Topic 5: Genes and genomes

Created	@November 9, 2023 7:00 PM
Reviewed	☐

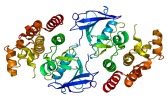
In this practical we will use the **Ensembl Comparative Genomics** data in four different exercises, from the analysis of homologs and protein families, to whole genome alignments:

1. **Orthologues, paralogues and gene trees** for the human *BRAF* gene
2. **Sequence conservation and constraint** for the human *RHO* gene
3. **Synteny** between human and horse in the region of the *RHO* gene
4. **Whole genome alignments** in the region of the human *BRCA2* gene

Start at the Ensembl website to complete these exercises. Answers to the exercises are prepared for using Ensembl version 110 (June 2023).

Exercise 1. Orthologues, paralogues and protein families for the human *BRAF* gene

BRAF is a human gene that makes a protein called **B-Raf**. The gene is also referred to as *proto-oncogene B-Raf* and *v-Raf murine sarcoma viral oncogene homolog B*, while the protein is more formally known as serine/threonine-protein kinase B-Raf. The B-Raf protein is involved in sending signals inside cells, which are involved in directing cell growth. In 2002, it was shown to be faulty (mutated) in some human cancers. Certain other inherited *BRAF* mutations cause birth defects. Drugs that treat cancers driven by *BRAF* mutations have been developed. Two of these drugs, vemurafenib and dabrafenib, are approved by FDA for treatment of late-stage melanoma.



(a) How many orthologues are predicted for this gene in primates?

Go to www.ensembl.org, choose *human* and search for *BRAF*. Click on *Orthologues*

(link below the first entry) to see all the orthologous genes. There are 23 primate species in the Ensembl database. For how many of these, there exists a 1:1 orthologue for the human *BRAF* gene?

ANS: 22

(b) How much sequence identity does the tarsier (*Carlito syrichta*) protein have to the human one?

Click on the *Show details* checkbox for Primates. A table with details for all primate orthologues will appear below. Look for the tarsier orthologue protein by the end of the table. The column *Target %id* indicates the percentage of the target sequence (Tarsier B-Raf) matching the query sequence (human B-Raf), while the column *Query %id* indicates the reverse (the percentage of the query sequence matching the target sequence). Note the difference in the values of the Target and Query %id; how would you explain this difference?

What percentage of the Tarsier B-Raf protein is identical to the human B-Raf protein? Enter your answer below (whithout the "%" symbol).

Click the link *Compare Regions* in the *Orthologue* column. You should see an alignment between the human *BRAF* gene region and the *BRAF* gene region for the tarsier.

If you go back to the previous page and click on the *View Sequence Alignments -> View Protein Alignment* link in the *Orthologue* column, you will be able to view a protein alignment in ClustalW format.

ANS:

(c) Is there a genomic alignment between marmoset (*Callithrix jacchus*) and human for the region surrounding the *BRAF* gene?

Look for the marmoset orthologue in the human *BRAF* orthologues table.

Choose all options that apply from below: ANS:

(d) Number of copies of the *BRAF* gene

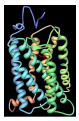
Once in the *Gene: BRAF* tab, click on the *Gene gain/loss tree* link on the left side menu. Numbers indicate the number of copies of the gene in each genome.

Which species from the clade Otomorpha (clade containing Herrings) has the highest number of copies of the *BRAF* gene in its genome?

ANS: goldfish

Exercise 2. Sequence conservation and constraint for the human *RHO* gene

Rhodopsin (**RHO**), also known as visual purple, from Ancient Greek ῥόδον (rhódon, “rose”), due to its pinkish color, and ὄψις (ópsis, “sight”), is a light-sensitive receptor protein. It is a biological pigment in photoreceptor cells of the retina. Rhodopsin is the primary pigment found in rod photoreceptors. Rhodopsins belong to the G-protein-coupled receptor (GPCR) family. They are extremely sensitive to light, enabling monochromatic vision in low-light conditions. Exposed to light, the pigment immediately photobleaches, and it takes about 45 minutes to regenerate fully in humans.



(a) The *RHO* gene in 10 primates

Go to www.ensembl.org, choose human and search for *RHO*. Click on the tab *Location* (first tab on the left). Could you tell where this gene is located in the human genome?

Click on *Alignments (Image)* in the left side menu and click on the button *Select an Alignment,* then *Multiple* and select the *10 primates EPO* alignment from the *Multiple* option. Click Apply.

Do all 10 primates show a gene in these alignments? Which primate genomes are represented and have an annotated *RHO* gene in the alignment? Select all that apply from below.

The same results can also be seen in the *Alignments (text)* page. Click on *Alignments (text)* in the left side menu. Aligned sequences are highlighted in blue.

You can export the alignments from either *Alignments (images)* or *Alignments (text)* pages. Click on the blue *Download alignment/data* button, and choose CLUSTALW from the list.

ANS:

(b) Sequence conservation and constraint of the *RHO* gene

Comparisons of orthologous genomic DNA sequences can be used to characterize regions that have been subject to purifying selection and are enriched for functional elements. Genomic Evolutionary Rate Profiling (**GERP**), identifies **constrained elements** in multiple alignments by quantifying substitution deficits. These deficits represent substitutions that would have occurred if the element were neutral DNA, but did not occur because the element has been under functional constraint. We refer to these deficits as *Rejected Substitutions*, which are a natural measure of constraint that reflects the strength of past purifying selection on the element.

In a given alignment, sites are scored independently. **Positive scores** represent a substitution deficit (i.e., fewer substitutions than the average neutral site) and thus indicate that a site may be under evolutionary constraint. Positive scores scale with the level of constraint, such that the greater the score, the greater the level of evolutionary constraint inferred to be acting on that site. **Negative scores** indicate that a site is probably evolving neutrally; negative scores should not be interpreted as evidence of accelerated rates of evolution because of too many strong confounders, such as alignment uncertainty or rate variance.

Click on *Region in detail* in the left side menu. Then click *Configure this page* in the left side menu. Click on *Conservation regions* under the *Comparative genomics* menu. Check to select:

- *Conservation score for 91 eutherian mammals EPO-Extended*
- *Constrained elements for 91 eutherian mammals EPO-Extended*
- *Conservation score for 65 amniota Mercator-Pecan*
- *Constrained elements for 65 amniota Mercator-Pecan*

Click on the tick on the top right corner.

Take a look at the results and answer the following questions:

- What do the *GERP scores* mean?

- Which *GERP score* values take exons?
- Which regions of the gene do most of the constrained element blocks match up to?

Select all options that apply from below: ANS:

Exercise 3. Synteny between human and horse in the region of the *RHO* gene

We continue with the *RHO* gene example in this exercise, but in this case we are going to examine syntenic regions with the horse chromosomes for this region.

(a) Syntenic regions between human and horse chromosomes

Click *Synteny* at the left hand menu. Then change the species to **Horse** next to the image and click **Go**.
Are there any syntenic regions between human and horse? Which chromosomes in horse have syntenic regions to human chromosome 3? Select all that apply from below: ANS:

(b) Other genes on the syntenic block between humans and horse where the *RHO* gene is located

Stay in the *Synteny* view but scroll down to the bottom of the page. Is there a homologue in horse for human *RHO*? What is it named? Where in the horse genome is located?

Compare the genes between human and horse in this syntenic block. Are there more genes in this syntenic block with homologues? You can display a comparison of this region between human and horse by clicking *Region Comparison*.

Select all options that apply from below:

ANS: The neighbors of the RHO gene are: IFT122 (upstream) and H1-8 (downstream), both in humans and in horses

Exercise 4. Whole genome alignments in the region of the human *BRCA2* gene

BRCA2 is a human tumor suppressor gene (specifically, a caretaker gene), found in all humans. Its protein, also called by the synonym **breast cancer type 2 susceptibility protein**, is responsible for repairing DNA. *BRCA2* and *BRCA1* are normally expressed

in the cells of breast and other tissue, where they help repair damaged DNA or destroy cells if DNA cannot be repaired. They are involved in the repair of chromosomal damage with an important role in the error-free repair of DNA double strand breaks. If *BRCA1* or *BRCA2* itself is damaged by a BRCA mutation, damaged DNA is not repaired properly, and this increases the risk for breast cancer. The human reference *BRCA2* gene contains 28 exons, and the cDNA has 10,254 base pairs coding for a protein of 3418 amino acids.

(a) *BRCA2* in the human genome

Go to www.ensembl.org, choose human and search for *BRCA2*. Click on the *Location* tab (first tab on the left). Where is this gene located in the human genome?

ANS: In the long arm of chromosome 13, on the forward strand

(b) Conservation of *BRCA2* between humans and anole lizard, chicken, chimpanzee, mouse, platypus, and zebrafish

Click *Configure this page* in the left side menu. Click on *BLASTZ/LASTz alignments* under the *Comparative genomics* menu on the left. Check to activate the following alignments:

- *Abingdon island giant tortoise (Chelonoidis abingdonii) - LASTZ net -> Normal*
- *Chimpanzee (Pan troglodytes) – LASTZ net -> Normal*
- *Mouse (Mus musculus) – LASTZ net -> Normal*
- *Platypus (Ornithorhynchus anatinus) - LASTZ net -> Normal*
- *Zebrafish (Danio rerio) - LASTZ net -> Normal*

Click on the tick on the top right corner and look for the activated conservation tracks in the image.

Does the degree of conservation between human and the various other species reflect their evolutionary relationship? Which parts of the *BRCA2* gene seem to be the most conserved? Did you expect this? Select from below all options that apply:

ANS:

(c) Conservation of the *BRCA2* gene in 91 eutherian mammals and 65 amniota vertebrates

Click *Configure this page* in the side menu. Click on *Conservation regions* under the *Comparative genomics* menu. Check to select:

- *Conservation score for 91 eutherian mammals EPO-Extended*
- *Constrained elements for 91 eutherian mammals EPO-Extended*
- *Conservation score for 65 amniota Mercator-Pecan*
- *Constrained elements for 65 amniota Mercator-Pecan*

Click on the tick on the top right corner.

Have a look at the *Conservation scores* and *Constrained elements* tracks for the set of 91 eutherian mammals and the set of 65 amniota vertebrates. Do these tracks confirm what you already saw in the tracks with pairwise alignment data? Choose all options that apply from below:

ANS:

(d) Retrieve the genomic alignment for a constrained element

Click on a constrained element (one of the blocks in the *Constrained elements for 65 amniota vertebrates Mercator-Pecan* track). A pop-up window will appear with more information about the
element, including its coordinate positions in the genome, length, score and statistical significance. Click on *View alignments (text)* in the pop-up menu. Take a look at the output page.

Is there an alignment for all 65 species for this region? Why? Choose the most appropriate answer from below.

Click *Configure this page* in the side menu. Check the option *Show conservation regions*. Click on the tick. Bases that match in >50% of the species in the alignment will be highlighted in light blue.

ANS: