Your name: Madison Stevens

Bioinformatics Spring 2024

Genome Assembly Lab

Setting up the genome

1. Use ls to view the contents of the genome/ directory and copy/paste the output in your shell here.

DRR312053.lite.1 GCA\_900604845.1 ncbi\_dataset ncbi\_dataset.zip README.md

Thermus\_thermophilus\_T…R1.1bt2 Thermus\_thermophilus\_T…R1.2bt2 Thermus\_thermophilus\_T…R1.3bt2 Thermus\_thermophilus\_T…R1.4bt2 Thermus\_thermophilus\_T…AR1.dict Thermus\_thermophilus\_T…NAR1.fa Thermus\_thermophilus\_T…R1.fa.fai Thermus\_thermophilus\_T…rev.1.bt2

Thermus\_thermophilus\_T…ev.2.bt2

*\*Couldn’t find this myself, had to ask someone else what they found. I looked through every screenshot in that file and still couldn’t find this.*

1. How many sequences are in the genome assembly for this bacterium?

4

1. What cellular structures contain the genomic information (hint: look at the sequence names)?

1 chromosome and 3 plasmids

1. Why do we have to index a genome before mapping?

Since the file is massive, it is important to index the genome before mapping, because it will help narrow down the section of the genome or even retrieve the reference sequence information that will be mapped, which will save time compared to looking at the entire thing.

1. The output from bowtie is a .bam file? What is a .bam file?

Bam stands for binary alignment and map where the output of that file is a sequence alignment of the aligned reads.

Get sequence reads

1. Use ls to view the contents of the fastq/ directory and copy/paste the output in your shell here.

SRR5324768\_pass\_1.fastq.gz SRR5324768\_pass\_2.fastq.gz

1. What are the read lengths?

101 bp

1. What do the 4 lines for each read in a fastq file indicate?

1. sequence name starting with “@”

2. Nucleotide sequence

3. Empty line except for “+”

4. Quality score information.

1. Look at the read names for pass\_1 and pass\_2. What information is the same, and what is different?

The lengths are the same and the read names. The base pairs and symbols are different, and it has the first read as 1. 1. 1. in pass 1 and 1. 2. 1. in pass 2.

1. How do you explain the differences in the read names between the two files?

One of them is the forward sequence and the other is the reverse sequence

Alignment Time

1. Use ls to view the contents of the alignment/ directory and copy/paste the output in your shell here.

SRR5324768.bam SSR5324768.bam.bai

1. This set of commands involves the use of pipes. What is the utility of this?

Pipes or piping is a way to use that package to run that code. So, it specifies to use that package to run it, without the piping it might not be able to run the code. It also allows for multiple lines of code to run without having to write each line individually.

1. How many reads were in the fastq files?

250,803

1. How many reads aligned concordantly?

170,147 (67.84%) aligned concordantly exactly 1 time

1. What is the meaning of 'concordantly' and 'discordantly'?

Concordantly means the directions of the reads are the same and discordantly means the reads do not match or are in opposite directions.

Pileup format is a text-based format for summarizing the base calls of aligned reads to a reference sequence.

1. What do the dots mean?

Represent the forward sequence bp that match the reference sequence

1. What do the commas mean?

Represent the reverse sequence bp that match the reference sequence

1. What does uppercase mean?

Represent the forward sequence bp that don’t match the reference sequence

1. What does lowercase mean?

Represent the reverse sequence bp that don’t match the reference sequence

1. What does an asterisk mean?

Refers to a base pair missing for that sequence, which could have been deleted or missed when sequencing.

1. What do colors mean?

The colors represent the Phred scores or quality scores. Blue means a lower quality.

1. What does the underline mean?

This refers to orphan reads or untrimmed reads.

Variant calls with GATK

1. Use ls to view the contents of the variants/ directory and copy/paste the output in your shell here.

SRR5324768.vcf SRR5324768.vcf.idx

1. Open the .vcf file using less. Scroll down past the headers using the arrow key. Look in the REF and ALT columns (4th and 5th) - what are the meanings of these columns and how do you interpret them (particularly LR027517.1:574 and LR027517.1:578)?

The REF column is referring to the reference genome, which we are comparing to the target (which is the ALT column). When comparing to the reference and target genomes, the extra bases that can be seen for LR027517.1:574 and LR027517.1:578 might be either insertions or deletions.

1. Look in the sample-level information (columns 9 and 10): why is GT always 1? Check the .vcf manual for more information: https://samtools.github.io/hts-specs/VCFv4.1.pdf

The GT is the organism’s genotype that you’re looking at, so it makes sense that it would be one since this organism is haploid.

1. What would you expect the possibilities for GT to be if this were a human genome?

For a human genome the GT should be 2 because we are diploid.

1. What does AD mean and why is it always 0? (hint: try google)

This means allele depth, which is almost always 0 because the reads are too few.

1. What is the range for DP (just scroll up and down and give a reasonable ballpark answer)?

8-22

1. What does DP mean?

This is the depth of the reads for the target genome.

1. What do you think this .vcf file be useful for in the future, if it was for your project?

This type of file can be used for detecting single base sequence differences. This could help identify substitutions, deletions, and/or insertions for that sequence compared to the reference sequence.