the relevant sequence page (see lefthand menu) and use the new "Download sequence" button

Gene to export:

ENSG00000139648.7 (KRT71)

Output:

FASTA sequence *

Strand:

Feature strand *

5' Flanking sequence (upstream):

0 * (Maximum of 1000000)

3' Flanking sequence (downstream):

0 * (Maximum of 1000000)

Next >

Fields marked * are required

Options for FASTA sequence

Genomic:

None *

Select/deselect all:

☐

cDNA:

☐

Coding sequence:

☐

Peptide sequence:

☒

5' UTR:

☐

3' UTR:

☐

Exons:

☐

Introns:

☐

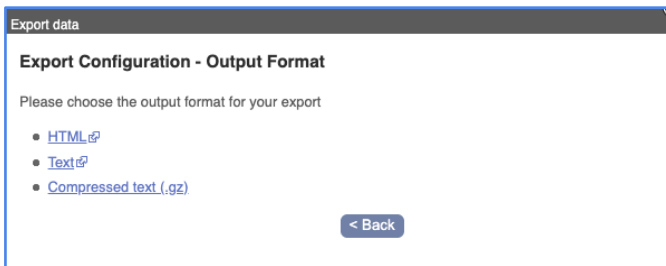
Fields marked * are required

Select "None"

Choose Peptide sequence only

You can export in different formats, we'll go for [Text](#).

yourgenome.org



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
 MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLSLGGVRSINVA
 SSGSGKSGGYGFRGRASGFAGSMFGSVALGPVCPTVCPGGIHQVTVNESLLAPLNVELD
 PEIQKVRQEREQIKALNNKFASFIDKVRFLQQNQVLETKWELLQQDLNCKNNLEPI
 LEGYISNLRKQLETLSGDRVRLDSELNRNVRDVVEDYKKRYEEEINKRTAAENEFVLLKKD
 VDAAYANKVELQAKVESMDQEIKFFRCLFEAEITQIQSHISDMSVILSMDNNRNLDSL
 IDEVRTQYEEIALKSKAEAEALYQTKFQELQLAAGRHGDDLKNTKNEISELTRLIQIRIS
 EIENVKKQASNLETAIADAEQRGDNALKDARAKLDELEGALHQAEEELARMLREYQELMS
 LKLALDMEIATYRKLLSEECRMSGFPPSPVSIISIISTSGGSVYGFPRPSMVSGGYVANS
 SNCISGVCSVRGGEGRSRGSANDYKDTLKGSSLSAPSCKTSR

← **Amino acid
1-letter
codes**

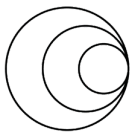
If you have something different, such as extra sequence beyond these 10 lines, go back to the export menu and make sure you've deselected some of the extra options, especially the **Genomic** option.

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.



Tools

[All tools](#)

BLAST >

Search our genomes for your
DNA or protein sequence

BLAST

Search

Camarhynchus parvulus (Small ▾) for

Go

e.g. [Camarhynchus parvulus 2:361680-384534](#) or [Clytia hemisphaerica IPR001650](#)

Ensembl Rapid Release is a new site designed to make our data available more quickly. Release of data occurs on a two-week cycle, meaning we can make our gene sets available with minimal delay once the annotation is complete. For each species we provide a gene set along with additional features such as protein feature annotation and BLAST functionality.

It is important to note that Ensembl Rapid Release is by nature not as fully featured as a typical data release on www.ensembl.org. Currently we do not provide homology data, gene symbol assignment (which uses homology data), data archiving or programmatic access. We are however working on adding more functionality over the coming months to further improve usability.

[More details](#) about Ensembl Rapid Release and the current and planned features.

All genomes

-- Select a species -- ▾

[View and download available data for all species](#)

Favourite genomes



Camarhynchus parvulus (Small tree finch) -
GCA_902806625.1
Camarhynchus_parvulus_V1.1



Clytia hemisphaerica (Z4C2) -
GCA_902728285.1
GCA902728285v1

Latest Genomes

We have 10 new species this release:

- [Carassius auratus red var.](#) (Red crucian carp) - GCA_009069565.1
- [Corvus moneduloides](#) (New Caledonian crow) - GCA_009650955.1
- [Daubentonia madagascariensis](#) (Aye-aye) - GCA_004027145.1
- [Denticeps clupeioides](#) (Denticle herring) - GCA_900700375.2
- [Ficedula albicollis](#) (Collared flycatcher) - GCA_000247815.2
- [Hipposideros galeritus](#) (Cantor's roundleaf bat) - GCA_004027415.1
- [Meleagris gallopavo](#) (Turkey) - GCA_000146605.4
- [Ovis aries](#) (Sheep) - GCA_000298735.2
- [Verasper variegatus](#) (Spotted halibut) - GCA_013332515.1
- [Zalophus californianus](#) (California sea lion) - GCA_009762305.2

[View all species and download data](#)

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

BLAST search ⓘ

New job Clear form

Sequence data:

Maximum of 30 sequences (type in plain text, FASTA or sequence ID)

Paste in the sequence

Or upload sequence file Choose file No file chosen

Search against:

Cammarhynchus_par... X

Choose the species to search

[Add/remove species](#)

☒ DNA database Genomic sequence ▼

☐ Protein database Proteins (Ensembl) ▼

Choose what sequences to match to

Search tool: BLASTN ▼

Search Sensitivity: Normal ▼

Description (optional):

Additional configurations:

General options ⊕

Scoring options ⊕

Filters and masking options ⊕

Run ›

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.

Sequence data:

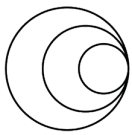
>ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:prc
MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLYSLGGVRSNLVA
SGSGKSGGYGFRGRASGFAGSMFGSVALGPVCPTVCPGGIHQVTNVESSLAPLNVELD
PEIQKVRQEREQIKALNNKFASFIDKVRFLQONQVLETKWELLQQLDNNCKNNLEPI
LEGYISNLRKQLETLSGDRVRLDSELNRNVDVVEDYKKRYEEEINKRTAAENEFVLLKGD
VDAAYANKVELQAKVESMDQEIKFRCLEAEITQISHISDMSVILSMNNRNLDDLSI
IDVVRTQYEEIALKSKAEAEALYQTKFQELQLAAGRHDGDLKNTKNEISELTRLIQRIS
EIEENVKQASNLETAIADAEQRGDNALKDARAKLDELEGALHQAEEELARMLREYQELMS
LKLALDMEIATYRKLLSEECRMSGEFFSPVSIISSTSGGSVYGFPRSMVSGGYVANS
SNCISGVCSVRGGEGRSGSANDYKDTLKGSSLSAPSKKTSR

[Add more sequences](#) (1 sequence added, 29 more sequences allowed)

☐ DNA

☒ Protein

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.

Search for a species




Species you've selected

Choose species by type

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.

Selected species 3	
	Gadus morhua (Atlantic cod) - GCA... X
	Cygnus atratus (Black swan) - GCA... X
	Enhydra lutris kenyon (Sea otter) - ... X

Click on [Apply](#) to select these species.

These species now appear in the BLAST interface.

Search against:

 Gadus morhua (Atl... X
  Cygnus atratus (Bla... X
  Enhydra lutris ken... X

[Add/remove species](#)

Step 6: choose the database to search

Select [Protein database](#) to compare the human protein sequence to proteins from these species.

☐ DNA database

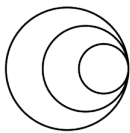
Genomic sequence ▼

☒ Protein database

Proteins (Ensembl) ▼

Step 7: Run BLAST

This is what your input form should look like now:



BLAST search

New job Clear form

Sequence data:

```
>human_KRT71
MSRQFTCKSGAAAKGGFSGCSAVLSGGSSSSFRAGSKGLSGGFGSRSLYSLGGVRSNLVA
SGSGKSGGYGFRGRASGFAGSMFGSVALGPVCPVPGGIHQVTNLSLLAPLNVELD
PEIQKVRQEREQIKALNNKFASFIDKVRFLQQNQVLETKWELLQQLDNNCKNNLEPI
LEGYISNLKQLETLSGDRVRLDSELNRVDDVEDYKKRYEEEINKRTAENEFVLLKLD
VDAAYANKVELQAKVESMDQEIKFRCLEAEITQIQSHISDMSVILSMDNNRNLDSI
IDVVRTQYEEIALKSKAEAEALYQTKFQELQAAGRHDGLKNTKNEISELTRLIQIRIS
EINVKKQASNLETAIAADAEQRGNALDKARAKLELEGALHQAKEELARMLREYQELMS
LKIALDMEIATYRKLLSEECRMSGEFPSPVSIISSTSGGSVYGFPSMVSGGYVANS
SNCISGVCVSRGGEGRSGSANDYKDTLKGSSLSAPSCKTSR
```

[Add more sequences](#) (1 sequence added, 29 more sequences allowed)

☐ DNA
☒ Protein

Search against:

Gadus morhua (Atl...
 Cygnus atratus (Bla...
 Enhydra lutris keny...

[Add/remove species](#)

☐ DNA database
☒ Protein database

Genomic sequence
Proteins (Ensembl)

Search tool: BLASTP

Search Sensitivity: Normal

Description (optional):

Additional configurations:

[General options](#)
[Scoring options](#)
[Filters and masking options](#)

[Run](#)

Human sequence added

Three species chosen

Protein selected

Click on the green [Run](#) button at the bottom to run BLAST.



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

Recent jobs

7 Refresh

Show/hide columns (1 hidden) Filter

Analysis	Jobs	Submitted at
BLASTP	Job 1: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding Running	14/10/2020, 11:51 (BST)
	Job 2: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding Done: 92 hits found View results	
	Job 3: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding Done: 100 hits found View results	

Some jobs are still running

Click to see the results of completed jobs

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 [ENSELKPP0000012757](#). 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

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: ENSG00000139648:ENST00000267119.6 peptide: ENSP00000267119 pep:protein_coding

Species: Enhydra lutris kenyonii (Sea otter) - GCA_002288905.2

Assembly: ASM228890v2

Search type: BLASTP (NCBI Blast)

Download results file New job

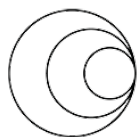
Results table

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	Reverse	45	523	481	756	0.0	92.93
ENSELKP00000013176	ENSELKG00000009999	46	540	Forward	KZ291803.1:10767343-10777116	Reverse	46	523	495	686	0.0	78.38
ENSELKP00000013095	ENSELKG00000009892	44	511	Forward	KZ291803.1:10747552-10758829	Reverse	43	523	494	628	0.0	72.67
ENSELKP00000012948	ENSELKG00000009752	44	535	Forward	KZ291803.1:10731961-10738939	Reverse	42	523	493	608	0.0	72.21
ENSELKP00000012936	ENSELKG00000009752	44	543	Forward	KZ291803.1:10731961-10738939	Reverse	42	523	501	600	0.0	71.06
ENSELKP00000013114	ENSELKG00000009892	44	469	Forward	KZ291803.1:10747552-10758829	Reverse	43	523	494	536	0.0	65.18

Which other proteins match human KRT71?

How much of the two proteins are the same?



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.

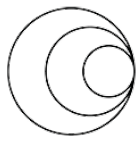
Species	Top hit identifier	% identity	E-value
Sea Otter	ENSELK P00000012757	92.93	0.0
Black Swan	ENSACYP00000008445	62.9	4e-170
Atlantic cod	ENSGMOP000000026268	58.31	9e-142

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?

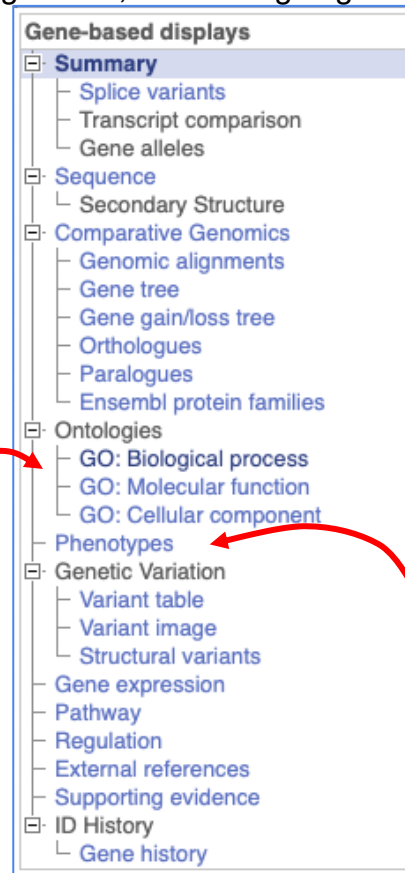
Gene: KRT71 ENSG00000139648

Description

keratin 71 [Source:HGNC Symbol;Acc:[HGNC:28927](#)]

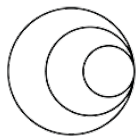
The menu on the left-hand side of the page lists different things we can look at for the gene. There are lots of things listed, but we're going to explore GO terms and Phenotypes.

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 ⓘ

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

Phenotype, disease and trait	Source	Species	Gene
waved hair	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71
whorled hair	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71
abnormal hair follicle inner root sheath morphology	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71
abnormal vibrissa morphology	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71
abnormal hair follicle morphology	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71
alopecia	MGI ⓘ	Mouse (<i>Mus musculus</i>)	ENSMUSG000000051879 Krt71

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

Effects
seen in
other
species

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)				Filter	
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?