Variant Filtering and Annotation

> View the VCF file *variants.vcf* using the ***less*** command and familiarise yourself with the format. This file contains a subset of variants from a sample and was generated by GATK haplotype caller.

- Can you tell the difference between heterozygous and homozygous calls (0/1 and 1/1)?

- Can you spot both SNVs and Indels?

- How many variants are in the file before filtering (HINT: you could use ***grep "^chr" variants.vcf | wc –l*** to count the lines in the file that start with chr (i.e. excluding the header lines))?

> Run ***perl Extract_pass_variants.pl*** to extract only variants that pass the GATK filters (remember back to lecture on best practice filters). In the initial file look in the FILTER column to try and identify the filtered variants. This code outputs the file *GATK_filtered.vcf*.

- How many variants remain after this filtering?

> Next we are going to apply filters for coverage and strand bias – run ***perl Filter_by_coverage.pl*** to generate *GATK_coverage_filtered.vcf* followed by ***perl Filter_by_strand_bias.pl*** which generates *FINAL_filtered.vcf*.

- How many variants were removed by each filtering step and how many remain?

- Can you work out from the code/output file what thresholds of coverage and strand bias are used?

> You are now going to spend the rest of the practical annotating the remaining variants using publically available tools. You can use whichever tools you like but here are a few suggestions:

- Ensembl variant effect predictor [http://www.ensembl.org/Homo_sapiens/Tools/VEP] (can upload a VCF – lots of useful annotations)

- UCSC [http://genome.ucsc.edu/cgi-bin/hgGateway] (search for chr:position and play with the options toward the bottom of the screen. HINT: Under *Variation* set *AllSNPs138* (dbSNP 138) to *full* and if the variant is in dbSNP click on it to see more information)

- SIFT [http://provean.jcvi.org/genome_submit_2.php]

- PolyPhen2 [http://genetics.bwh.harvard.edu/pph2/]

- dbSNP [http://www.ncbi.nlm.nih.gov/SNP/] (search for rsXXXX number found in VCF or in UCSC)

- If you have time try looking for publications that link the variant to any disease/trait

> This VCF file was a section of that reported from an individual with Dilated Cardiomyopathy (heart defect that causes sudden cardiac death) – are there any variants in your final filtered file that you think might be pathogenic (disease causing)?

NB: in reality a sample would include many more variants than this and these filters would not remove this many variants. The number was cut down to allow filtering and annotation within our time constraints