Department of Computing and Information Systems
COMP 90016
Assignment 2
Release date: 10th of April 2018
Due date: 30th of April 2018 (11:59pm)
This assignment is worth 15 marks, or 15% of your final score

This assignment is introducing a few more practical issues around SNV discovery and genotyping to you. You are going to extend the work from the tutorials to a more comprehensive caller, and an automated haplotype phaser. Finally, you are going to interpret the data in light of the findings of a publication about human eye colour.

There are three sets of data for you to analyse throughout this assignment: *sample1.bam-sample3.bam.* Each BAM file is accompanied by a BAM index (*.bai*). The alignments are a very limited subset from a WGS experiment of three different **human** individuals.

Tasks:

1. One of the challenges around variant calling is the fact that sequencing data contains errors, which may present exactly like a variant would: As a heterogeneous position in the pileup data. The findings from the workshops and from Assignment 1 show that the base quality of sequencing errors may be slightly different to those of actual polymorphisms.
   Leverage this fact by implementing a strict quality filter to the data. Extend the Workshop 5 program or implement your own solution for a SNV caller that reports all heterozygotes where two bases present at the same position at a frequency between 20-80% each. Implement a quality filter that only counts nucleotides with a Phred base quality of at least 20.
   Compare the results of the original implementation to your more stringent solution. Discuss how the results change quantitatively and qualitatively for each of the three samples.
   Task 1 is worth 3 marks
2. The SNV caller developed so far can only detect heterozygous genotypes. In reality, we would like to identify all differences to the reference genome, including homozygous variants. This could be accomplished with the information from the BAM files alone, but it is slightly complicated.
   You can give your tool access to the file *human_reference.fa*, which contains the reference sequence for most of the positions covered by each of the BAM files (**not all!**). The range of bases (1-based) is stated in the name line of the FASTA file.
   - Equipped with the knowledge of the reference sequence, extend your method from Task 1 to call any position where **at least one** base different from the reference presents at a frequency of 20% or higher.
   - Further, report the **genotype** of each such position.
   - Follow the VCF specification for your output format using the fields *"#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE*". Set *ID* to ".", *QUAL* to the average

mapping quality of the non-reference allele, *FILTER* to "PASS", *INFO* to, for example, "AF=0.5" (showing the allele frequency of the non-reference allele), *FORMAT* to "GT", and *SAMPLE* to the genotype. See https://samtools.github.io/hts-specs/VCFv4.2.pdf for further specifications, and Section 1.4.2 for genotypes in particular. Also, order the genotype in ascending order (i.e. 0/1 or 0/1/2, instead of 1/0) and the ALT alleles lexicographically (A,G instead of G,A).

Task 2 is worth 3 marks

3. Phasing can be useful to identify specific haplotypes, instead of independent genotypes. It can also be utilized to identify non-congruent data.

   Write a Python program *phaser.py* with the aim to phase heterozygous genotypes. The input to the program is the output of the program from Task 2.

   - The algorithm is to identify any heterozygotes that are overlapped by the same read or paired reads from the same fragment.
   - Establish a haplotype for two close variants if there is a consensus of at least 90% to its phasing: For example, if 10 reads overlap **each allele** of two variants, then at least 9 of them have to support the same phasing for **each of the haplotypes.**
   - Build haplotypes of more than 2 variants by extending the phasing in order only from the right-most variant (i.e. you don't have to establish all possible pairs of variants, as consensus phasing is transitive). In other words, if variants V1, V2, and V3 are occurring along the chromosome in ascending order, and V1 is phased with V2, then it is enough to phase V2 with V3 only to extend the haplotype.
   - Print the haplotypes and the rejected haplotypes to the command line with some useful metrics. Note that the VCF format offers haplotype annotations, but you are not required to adhere to the specifications for this task.

   Please note that this task is difficult due to potential indels in the read alignments. Take good care in identifying the correct base corresponding to the variant in each read.

   Discuss your approach and the results for each of the samples. How many haplotypes are detected? How many potential haplotypes are rejected due to missing consensus? How can this information be used in further analyzing the data?

   Task 3 is worth 6 marks

4. The work by Sturm et al. titled "A Single SNP in an Evolutionary Conserved Region within Intron 86 of the *HERC2* Gene Determines Human Blue-Brown Eye Color" describes that eye colour in humans can be determined with reasonable confidence by single SNPs: two highly predictive variants named rs12913832 and rs1129038 are explored in the paper.

   - Given the output of your program (Task 2), state the genotype for rs12913832 and rs1129038 (their respective 1-based genomic coordinates are chr15:28365618 and chr15:28356859).
   - Argue what the most likely eye colour for each of the samples is (utilize Table 3 in the paper) and how confident this assertion is. Use the numbers and findings from the paper to make your point clear. Discuss multiple factors that affect your confidence in the eye colour being determined correctly.

   Task 4 is worth 3 marks

Notes:

- Use the quality filter for Task 2 as well, if you finished Task 1, otherwise proceed without.
- The HERC2 and OCA2 genes are on the reverse strand of the reference. The genotypes shown in the paper are stated with this in mind (from the point of view of the sense-strand of the genes). You have to consider this when comparing your genotypes with the genotypes in table 3.
- Remember that the reference contains lower-case characters denoting repeats. Convert these to upper case to make them consistent with read bases.