Genomics

Freie Universität Berlin, Institut für Informatik
Knut Reinert, Peter Robinson, Sebastian Bauer
Wintersemester 2012/2013

7. Übungsblatt vom 4. Dezember 2012
Diskussion am 13. Dezember 2012

Exercise 1.

Bayes' theorem is often used for a set of n mutually exclusive events $E_1, E_2, \dots, E_n$ such that $\sum_i P(E_i) = 1$, and can then be written as

$$P(E_i|B) = \frac{P(B|E_i)P(E_i)}{\sum_i P(B|E_i)P(E_i)}. \tag{1}$$

An example that is seen in many textbooks describes the situation in which a chest X-ray (CXR) is obtained as a part of a routine checkup and a shadow is seen in one of the lungs that is suspicious for lung cancer. The referring physician consults a study that documents the following conditional probabilities:

- P(CXR=positive|Lung cancer=present)=0.8
- P(CXR=positive|Lung cancer=absent)=0.02

That is, the CXR has a false-negative rare of 20%, and a false-positive rate of 2%. Assuming for now that the prior probability of a patient having lung cancer is 1 in 500. That is, P(Lung cancer=present)=0.002 and P(Lung cancer absent=0.998). Calculate now the posterior probability for the diagnosis of lung cancer in a person without any other symptoms who presents with a positive chest X-Ray.

Exercise 2.

a) Consider the probability (according to MAQ) that a read z comes from position u of a reference sequence $\mathcal{R}$

$$p(z|\mathcal{R}, u) = \prod_{i \in \text{mismatched bases}} 10^{-\frac{q_i}{10}}$$

Which mapping is more probable:

- Read z is assigned to position u_1 with mismatches at position 7 (PHRED = 10), 23 (PHRED 15), and 37 (PHRED (20)
- Read z is assigned to position u_2 with only one mismatch at position 39 (PHRED 50)

Exercise 3.

Why does MAQ use a different prior probability for observing heteroyzgous genotypes when considering new SNPs (i.e., previous unobserved single-nucleotide polymorphisms) and known SNPs (previously observed variants already present in databases such as dbSNP)? For your answer, describe the effects of different priors on the MAP variant calling algorithm described for MAQ, and consider why one might assume different priors for the two classes of SNP (new vs. known).

Exercise 4.

This exercise is more of a practical, and it is hoped that it will be fun. We will download a publically available variant file from an exome performed on a healthy control. The exome is part of the supplementary material for the article *Low budget analysis of Direct-To-Consumer genomic testing familial data* published in the online journal **f1000 Research**. You can easily find the article by googling on it, or enter the following URL: `http://f1000research.com/articles/page/6/`. If you then go the the table entitled *Son exome files*, click on *Show all items*, you will be able to download `Son's VCF file.vcf (8.76 MB)`.

The exercise will consist in extracting profiles on the number of variants and their qualities in this file. You are free to use any programming or scripting language you like, but we will show how to do this with the unix tools awk and wc, which you should probably be familiar with anyway.

For simplicity, we will rename the files

```
$ mv Son\'s_VCF_file.vcf Son.vcf
```

awk is an incredibly powerful that can be used to scan and filter tab-separated lines in files. wc (word count) does what its name suggests. For instance, to count just lines that represent chromsome 18 variants, enter the following

```
$  awk '!/^[\t]*#/&&NF{if ($1==18){++c}} END {print  c}' Son.vcf
```

I got 628. Note that the first part of the command skips comment lines whose first symbol is "#". The variable c is autovivified (similar to in perl).

All in all, the command is not much niftier than grep

```
$ grep -c ^18 Son.vcf
```

But we can do much more with awk. Some examples follow. Count lines for chromosomes 18 and 17 (2691)

```
$  awk '!/^[\t]*#/&&NF{if ($1==18|| $1==17){++c}} END {print  c}' Son.vcf
```

Count lines for variants with an "rs" number (dbSNP accession id, i.e.,common, known SNPs). Note that the third column of a VCF file can be used to hold accession numbers (usually from dbSNP) for the variants.

```
$ awk '!/^[\t]*#/&&NF{if (index($3,"rs")==1){c++}} END {print  c}' Son.vcf
```

I got 35293 such SNPs, and using an obvious change to the awk command (“!=” etc.), I got 2524 SNPs not in the dnSNP database (6.7%). We can use the following commands to compare the quality (column 6) for variants with and without rs numbers:

```
$ awk '!/^[\t]*#/&&NF{if (index($3,"rs")==1){q+=$6;++d}}END{qua=q/d;print qua}' Son.vcf
450.662
$ awk '!/^[\t]*#/&&NF{if (index($3,"rs")!=1){q+=$6;++d}}END{qua=q/d;print qua}' Son.vcf
445.173
```

OK, so you get the idea. Search the file for the following data for this assignment. Consider that a dash (“-”) can be used to show a nucleotide is missing.

1. Which base is most commonly mutated amonst all variants affecting a single reference nucleotide?

2. Which base is the most mutated base amonst all variants affecting a single reference nucleotide?

3. What is the average quality of the variant calls for deletions of more than 2 nucleotides (hint: consider the length function. Do we show the deleted bases in the reference or the variant column?)?