Zebrafish Open Door Workshop – Madison 2018

Module 2 Exercise Answers

Comparative Genomics

Exercise C1 – Zebrafish orthologues

(a) Start in the Location tab (region in detail) for *sardh*. Click on Alignments (Image) at the left, and select the 11 fish EPO alignment in the pull-down menu in the view.
The zebrafish, cave fish, cod, tilapia, Amazon molly, platyfish, spotted gar, stickleback, medaka, fugu, and tetradon are shown in this region. All the species show a gene in the aligned region. This can also be seen in the Alignments (text) page (the exons are highlighted in red).

(b) You can export the alignments from either the Alignments (text) or Alignments (image) pages in the Location tab. Click on the blue Download alignments button at the top of the text page, or the icon at the top of the image, and choose ClustalW from the list.

(c) Click on Region in detail in the left hand menu. Turn on the multiple alignment, constrained elements and conservation score for 11 fish EPO tracks, all under the Comparative genomics menu by configuring the page.

The 11 fish EPO track just shows that the whole region for the *sardh* gene can be aligned among those eleven species of fish. The Constrained elements and Conservation score tracks show the conserved sequence is located where in the alignment.

Higher conservation regions match up with exonic regions (exons tend to be highly conserved) of the gene.

Hover over the the Track name and the ⓘ (information button) to read more about constrained elements (or any other data track).

Exercise C2 – Orthologues, paralogues and gene trees for the human *BRAF* gene.

(a) Go to www.ensembl.org, choose human and search for *BRAF*. Click through to the Gene tab view.

On the gene tab, click on Orthologues at the left side of the page to see all the orthologous genes.
There are orthologues in 10 primates.

The percentage of identical amino acids in the Tarsier protein (the orthologue) compared with the gene of interest, i.e. human *BRAF* (the target species/gene), is 69%. This is known as the Target %ID. The identity of the gene of interest (human *BRAF*) when compared with the orthologue (Tarsier *BRAF*, the query species/gene) is 62% (the query %ID).

Note that the difference in the values of the Target and Query % ID reflects the different protein lengths for the human and tarsier *BRAF* genes.

(b) There is more than one way to get to the answer.
Option 1: Go to the orthologues page and click on the marmoset orthologue to open the gene tab.
Click Genomic alignments at the left. Then select Alignment: Human (Homo sapiens) – lastz and click Go. Choose Block 1 to get the largest block of aligned sequence.
The red sequence is present in exons, so there is a gene in both species in this region. You can find where the start and stop codons are located if you configure this page and select START/STOP codons.

Option 2: Go to location tab of the marmoset *BRAF* gene and then click on Region Comparison view at the left. Click on Select species or regions at the left and click on the + to select Human (Homo sapiens) – lastz then save and close. You should see an alignment between the human *BRAF* gene region and the *BRAF* gene region for the marmoset.

(**Note**: To see a blue line connecting homologous genes in the Region Comparison view page, click on Configure this page and under Comparative features select Join genes. Zoom out on the location view to see blue lines connecting all the homologous genes between marmoset and human genes in that region).

Exercise C3 – Whole genome alignments

(a) Go to the Ensembl homepage (http://www.ensembl.org/).
Select Search: Zebrafish and type brca2 in the search box.
Click Go.
Click on 15: 31911989-31928519 below BRCA2.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:
Click Configure this page in the side menu.
Click Reset configuration.
SAVE and close.

(b) Click Configure this page in the side menu
Click on BLASTZ/LASTZ alignments under the Comparative genomics menu. Select Cave fish (Astyanax mexicanus) - LASTZ net, Cod (Gadus morhua) – LASTZ net, Human (Homo sapiens) – LASTZ net, Medaka

(Oryzias latipes) -LASTZ net and Stickleback (Gasterosteus aculeatus) - LASTZ net.
SAVE and close.
Yes, the degree of conservation does reflect the evolutionary relationship between human and the other species; the highest degree of conservation is found in cavefish, followed by the other fish, and finally human. Especially the exonic sequences of *BRCA2* seem to be highly conserved between the various species, which is what is to be expected because these are supposed to be under higher selection pressure than intronic and intergenic sequences.

(c) Click Configure this page in the side menu.
Click on Conservation regions under the Comparative genomics menu.
Select Conservation score for 11 fish EPO_LOW_COVERAGE, and Constrained elements for 11 fish EPO_LOW_COVERAGE.
SAVE and close.
Both the Conservation score and Constrained elements tracks largely correspond with the data seen in the pairwise alignment tracks; all exons of the *brca2* gene show a high degree of conservation (Note the UTRs which are not conserved).

(d) Click on a constrained element (brown block).
Click on View alignments (text) in the pop-up menu.
Click Configure this page in the side menu.
Select Show conservation regions.
SAVE and close.

The conserved regions will be shown in light blue.

Variation

Exercise V1 – Exploring a SNP in zebrafish

(a) Please note there is more than one way to get this answer. Either go to the Variation Table for the zebrafish *ift88* gene, and Filter variants to the missense variants, or search Ensembl for rs180018766 directly.

(b) Once you're in the Variation tab, click on the Genes and regulation link or icon.

This SNP is found in four transcripts from two genes (ENSDART00000110943, ENSDART00000132032, and ENSDART00000038923 from ENSDARG00000027234 and ENSDART00000018251 from ENSDARG00000010700). It is a missense variant in two of these transcripts, an intron variant in one of the transcripts and an upstream variant of the fourth transcript.

(c) From the **Genes and Transcript consequences** table, you can see that rs180018766 is an upstream variant of transcript ENSDART00000018251 from the ENSDARG00000010700 (interleukin 17d) gene.

(d) Select Population genetics from the side menu.
From the Frequency data table, the Sanger Stemple submission shows that C is the major allele (50.6% of the population) compared to T (49.4% of the population). Note that T is the reference allele, irrespective of population prevalence.

Exercise V2 – Exploring variants of the zebrafish gene tbx16

(a) Go to www.ensembl.org. Type tbx16 in the Search box for zebrafish, then click Go. Select variant table from the side menu of the *tbx16* gene page.

Select splice region variant from the consequence type filter.

(b) Select rs180149499.
The reported alleles are A/C/T. A is the reference allele. C and T are both reported variants.

(c) In Ensembl, the allele that is present in the reference genome assembly is always put first (A is the allele for the reference zebrafish genome.

(d) The variant is located on Chromosome 8, coordinates 51753117.

(e) Click on HGVS names to reveal information about HGVS nomenclature. This SNP has got four HGVS names (two for each variant allele):

Variant allele C

- NC_007119.7:g.51739646A>C
- ENSDART00000007090.9:c.953+6T>G

Variant allele T

- NC_007119.7:g.51739646A>T
- ENSDART00000007090.9:c.953+6T>A

Exercise V3 – Human population genetics and phenotype data

(a) Please note there is more than one way to get this answer. Either go to the Variation Table for the human *TAGAP* gene, and Filter variants to the 5'UTR, or search Ensembl for rs1738074 directly.

Once you're in the Variation tab, click on the Genes and regulation link or icon.
This SNP is found in three transcripts (ENST00000326965, ENST00000338313, and ENST00000367066).

(b) Click on Population genetics at the left of the variation tab. (Or, click on Explore this variation at the left and click the Population genetics icon.)
In Yoruba (HapMap-YRI population), the least frequent genotype is CC at the frequency of 9.7%. This is also the least frequent genotype in other populations (to find out what the three letter populations are, hover over the names).

(c) Click on Phylogenetic context.
The ancestral allele is T and is inferred from the alignment in primates.

Select the 40 eutherian mammals EPO LOW COVERAGE alignment and click on Go.
A region containing the SNP (highlighted in red and placed in the centre) and its flanking sequence are displayed. The T allele is conserved in all but two of the 40 eutherian mammals displayed. Note that two species have no alignment in that region and many other species have no variation database.

(d) Click Phenotype Data at the left of the Variation page.
This variation is associated with multiple sclerosis and coeliac. There are known risk alleles for both multiple sclerosis and coeliac and the

corresponding P values are provided. The allele A is associated with coeliac disease. Note that the alleles reported by Ensembl are T/C. Ensembl reports alleles on the forward strand. This suggests that A was reported on the reverse strand in the original paper. Similarly, one of the alleles reported for Multiple sclerosis is G.

Exercise V4 – Exploring a SNP in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/).

Type rs1801133 in the Search box, then click Go.
Click on rs1801133.

(b) Click on Genes and Regulation in the side menu (or the Genes and Regulation icon).
No, rs1801133 is Missense variant in four *MTHFR* transcripts. It's a downstream gene variant of ENST00000418034.

(c) In Ensembl, the alleles of rs1801133 are given as G/A because these are the alleles in the forward strand of the genome. In the literature and in dbSNP, the alleles are given as C/T because the *MTHFR* gene is located on the reverse strand. The alleles in the actual gene and transcript sequences are C/T.

(d) Click on Population genetics in the side menu.
In all populations but two (from the 1000 genomes and HapMap projects), the allele G is the major one. The two exceptions are: CLM (Colombian in Medellin; 1000 Genomes), HCB (Han Chinese in Beijing, China; HapMap).

(e) Click on Phenotype Data in the left hand side menu.
The specific studies where the association was originally described is given in the Phenotype Data table. Links between rs1801133 and homocysteine levels were described in two papers. Click on the pubmed IDs pubmed:20031578 and pubmed:23824729 for more details.

(f) Click on Phylogenetic Context in the side menu.

Select Alignment: 8 primates EPO and click Go.
All eight primates, including human, have a G in this position.

VEP

Exercise V5 – VEP

Go to www.ensembl.org and click on the Variant Effect Predictor link on the homepage. Click Launch VEP.

Choose zebrafish as the species. Enter the three variants as below:

9 33147350 33147350 A/G 1
9 33148752 33148752 T/C 1
9 33147257 33147257 T/C 1

Note: Variation data input can be done in a variety of formats. See more details here
http://www.ensembl.org/info/docs/variation/vep/vep_formats.html

Click Run.

When your job is listed as Done, click View Results.

You will get a table with the consequence terms from the Sequence Ontology project (http://www.sequenceontology.org/) (i.e. synonymous, missense, downstream, intronic, 5' UTR, 3' UTR, etc) provided by VEP for the listed SNPs. You can also upload the VEP results as a track and view them on Location pages in Ensembl.

All three variants affect the same three transcripts. The affected transcripts are: ENSDART00000141849, ENSDART00000143453, ENSDART00000005879.

One variant is a novel variant but the other two have previously been described (rs180140691 and rs40727775) in dbSNP.