Finding genes in practice

Practical for FOS course Bioinformatics: Computational biology of complex diseases and ageing, Tuesday October 23, 2018

Yolande Ramos and Rodrigo Coutinho de Almeida (Molecular Epidemiology)

In this assignment write down which choices you make and, more importantly, why you make these choices (i.e. most relevant, most interesting, most tantalizing etc.)!

1. Selection of genes of interest

1. Check one of the files [Data_(1-12).txt] of the (part of a) whole-genome linkage scan for Osteoarthritis OA) and write down the highest score and marker (DxSxx) observed in your data. Does this give you insight into genetic cause of OA?
2. Go to the UCSC database (**http://genome.ucsc.edu/**) and click on the '**Tools**' → '**Genome Graphs**' option.
3. Set '**assembly**' to Mar. 2006 (NCBI36/hg18).
4. Click '**upload**' so you can upload your own linkage datafile.
5. Enter a name and description and don't change any of the '**best guess**' fields.
6. Upload the file and select it to be displayed in **blue** instead of the --nothing-- in the '**graph**' menu.
7. Set the '**significance threshold**' at the highest LOD score in the data minus 1 (representing all the genes in the 1-LOD drop interval). So, if the highest LOD score is 3.5 the '**significance threshold**' should be set to 3.5 – 1 = 2.5.
8. Browse the region with highest peak to see the genetic area in which this LOD score is observed. Which is the causal gene?
9. Go back in the browser (use the '**back**' button of internet explorer) and then click '**sort genes**' to go to the Gene Sorter.
10. In the next menu configure the Gene Sorter to display the Gene Ontology column and submit again.
11. In the new column (Gene Ontology) a more detailed description of proteins coded under the linkage peaks is given. Explore descriptions.
12. Set '**display**' to all.
13. Go through the genes and check for genes relevant for the disease we are investigating (Osteoarthritis).
14. Select 2 of these genes, which, from here onwards, will be your 'genes of interest'. After you have selected your genes please contact Yolande or Rodrigo.

2. Online databases

For this assignment you will explore several databases using the gene of your choice.

15. Go to http://www.ensembl.org and enter your gene of choice (e.g. the C-reactive protein gene 'CRP').
16. If prompted choose '**gene**' in feature type and '**homo sapiens**' in species type.
17. When several options are presented (depending on whether the gene has multiple transcripts or the abbreviation is not unique) carefully select the gene you intend to review and click on '**region in detail**'.
18. You now see a view of your gene in its genetic context and neighboring genes and features.
19. Using the menu on the left check alignments in either graphics or text to several other species. Try alignment to both near and distant species, and determine the amount of interspecies overlap.
20. Open a second tab or browsing window and go to http://genome.ucsc.edu.
21. Click the '**Genome Browser**' option and in position/search term enter the gene name again.
22. Select the right transcript for your gene and click on the link for this transcript.
23. Review the summary screen and identify which tracks are similar to the Ensembl browser.
24. Look up the conservation tracks in USCS, and click on the hyperlinks '**Conservation**' or '**28-Way Cons**' in the menu to enter the configuration of these tracks.
25. Enter a few additional species to this track and view the conservation level of these by submitting the changes.
26. Decide which of the databases shows you the most intuitive view of the conservation level across species.
27. In the Ensembl browser a tab '**gene: [xxx]**' is present, review this tab and click on the option '**sequence**' in the left menu
28. Look through the sequence and identify the number of exons in the sequence (highlighted in red).
29. Click '**configure this page**' and switch on '**Yes and show links**' behind '**Show variations**' then save and close.

30. In the next view you can now identify common SNPs. Try to identify whether exon mutations, polymorphic in the EUR population, are present in your gene (Hint; click on the SNP in the sequence and then look under '**Population genetics**').
31. Click one of the variants to get more information of the variant, such as flanking sequence, which may be important for designing PCR primers and whether the SNP is on one of the many SNP arrays such as the Illumina OmniExpress array.
32. Go back the '**gene: [xxx]**' tab and click one of the transcripts.
33. In the next summary view you can review the cDNA sequence, exons and coded protein information by clicking in the left menu. In the cDNA view try to find out what the alternating blue and black text parts indicate.
34. In the UCSC browser, scroll down to the variations and repeats screen and change the SNPs (131) to '**full**'.
35. Identify SNPs which you also found in the Ensemble browser and click on one of these SNPs to get more information. Note that this also provides the flanking sequences.
36. Click on '**view DNA for this feature**' and add 200 basepairs to both 5' and 3' sides of the polymorphism.
37. Go back to the genetic context view of UCSC and review what the SNP colors mean (hint, click the gray bar next to the SNP names to get background information).
38. Open a third window and go to http://bioinfo.ut.ee/snpmasker/. Enter the region coordinates in the right box and click '**Run SNPmasker**'.
39. Click the '**check results**' option and open file to view the basic information on the SNPs. Find the number of SNPs recorded in this database for your gene of interest: is your SNP also listed? If not, what could be a potential problem?
40. Now check characteristics of your SNP in dbSNP (go to https://www.ncbi.nlm.nih.gov/pubmed, set menu on '**SNP**' and enter rs-number in the search menu).
41. Is the information comparable to the other webtools? Do you find any deleterious/non-synonymous SNP close by that previously have been found in association to any disease?
42. Because SNPs are often used as markers, i.e. to capture genetic information and not necessarily the SNP itself as a functional variant it is important to use the Linkage Disequilibrium (LD) information between SNPs. This can be done by use of the HapMap database, where LD information of several populations is recorded.

43. Go to http://www.internationalgenome.org/home and explore the website: what information can be found here?
44. In the ‘Links’ menu select ‘**Software tools**’: you can see software for SNP calling and Imputation are available. Explain in your own words what is imputation? What is it based on?

3. Haploview

45. To visualize the LD in a region the data has to be loaded in Haploview. Open 'Haploview' and select the '**HapMap Format**' in the first screen. Open the '**Example**' file in this program. You will see a list of the genotyped SNPs in the dataset, with several information fields.
46. Go to the tab LD plot; you now see a visual representation of all the genotyped SNPs in the dataset in an LD plot. How many blocks are located in this region? Full information on the LD color scheme is available through the help section.

In a research setting there is often the need to maximize output and minimize cost for generating the output. In this program a tool is included that will theoretically yield a maximized genetic information for a minimum of genotyped SNPs. The tagger tab is where this can be configured.

47. Select the tab '**tagger**' and click the button '**uncapture all**'. Randomly select 5 SNPs from the list and then hit '**Run tagger**'. A new results-tab opens automatically and this shows your selected SNPs and tagging efficacy.
48. Write down which 5 SNPs you selected and then reset the table, set a maximum of 5 tags in the lower part of the program and under the '**force include**' select the 5 tags you randomly choose. Run the tagger again and in the results tab you can review what percentage of total genetic variation present you have tagged by the selection.
49. Go back to the configurations page, click '**reset table**', and then rerun the tagger. Is the percentage of captured variation increased by the application-chosen SNPs?
50. Click on the SNPs in the left pane to see which information these capture.

    NOTE: this example-file contains SNPs from a specific region which limits the possibilities!
51. Explore the databases in higher depth (e.g. you can check the screenshots from Haploview of the 1000 Genome project), and take a look at several other online databases listed below:

- http://www.ncbi.nlm.nih.gov/omim - Human specific gene characteristics
- http://biogps.org/#goto=welcome - Gene information (o.a. gene expression)
- http://gvs.gs.washington.edu/GVS/ - The Seattle SNP server
- https://omictools.com/microsniper-tool - Predicts effect SNP on miRNA targets