Practical: Thursday morning

This practical focuses on public resources that you can use to learn about the frequencies and consequences of variants ascertained through sequencing studies. This afternoon we will be introduced to filtering and analyzing variant calls from sequencing data using Hail.

Depending on time, make sure you focus on the instructions and questions in red in what follows.

Part 1: Annotation using the Variant Effect Predictor (VEP)

As its name suggests, the VEP is a tool for predicting the consequences of sequence variants (http://uswest.ensembl.org/info/docs/tools/vep/index.html). It can be run on your own server using a Perl script, or by uploading your variant data to the web server. If you have a large number of variants (e.g. a whole exome's worth), it is probably better to download the Perl script and install the necessary caches and databases on your own servers (get someone else to do this if you possibly can), but in this practical we will just use the web server to annotate a few variants.

First take a look at the file `minivcf_for_VEP_practical.vcf` in the directory for this practical. You might want to make your own new directory for the practical and copy the VCF over. Note that this is a fake VCF, and genotype columns have been left out because they are not required for the annotation step.

You'll notice that the file begins with a lot of header lines beginning with "#". A lot of these are uninformative and get in the way. You may want to use the `grep -v` command to exclude the lines containing "GVCF", "GATK" and "contig" when viewing the file.

Take a look at the lines beginning "##INFO=<ID=", which describe various metrics in the INFO column, and think about what these metrics are (some were mentioned in the lecture earlier).

Now we're ready to annotate the VCF. Go to:
http://grch37.ensembl.org/Homo_sapiens/Tools/VEP
(**N.B. It's important that we annotate variants with respect to the correct reference genome.** In this case, the variant calling was done against GRCh37, not the most recent version, GRCh38.)

(There is a description of the input options here, if you need it:
http://uswest.ensembl.org/info/docs/tools/vep/online/input.html)

Select "minivcf_for_VEP_practical.vcf" from the directory for this practical in the "Upload file" section.
Select "Ensembl transcripts" under "Transcript database to use".
Select "Gene symbol" and "Protein" under "Identifiers".

Select “1000 Genomes continental allele frequencies” and “ExAC allele frequencies” under “Frequency data”.
Select “Transcript biotype”, “Protein domains”, and “Exon and intron numbers” under “Miscellaneous”.

Run the job, and when it’s finished, select “View results”. You could look at the results in the web browser, but, for a more authentic experience, you should download the results VCF and look at this using `less` in a Terminal window. Use `grep` to pull out the line that describes the CSQ field that has been added to the INFO column by VEP.

Now take a look at the CSQ annotations that have been added to the INFO column for each variant. You might want to use the `grep`, `cut`, and `tr` commands to visualize these more clearly. Make sure you understand the structure of this annotation string – it can be a bit overwhelming!

Question 1: What is the most severe consequence of each variant in the file? Hint: Take a look at the description of the various variant consequences here http://uswest.ensembl.org/info/genome/variation/predicted_data.html , and think about why they are ordered in this way.

Question 2: Compare the consequences of the two different ALT alleles at the site 1:202724482.

Question 3: How do the 1000 Genomes and ExAC frequencies of the different variants compare to the allele frequencies in the AF field given in the INFO column? List reasons that they might differ.

Part 2: Exome Aggregation Consortium (ExAC)

ExAC has released data from 60,000 exomes. It is possible to download the site VEP (i.e. giving genotype counts but no individual genotypes), but we will just take a look at the web server http://exac.broadinstitute.org .

The database contains both common and very rare variation found in the ~60,000 individuals’ exomes. You can enter any gene to see the variants detected in all these samples, as well as their frequencies.

Use the “LoF” button to show only the loss-of-function variants, which tend to be the most severe.

Question 4: Compare the loss-of-function variants observed in the genes *KMT2A* and *ATAD3C*. What do you observe? Why might this be?

Click on an individual variant and explore the information that is given on the next page.

Question 5: You have sequenced an individual with a rare disease and discovered a variant on chromosome 9 at position 127661645 (denoted 9:127661645). Why might you rule this out as being causal of her disease? What if you found a damaging missense variant at 1:151638470?

Question 6: What about the variant at 19:12203287? Would you be as confident in the frequency estimate for this variant compared to the variant at 19:12221167?

Part 3: Ensembl

Ensembl is a database and web server run by the European Bioinformatics Institute that contains information about the sequence and structure of all genes in the genome, as well as about sequence variation and its consequences. The Ensembl browser can be used to visualize a range of data about gene regulation (e.g. from the ENCODE project).

Go to http://www.ensembl.org/Homo_sapiens/ .

We will use the "Population Genetics" link and the "Genes and regulation" link on pages for individual SNPs to compare different types of information about them.

Question 7: What can you learn about the world-wide frequency of SNP rs34536443?

Has this SNP been associated with any disease, and if so, which?

What are its predicted consequence? (use the "See all predicted consequences" link)?

Take a look at the structure of the gene this variant falls into (click on the link to the Ensembl ID from the Consequences table, then click on the link to the gene's location). Zoom in on rs34536443 and add different layers of data so you can see where the annotations in the Consequences table came from. (Hint: click on the cog-shaped symbol on the top left of the transcript tracks panel, and configure the tracks).

Question 8: What about rs2476601 – what is notable about this SNP?

Question 9: What about rs4988235? What is the significance of this SNP? Why does its worldwide frequency distribution look like this?

Take a look at the LD patterns around this SNP in different populations by clicking on the "Linkage disequilibrium" tab, then clicking the links to "view plot" in the "LD plot" column of the table. Why do you see these differences?

Question 10: Why might you still be interested in looking at the 1000 Genomes frequency information, given that ExAC is so much bigger?