Homework 2

Lorenzo Bozzi,
{lorenzo.bozzi}@aalto.fi
Course Name: High-Throughput Bioinformatics
Course Code: CS-E5875
Due date: 20/11/2023

I. Introdution

THE following paper reports the solution to the first assignment.

The experiments for this assigment were conducted on a local machine, specifically a MacBook Air M1 (2020) equipped with 8 GB of RAM and the Apple-designed M1 chip. We utilized Jupyter Notebook (version 6.5.2) as the programming environment and employed R as the primary programming language.

II. Exercise 1

This section focuses on implementing a genotyper based on the naive Bayesian method of GATK, as described in lecture slides and documentation provided by the Broad Institute. The goal of this exercise is to use the genotyper to analyze a pseudo-pileup file named `input.txt` to infer the most probable genotype for specific genomic loci.

GATK, or Genome Analysis Toolkit, is a widely used software tool in genome analysis to identify genetic variants, such as Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (indels). If G_i denotes the genotype for $i = 1, \ldots, n$, and D represents the dataset, the model is based on Bayes' theorem:

$$\mathbb{P}(G|D) = \frac{P(G)P(D|G)}{\sum_i P(G_i)P(D|G_i)}. \quad (1)$$

Since the denominator is a constant and challenging to compute for each genotype type, the focus lies on the denominator term. In our case we focus evaluating the posterior probability of each of the ten possible genotypes "GG", "GC", "GA", "GT", "CC", "CA", "CT", "AA", "AT", "TT".

The `input.txt` file contains information regarding each genomic locus, including the chromosome, 1-based genomic coordinate, reference base, the count of reads covering the site, read bases, and base qualities. Base qualities are indicated as Phred scores calculated using the formula $q = -10 \log_{10} e$, where e represents the probability that the corresponding base call is incorrect. The dataset comprises 560 rows, with each row presenting detailed information about a genomic locus.

In the first part of the analysis, the focus was on selecting genotypes assuming uniform distribution for prior probabilities across all sites. The output showcases the selected genotypes along with the frequency of each genotype being identified as the most probable for the analyzed loci, as depicted in Table I.

Table I: Frequencies of Selected Genotypes.

Genotype	Frequence
AA	124
AT	10
CA	16
CC	68
CT	33
GA	34
GC	25
GG	146
GT	11
TT	93

In the table I, it's evident that certain genotypes have been identified more frequently than others. For instance, GG was detected 146 times, while AT was only observed 10 times. This indicates that, for these specific sites, GG appears to be the most prevalent genotype, followed by AA, TT, and so forth.

In the second part, the analysis was reiterated, comparing results across different populations: global (ALL), European (EUR), and Finnish (FIN). The outcomes are depicted in Table II.

It's noticeable that assuming uniform prior probabilities leads to varying outcomes for each site (except for site '29814971'). Additionally, it's interesting to observe that only for sites '47131885' and '47132180', all three populations exhibit the same pattern (i.e., the same most frequent genotype). Conversely, the Finnish and European populations show distinct behaviors only at site '29652851'.

Comparing against the uniform population emphasizes the critical role of selecting prior probabilities, significantly influencing the inferred most probable genotypes. While the Finnish population demonstrates a similar behavior to that of the European population, which is somewhat expected given that the Finnish population is inherently part of the European population, it reveals a different genotype at site '29652851'.

Furthermore, in the final section, the probabilities were normalised to ensure comparable histograms where genotypes were identified with the set $[n]$, following the order in which they were previously mentioned (e.g., GG=1, etc.), as depicted in figures 1 and 2 for reference.

Table II: Results of Genotype Analysis by Site and Population.

Site	Genotype	Population
29814971	GG	Uniform
29814971	GG	ALL
29814971	GA	EUR
29814971	GA	FIN
47131885	CT	Uniform
47131885	CA	ALL
47131885	CA	EUR
47131885	CA	FIN
29812725	CT	Uniform
29812725	CT	ALL
29812725	AT	EUR
29812725	AT	FIN
47132180	CC	Uniform
47132180	CA	ALL
47132180	CA	EUR
47132180	CA	FIN
29652851	TT	Uniform
29652851	CT	ALL
29652851	AT	EUR
29652851	CT	FIN

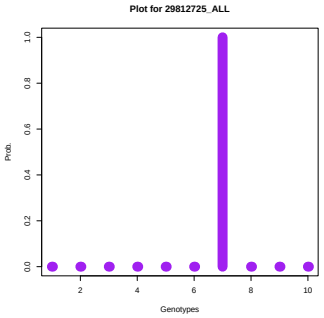

Figure 1: Plot for ’29812725’ ALL case.

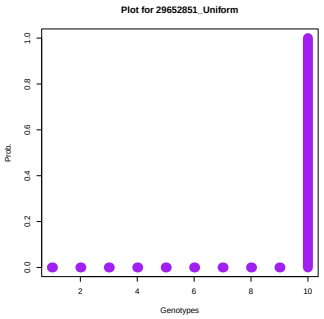

Figure 2: Plot for ’29812725’ Uniform case.