Short Project 1

Due date: 02/28/2020

In a paper published in 2008, Dohm et al. reported that short read data from high-thoughput instruments were prone to generating more reads from genomics regions corresponding to elevated GC content.

In this short project you will test whether the new protocols used in the Illumina sequencing technology are also biased towards generating more data from GC rich regions. You will be testing this hypothesis on a bacterial sequencing project of your choice. However, you are required to only use datasets from Illumina sequencing instruments release during or after 2012. See the following link for more details (https://www.illumina.com/content/dam/illumina-marketing/documents/company/investor-relations/ILMNQ318SourceBook.pdf)

The bioinformatics procedure you will need to follow to test this hypothesis is:

1. Identify a microbe of interest. Note that viruses are perhaps too small to observe any GC-related sequencing biases while fungi are perhaps somewhat too big to experiment with on a laptop. Consider using selecting a bacterial reference that has publicly available shotgun sequencing data. Include in your report a very brief explanation of why you selected reference. You can explore available microbiol genome on the NCBI's microbial genomes page (https://www.ncbi.nlm.nih.gov/genome/microbes/)

2. Find whole genome sequencing data associated with your selected reference genome. A good place to find raw sequencing data is the NCBI's SRA page (https://www.ncbi.nlm.nih.gov/sra)

3. Align the sequencing reads back against the reference genome using two different aligners of your choice. For each aligner, briefly explore the parameter space and report the parameter values which provided the best alignment statistics. Explain what the modified parameters do. Make sure your alignments are save as SAM files.

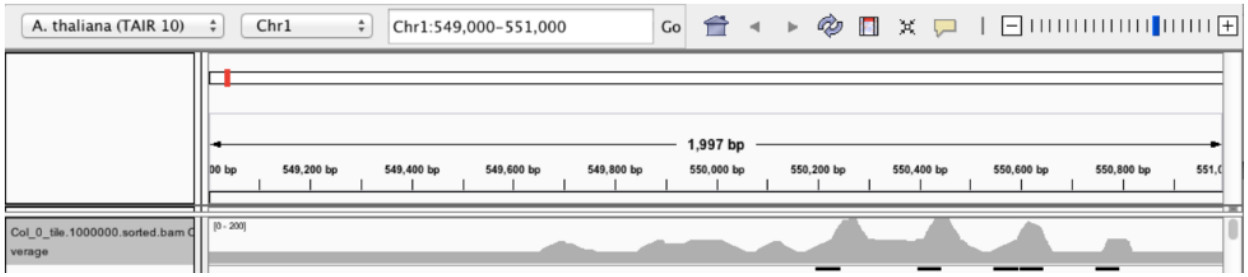
4. Use a sliding window of size 100 and compute the number of reads aligning within that window, i.e. alignment which start within the window, as well as the GC content within the window. The resulting table should look like the following:

Window number	Position in the genome	Number of reads starting in the window	%GC content
1	0-100	30	67
2	100-200	28	64
3	200-300	36	71
…			

Note that the easiest way to parse the resulting SAM file is by using a library such as pySAM (python) or Rsamtools (R) to parse the binary representation of your SAM file (BAM file).

5. Use a graphical viewer for next generation sequence alignments (ex. Tablet: https://ics.hutton.ac.uk/tablet/ or IGV: https://software.broadinstitute.org/software/igv/) to

inspect the alignment and see whether you can observe any inconsistencies or clear differences in coverage. Both these viewers can take a bam file and show a graph representing coverage of reds along the genome. See screenshot below where the peaks and valleys in the graph show variations in number of aligned reads along the displayed region.



6. Use a statistical test of your choice to test whether there is a statistically significant difference between regions of high GC and regions of low GC.

7. Use the graphical viewer of your choice to highlight an example of a GC rich and GC poor region to support your finding in step 6. Include screenshots of both regions in your report.

8. Include in your final report the GitHub link of the repo containing the code used to analyze the data as well as tab delimited table with your results for step 3.