Finding genes in practice

Practical for FOS course Biomedical Data Sciences, Friday October 23, 2020
Yolande Ramos (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 and write down the highest score for a marker (DxSxx) observed in your data.
2. Go to the UCSC database (http://genome.ucsc.edu/) and go to '**Tools - Genome Graphs**'.
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 to see the genetic area in which this LOD score is observed. Which is your gene of interest?
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 relevant genes 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 supervisor.

2. Online databases

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

15. Go to **http://www.ensembl.org/index.html** and enter your gene of choice (e.g. WW Domain Containing E3 Ubiquitin Protein Ligase 2 'WWP2').
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 '***17-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 (130) 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 '***Flanking sequence***' 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 **https://www.ncbi.nlm.nih.gov/snp/**. Enter a SNP and review the information screen (what is the genomic location of the SNP? Are there any other SNPs in the neighborhood and if so how are these related to the original SNP?).

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 LD information between SNPs. This can be done by use of the HapMap database, where LD information of several populations is recorded.

39. Go to **http://www.internationalgenome.org** to access the database. Go to '***Browser'*** and select '***Ensembl GRCh37***' which will lead you to Ensemble with incorporated 1000 Genome data. Here, you can perform a search for your gene of interest.
40. Click '***Variations in gene***' and explore the page you view. Select one of the SNPs and check its genomic location.
41. Now go back and click '***Region in detail***'. To get insight into the genomic structure of the gene, go to 'Linkage Data', select population '***CEU***'. Explore the Linkage structure in this region.

42. Open **https://ldlink.nci.nih.gov/**, go to LDproxy, select population '***European***', and enter your SNP. you now see a visual representation of all the genotyped SNPs in the dataset in an LD block-plot. Full information on the LD color scheme is available below the plot.
43. How many SNPs in LD can you find? How does the information help you further finding causal variants?

In a research-setting there is often the need to maximize output and minimize costs 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.

44. Finally, go to **https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php**. Upload the SNP of interest in '***Build Query***' and submit (NOTE: if the output shows a long list of SNPs or very few SNPs you can change the LD threshold in the '***Set Options***' tab).
45. What information can you obtain from the output?
46. Explore the databases in higher depth, and take a look at several other online databases listed below. What would you conclude from this exploratory exercise?

• 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://gtexportal.org/home/	-Tissue-specific gene expression and eQTL