

Assignment 5 / Biometry and bioinformatics I / 2013

Deadline for submission: 31.12.2013

If you are not very familiar with (human) molecular genetics and you are now participating Molecular genetics reading group (or otherwise studying the subject), do this assignment later during this Autumn, not immediately. However, info about unknown concepts etc. can easily be obtained from various web-sources.

The goal of this assignment: Getting familiar with different databases and learning how to get various kind of information.

NCBI and Ensembl genome browsers to examine human (disease) gene.

Choose a human disease, or some other trait of interest and then use the OMIM site identify a gene that is implicated in the etiology of the disease/trait.

Then use the NCBI, UCSC and Ensembl genome browsers to answer the following questions about the gene:

1. What are the various identifiers (aliases) for the gene
2. Where is the gene located on the chromosome (cytologically and physically)
3. What is the Reference sequence (RefSeq) for the gene
4. How many exons are there in the major transcript, and how long is it
5. What is known about the function of the gene
6. Using either the NCBI or Ensembl browser, explore the structure of the gene in a mouse and human.
7. From what other animals there is information about the gene

An example about a trait "asthma" and a gene, Interleukin-13 (IL13).
You cannot use this particular gene in your assignment!

- Go to <http://www.ncbi.nlm.nih.gov>, click OMIM, and type the disease (or any trait, try what kind queries work!) for example "asthma" into the search field.
- Scroll through the genes until you find one of interest, for example Interleukin-13 (IL13), and click on the blue number (*147683).
- This will bring up a page with lots of information, but you can also link to various sites, including Entrez Gene, which tells you that LocusID for IL13 is 3596 and that its cytological location is 5q31 (chromosome 5, long arm, position 132.1 Mb).
- The gene is also called ALRH, BHR1, and P600. The RefSeq is NM_002188, an mRNA sequence.

- There are four exons in the major transcript, which is 1282bp long and codes a protein of 146 amino acids, 34 of which are for the signal peptide.
- The protein is an immunoregulatory cytokine that regulates B-cell maturation and differentiation and the inflammatory response in the presence of allergens.
- This information can also be obtained by linking out to primary source from the Human Genome Organization Gene Nomenclature Committee (HGNC:5973), which links to the <http://www.ensembl.org> and <http://genome.ucsc.edu> pages.
- Alternatively, you can access these sites directly and search for the genes by typing "IL13" into the search engines.
- The Ensembl gene identifier is ENSG00000169194, and the gene view indicates that the transcript spans almost 3 kb. A number of gene ontology functions are listed as well.
- The UCSC site displays polymorphisms and conservation plots as well as a large number of other tracks.
- Human – mouse comparison. From the Ensembl homepage (<http://www.ensembl.org>) find the Gene Report for Human IL13 by typing ENSG00000169194 into the search box.
- The top hit should link to the GeneView page for Human IL13, which allows you to choose alignments with various species or combinations of species.
- Click on "View genomic alignment with Mus musculus" which will bring up graphics of chromosome 5, an overview of a 1-Mb window around IL13, and a "Detailed View" of 3 kb encompassing the gene.
- You can use the "Zoom" feature to explore the alignment in lesser or greater detail, or navigate to other multiple alignments and sequence feature annotations.