CS-E5875 High-Throughput Bioinformatics
Sini Rautio (modified by Maja Ilievska)
08.11.2018

Exercise 2

After the exercise session, you should submit a short report to MyCourses in PDF format. The report should contain a brief (max one page) description of your method and a description and discussion of your results. You may include equations and code snippets as you find necessary. Include a table of the number of times each genotype was selected and bar plots. Also include your entire code at the end of the report. **Submitting only the code with some comments is not enough!** The deadline for the report submission is on **16th November at 12:00**.

In this exercise you are asked to implement a simple genotyper based on the GATK naive Bayesian method explained in the lecture slides. To simplify the task, a pseudo-pileup file is provided as input file. Each row in the file contains the following information: chromosome, 1-based coordinate, reference base, the number of reads covering the site, read bases and base qualities. The column containing base qualities contain the actual Phred quality scores (instead of an ASCII character encoding), which measure sequencing error. Phred score q is given by,

$$q = -10 \log_{10} e,$$

where e is the probability that the corresponding base call is incorrect.

The data in the columns "read bases" and "base qualities" are comma-delimited and each read base-base quality pair is obtained from one sequencing read (from the pileup). Your genotyper should compute $P(G|D)$ of all 10 genotypes {"GG", "GC", "GA", "GT", "CC", "CA", "CT", "AA", "AT", "TT"} where D is the data, i.e the information in the columns "read bases" and "base qualities" and G denotes a genotype. The genotyper should determine the maximum a posteriori (MAP) estimate of the genotype.

In the first part, assume that the prior probabilities of the genotypes have uni-

form distribution for all sites (3 points). In the second part, repeat the analysis for the selected DNA sites below, using the indicated population genotype frequencies (ALL=whole population, EUR=european, FIN=finnish) as prior probabilities. Compare the results from the different population frequencies and with the results in part 1 (2 points). Did the results change? Why did they change/why didn't they change? For sites 47131885 and 29812725, plot as bar plots posterior probabilities of each genotype using uniform and "ALL" population priors (1 point).

Coordinates	**G**	**ALL**	**EUR**	**FIN**
29814971	GG	0.308	0.239	0.202
	CC	0.294	0.237	0.232
	GC	0.385	0.525	0.566
	CT	0.011	-	-
	GT	0.003	-	-
47131885	CC	0.849	0.899	0.909
	CT	0.146	0.099	0.091
	TT	0.005	0.002	-
29812725	TT	0.116	0.161	0.131
	AA	0.031	0.012	0.010
	AT	0.853	0.827	0.859
47132180	CC	0.917	0.913	0.980
	GC	0.082	0.087	0.020
	GG	0.001	-	-
29652851	TT	0.919	0.799	0.788
	CC	0.005	0.010	0.020
	CT	0.075	0.191	0.192

To help you with some of the data pre- and post-processing steps, below are some possibly useful code/pseudocode.

```
data<-read.table(infile)
for (i in 1:nrow(data)) {
    bases<-unlist(strsplit(as.character(data[i,5]),
    split=","))
    quals <- ...
    ps <- compute_p(bases,quals,priors)
    selected_genotypes[i] <-genotypes[which.max(ps)]
```

```
}
table(selected_genotypes)
```

For the second part of the exercise, you can generate a 3-dimensional array (10 x 3 x 5) with the provided prior probabilities in the table above. The first dimension represents the genotypes, the second and third dimension correspond to the populations (ALL, EUR and FIN) and the coordinates respectively. If you find it difficult, you can use the code provided in the script generate_prior_array.R. You can run the script with the command source("generate_prior_array.R"). Finally, you can use the command 'cat' for printing a table of the selected genotypes on screen or in a file.