

### Lab 1: *E. coli* descriptive statistics

In this lab we will be using data from the completely sequenced genome of *E. coli*, more precisely strain K-12, substrain MG 1655, version M52, M52 for short. All the data you need will be found in the directory `/class/data/s141c/ecoli`.

#### 1. Look!

Go to <http://www.genome.wisc.edu/k12.htm>, scroll down to the very last line of that page, which says "DOWNLOAD the uncompressed ASCII file (4.6MB)". Click on that last download and wait a little while. Then look at the sequence M52. (Do not download the 9.5 MB version mentioned a few lines earlier.)

#### 2. Composition

In the file `123nt.dat` in the class lab directory you will find the base composition of M52, that is, the number of Adenines, Guanines, Cytosines and Thymines (As, Gs, Cs and Ts). Also there is the dinucleotide composition and the trinucleotide composition, that is, the number of AAs, AGs, ...AAAs, AAG,s etc., where AA denoted consecutive As along one strand.

- (a) If you had to assign a probability to observing an A at a stated position on the *E. coli* genome, what figure would you use and why? Under what assumptions does this seem appropriate?
- (b) You are told there is an A at a position along the *E. coli* genome. What probability would you assign to it being followed by a G, and why?
- (c) Does the *E. coli* composition data suggest that the event we observe a G at one site is independent (in some suitable sense) of the previous two bases? Explain fully, illustrating with appropriate data.

#### 3. Purine counts.

In the class lab directory you will find the number of purines (i.e. A or G) in about 46,000 blocks of 100 base pairs. Also there are the aggregated numbers of purines in about 4,600 blocks of 1,000 bp, and in about 460 blocks of 10,000 bp.

- (a) For each set of counts, calculate the mean and standard deviation of the proportion of purines per block, and draw histograms of these numbers.
- (b) Compare the results of (a) across the different block sizes and comment.

- (c) For each block size, calculate the fraction of the counts within 1, 2 and 3 standard deviations of the mean.
- (d) Repeat (a), (b) and (c) for *proportions* (rather than counts) of purines in each block.

#### 4. Hexamer counts.

In class lab directory you will find counts of the numbers of occurrences of the hexamers GTATTG and TATAAT from 927 blocks of 5,000 bp. Also there are aggregated counts from 231 blocks of 20,000 bp.

- (a) For each set of counts, calculate the mean and standard deviation of the numbers of the hexamers per block, and draw histograms of these counts.
- (b) Compare the results of (a) across the two block sizes and comment.
- (c) Compare the results for the two different hexamers and summarize your conclusions.

#### 5. Gaps between hexamers

In class lab directory you will find lists of gap sizes (in bp) between occurrences of the hexamers CACTTT, GCATGC and TATAAT.

- (a) For each of the 3 hexamers, calculate the mean and standard deviation of the gap lengths, and draw histograms of these lengths.
- (b) Compare the results for the three hexamers and summarize your conclusions.

For each hexamer and value of  $k = 1, 2, \dots$ , let  $s_k$  denote the proportion of gaps exceeding  $k,000$  bp in length.

- (c) Plot  $-\log s_k$  against  $k$  and estimate its slope. What does this tell us about the tail of the gap distribution?
- (d) Compare the results for the 3 hexamers.