

Bioinformatics: Life Sciences on Your Computer

Practice Problems - 2

1. Go to the Zebrafish home page in Ensembl.
 - a. In the search box, enter 1:8365000-8385000 (a region of chromosome 1). How many Ensembl genes are listed in Region in detail (each box is a gene)?
 - b. Look for the box for "uncx4.1." Click Drag-select at the upper right and then draw a box with your mouse around the uncx4.1 box. Then click Jump to region. Look at the bottom graphic for the detail. How many exons are in the uncx4.1 gene?
 - c. Is the gene on the forward or reverse strand?
 - d. Go to Configure this page (left sidebar). Turn on the Prediction Transcripts track to add the Genscan prediction track. Does it match the known gene?
Try to export and view the image as a PDF (click right icon on bottom).
2. In Ensembl, find the *Caenorhabditis elegans* lin-14 gene.
 - a. How many different transcripts are listed?
 - b. How can the shorter transcript produce a longer protein?
 - c. List the charge and isoelectric point for each protein variant.
 - d. Click the Gene tab and click Protein families. How many members of this protein family are listed?
3. Exploring a gene in Ensembl:
 - a. Find the human F9 (Coagulation factor IX) gene. Open the Ensembl gene. On which chromosome and which strand of the genome is this gene located?
 - b. On what contig does this gene reside (blue bar)?
 - c. How many transcripts (splice variants) has Ensembl annotated for it?
 - d. What is the longest Ensembl transcript?
 - e. How many exons does it have?
 - f. How long is the protein it encodes?
 - g. Click the transcript identifier for the longest transcript and then "Exons" on the left menu. Which exons are either completely or partially untranslated (purple is untranslated, black is translated)?
 - h. Click the Protein Summary Link. In which part of the protein (N- or C-) encoded by ENST00000218099 does its catalytic peptidase activity, responsible for the cleavage of factor X to yield its active form factor Xa, reside?
 - i. Click Domains and Features for the protein. List the descriptions of the three Pfam hits that list descriptions.
 - j. What missense (nonsynonymous) SNP is found at amino acid 39 (Variation Table: left menu)? List the accession number, the nucleotide change and the amino acid change.
 - k. Go back to the Gene tab and click Orthologues. Is there a mouse ortholog predicted for the human F9 gene?
 - l. How many transcripts are there for the mouse F9 gene?
4. Go to OMIM.org and find the record with accession number +132810 to access an epoxide hydrolase record. On the right menu, click Allelic Variants (Under Table of Contents) and look at variants .0001 and .0002.
 - a. What are the amino acid changes of the first two variants (include position)?
 - b. Which variant shows a definitive correlation between mutant alleles and diminished enzymatic

activity?

- c. Smith and Harrison found out what about the slow phenotype and its relation to COPD?
- d. Click the dbSNP link by allelic variant .0002. What is listed for clinical significance?
- e. Click the ENSP link for 355802. Find allelic variant .001 in the sequence. How many nonsense (stop gained) SNPs are shown?
- f. Which (SIFT or PolyPhen) lists allelic variant 1 as "deleterious?"

5. Go to the human EPHX1 gene in [Ensembl](#).

- a. Click Variation Table on the side menu. How many synonymous SNPs are listed?
- b. How many missense SNPs are listed?
- c. Show all missense SNPs. Looking only at the first 50 amino acids, which SNPs at which locations are predicted by SIFT or PolyPhen to be potentially damaging (look for orange or red)?
- d. How many of the five known transcripts code for protein?
- e. How many paralogues are listed for this gene?

Answers

- 1. a. ~19
 - b. 3
 - c. forward strand (carat points right and above blue bar)
 - d. Yes, it matches perfectly--gene predictors do not predict UTRs.
2. a. 2
- b. The longer one can have longer UTRs
 - c. 5.0/7.4586 & 3.0/7.0861
 - d. 9
3. a. X Chromosome, forward (or +) strand
- b. AL033403.1
 - c. 3
 - d. F9-001 at 2780 bp
 - e. 8
 - f. 461 aa
 - g. Two--the first and last
 - h. C-terminus (right side)
 - i. EG-like domain; GLA_domain; Pepsinogen_s1
 - j. rs200505648, A/G, K/R (Note: K/R changes are common--arginine and lysine are both positively charged amino acids)
 - k. Yes
 - l. One transcript
4. a. Y113H or Tyr113His; H139R or His139Arg
- b. the first
 - c. Slow phenotype 4 to 5 times more common in patients with COPD
 - d. non-pathogenic
 - e. 5
 - f. SIFT

- 5.a. 134
- b. 233
- c. 32 and 49
- d. 4
- e. 0