Department of Computing and Information Systems
COMP 90016
Workshop 5

The genomics can finally begin once a set of reads has been aligned to a reference. In this workshop we are going to dabble in SNP calling.
Find the data *reads.bam* on the LMS or digitalis. The task is to write a rudimentary SNP caller to identify heterozygous genotypes. Specifically, for each base in the aligned portion of the genome, report those that have two or more bases representing at 20-80% each.

1. Use the *Pysam* library to extract pileup columns from the data. This will make it easy to access the relevant information for this task.

   Answer: example script is available in the week5 folder on the server. For more information on pileup, see:
   http://pysam.readthedocs.io/en/latest/usage.html?highlight=pileup

2. Have your program report statistics on the frequency spread and the average quality (optional) of each of the heterozygotes as well as their genomic coordinates.
   Do all of them look real?

   Answer: Ideally, if heterozygotes present in the genome, we should have 50/50 read depth for the alleles. However, we can't guarantee exact same number of reads aligned to the position of interest, thus the proportions can be a bit off, but shouldn't be too far apart. Some of the pairs have very different aligned portions, which are the ones that don't look real. The program to generate the output is in the week5 folder on the server.

3. Investigate the genomic coordinates of some of the SNPs on UCSC's genome browser. If a population polymorphism is known for this site, you should be able to see this in one of the SNP tracks in the browser, identifying the SNP id (rs…). The genome to select for this instance is the human genome version hg19.
   On NCBI's dbSNP you can gather more information about variants by searching for the SNP id, or you can directly click on the variant in the browser to get you to a summary page within the UCSC website.
   What effect do they seem to have on the phenotype?

   Answer: an example from the output is chr15:28356859 (position is 28356858 in the output as *pysam* has 0-based positions). Genome browser shows there being a SNP at this position, with the id of *rs1129038*. Publications about this SNP is listed on the summary page (http://genome.ucsc.edu/cgi-bin/hgc?hgsid=663044351_VxXTTt8C3aCZZEiq7auzNJFrTQpT&c=chr15&l=28356858&r=28356859&o=28356858&t=28356859&g=pubsMarkerSnp&i=rs1129038). It's easy to find from the publications that this SNP have an effect on human eye color.

4. For those heterozygotes identified above that are not known SNPs, can other genomic features in the genome browser explain their appearance?

   Answer: Many of the identified locations don't present a SNP in a genome browser, but are marked by RepeatMasker. It annotates regions with repetitive DNA sequences, which create ambiguities in alignment, and in turn produce biases and errors. A good read on repeat regions: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324860/.

The above strategy to call heterozygous variants in aligned read data is pretty crude, but not too far from what is actually done in reality (and definitely not from what used to be done in the early days). Given the results you should be able to appreciate that variant calling is a messy business with noise in the data due to mapping and sequencing errors distracting the bioinformatician from the truly interesting results.
Overlapping the results from SNP calling with all known population variants is standard procedure and can be done in an automated fashion with simple command line tools (such as Bedtools). We did not explore this option here as dbSNP database files exceed 3gb in size for the human genome.