QLSC 600 - Assignment 1B (25 points)

Mathieu Blanchette and Ken Dewar

Due date: Sept. 24, 2pm.
Solutions to be submitted on MyCourses.
To be done in teams of 3 students
Only one submission per team
Clearly indicate the name of all team members

0) Read the attached PDF about predicting genotype from sequencing data. It is the summary of today's lecture.

1) (15 points) Write a program that reads a BAM file from a diploid genome, as well as a DNA sequence for the reference genome, and uses the probabilistic approach seen in class to identify genomic positions that are heterozygous or homozygous-non-reference. To keep things relatively simple, focus only on substitutions. Use one of the following libraries to help import and handle BAM files:

Python: http://pysam.readthedocs.io/en/latest/api.html

Rsamtools: http://bioconductor.org/packages/release/bioc/html/Rsamtools.html

(Feel free to use any other package that may help you read BAM files, but please implement the genotype calling algorithm yourself).

Data to be analyzed:

- http://www.cs.mcgill.ca/~blanchem/QLSC600/17.41000000-42000000.HG00096.wgs.ILLUMINA.bwa.GBR.high_cov_pcr_free.20140203.bam

  This contains mapped reads for region chr17:41000000-42000000 of the human reference genome (assembly hg19).

- Reference genome chr17
  sequence: http://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/chr17.fa.gz

2) (10 points) Compare manually (or computationally) your predictions to those made IGV. Do they agree? Discuss what you think the source of the disagreement is, and what could be changed in your algorithm to correct the situation. Limit your discussion to a maximum of 0.75 page.