

Assignment-5

Here's the data that you will work with:

X chromosome data (.zip file, size = ~715 MB).

The "README" file in the directory has information on the files.

Your assignment is the following. Given the following:

- (a) reads from a genome – 3 million reads of the 150m we actually generated,
- (b) the reference sequence of chromosome X – 150m instead of 3b for the whole genome,
- (c) the BWT last column and the pointers back to the reference for chromosome X,
- (d) the locations of the exons of the red and green genes in chromosome X,

Align the reads to the reference sequence with up to two mismatches, and then count reads mapping to exons of the red and green genes, counting 1 for each read that unambiguously maps to one of the two genes and 1/2 for each gene for a read that maps ambiguously. (Note: Each 'N' in the reads file will be interpreted as an 'A'.)

For each of the possible red-green gene configurations in the presentation, determine the probability of generating these counts given that configuration and determine the most likely configuration that leads to colour blindness.