Intro to Activity #1

Activity #1: Calculating Coverage Along a Chromosome

- In genome analysis, the most often asked question by researchers is:
  - How many reads cover each site in the genome?
- Coverage (or depth) is calculated as read length (L) times the number of reads (N) divided by the genome size (G).
  - $C = \frac{N*L}{G}$

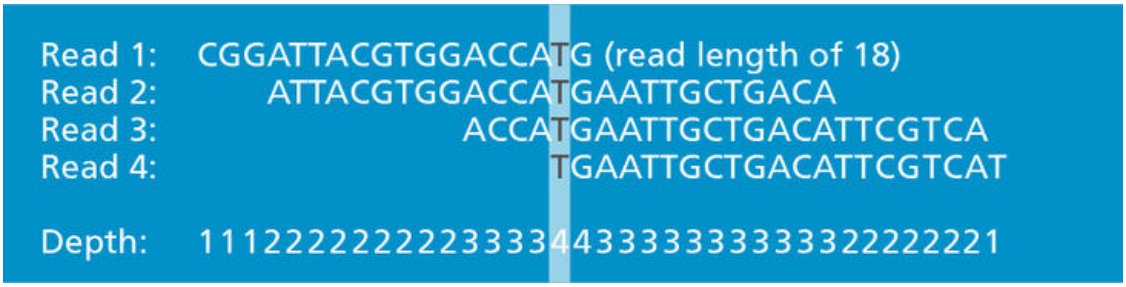

Image source: By Genomics Education Programme - Read, read length and read depth - read depth of '4', CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=58405954

Using samtools to get read depth output:

- A commonly used bioinformatics software for analyzing genome sequence alignment files (bam format) is samtools.
- Specifically, we have used `samtools depth` to output this information in long form.
- Because the output is so large, we will use R to read in the output and understand it.
- Today's dataset was aligned to the *Drosophila pseudoobscura* reference genome.
  - Specifically, you have been provided reads aligned to the genome, with the first 50 kilobases of each contig on the 4th chromosome subsetted.
  - The raw sequences used to make this file are from a published study, which is archived in NCBI's short read archive (SRP007802).

depth samtools depth [*options*] [*in1.sam|in1.bam|in1.cram* [*in2.sam|in2.bam|in2.cram*] [...]]

Computes the depth at each position or region.

Options:

-a	Output all positions (including those with zero depth)
-a -a, -aa	Output absolutely all positions, including unused reference sequences. Note that when used in conjunction with a BED file the -a option may sometimes operate as if -aa was specified if the reference sequence has coverage outside of the region specified in the BED file.
-b *FILE*	Compute depth at list of positions or regions in specified BED *FILE*. []
-f *FILE*	Use the BAM files specified in the *FILE* (a file of filenames, one file per line) []
-l *INT*	Ignore reads shorter than *INT*
-m, -d *INT*	Truncate reported depth at a maximum of *INT* reads. [8000]
-q *INT*	Only count reads with base quality greater than *INT*
-Q *INT*	Only count reads with mapping quality greater than *INT*
-r *CHR:FROM-TO*	Only report depth in specified region.

`samtools depth options`

What you will be doing:

- First, you will read the `samtools depth` output into R
- You will learn to manipulate data in R and make subsets
- You will calculate basic summary statistics of coverage
- You will examine the coverage information visually using a histogram
- Finally, you will plot coverage along a single chromosome

Reproducibility and bioinformatics

- During today's exercise, you will create a script to automate each step in your handout
- This script will enable you to re-do this analysis anytime in the future
- More importantly, you can change the input and use this script on any data file similar to the input provided
- Automating this process via a script means your results are directly comparable because the methods are identical
- Being able to get the same results every time is the definition of reproducibility and is fundamental to the validity of bioinformatics research