### Exercises Monday 25th April Personlig Medicin Master

By Frederik Otzen Bagger

Burrows-Wheeler Transformation

1. Make the Burrows-Wheeler Transformation (BWT) matrix of the word “BOOKKEEPER”. An example is also given in the paper Li et al. 2009: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/>.

The BWT matrix is the rotation of BOOKKEEPER$ where the first two rows are:

$BOOKKEEPER

R$BOOKKEEPE

… etc.

And then sort the rows of the matrix according to *B-rank* (“$” is the first letter of the alphabet).

In B-rank the first occurrence of a letter comes before the second occurrence etc. (e.g. if you have 3 A’s in your name, make sure that the first A comes first when doing the sorting). It can help to start by making an index of the occurrences of each character in your word. (e.g. FREDERIKKE -> F1R1E1D1E2R1I1K1K2E3 ; and when sorting the rotation “K1K2” comes before “K2E3” even though E is before before in the alphabet).

1. In order to make the ranking computationally cheaper, think about whether you would ever need to consider the letters after the “$”, when ranking?
2. Confirm that LF mapping rule holds true for your matrix.

LF mapping rule: *The ith occurrence of a character, c, in the last column, and the ith occurrence of c in the first column corresponds to the same occurrence in [your word]$.*

1. Think of a query to your BWT. Remember that you are allowed to have a query that uses last and first part of the reference-word.
2. See if you can use the last and the first column of the BWT in a way that you don’t have to search through all the rows.

GATK

1. Find the documentation for GATK haplotype caller (browse around and read the summary). What happens to base-quality scores below 18?
2. Use google to find the specifications for a VCF file

Below is an entry from a clinical VCF-file.

1. What is the position of the variant and what is the change compared to the reference genome?
2. You don’t need to understand all the annotation, but find the rs-number and look up the variants in ClinVar. Is it dangerous? How do you know? Could you have guessed this from the vcf entry?

chr1 13838 . C T 70.64 gnomAD\_AF\_too\_high AC=1;AF=0.5;AN=2;BaseQRankSum=0.431;DP=5;ExcessHet=3.0103;FS=4.771;MLEAC=1;MLEAF=0.5;MQ=23.63;MQRankSum=-1.383;QD=17.66;ReadPosRankSum=-0.674;SOR=2.209;CSQ=T|downstream\_gene\_variant|MODIFIER|WASH7P|653635|Transcript|NR\_024540.1|transcribed\_pseudogene||||||||||rs28428499|524|-1||SNV|EntrezGene||YES|||||||||||||C|C|OK||||||||||||||||||||||||||||||||||||||||-6.269|1.896|-4.373|-4.373|2.492|-2.109|0.382|0.382||||4.442|0.313180||rs28428499|3.61776e-01|2.34715e-01|3.87996e-01|3.81969e-01|4.04373e-01|4.23470e-01|3.38043e-01||||||||,T|non\_coding\_transcript\_exon\_variant|MODIFIER|DDX11L1|100287102|Transcript|NR\_046018.2|transcribed\_pseudogene|3/3||NR\_046018.2:n.1081C>T||1081|||||rs28428499||1||SNV|EntrezGene||YES|||||||||||||C|C|||||||||||||||||||||||||||||||||||||||3.895|7.656|-6.242|-0.370|-6.612|-6.612|-10.572|-0.319|-10.891|-10.891||||4.442|0.313180||rs28428499|3.61776e-01|2.34715e-01|3.87996e-01|3.81969e-01|4.04373e-01|4.23470e-01|3.38043e-01||||||||,T|regulatory\_region\_variant|MODIFIER|||RegulatoryFeature|ENSR00001164745|promoter\_flanking\_region||||||||||rs28428499||||SNV|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||4.442|0.313180||rs28428499|3.61776e-01|2.34715e-01|3.87996e-01|3.81969e-01|4.04373e-01|4.23470e-01|3.38043e-01||||||||;gnomADg\_AF=0.361776,0.361776,0.361776;GM\_force\_call=.,.,. GT:AD:DP:GQ:PGT:PID:PL:PS 0|1:2,2:4:78:0|1:13813\_T\_G:78,0,78:13813

Also look up on ClinVar the variant rs80358616.

Which gene does it affect? How big is the change?

What is the position on hg19 (GRCh37), and what is the position on hg38 (GRCh38)?

Without going down to the submitted info table, how many identifiers, aside for the rs-number, can you find for this one variant?

Make sure, and discuss with your colleagues that you understand the use and general meaning of the file-formats:

.fastq

.bam

.cram

.vcf

(addition of .gz)