Your name: William Holt

Bioinformatics Spring 2024

Genome Assembly Lab

**I could not get the program to work, so I used the screenshots to find the answers**

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

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

4 Sequences

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?

Indexing a genome before mapping helps to narrow down the genome to a certain section. It can also be done to retrieve the reference sequence that will be mapped. This saves time compared to mapping it without indexing.

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

.bam stands for Binary Alignment Map. It is a output file that makes an aligned sequence from 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?
2. Sequence name that starts with an “@” symbol
3. The nucleotide sequence
4. Empty line except for a plus sign
5. Quality score
6. Look at the read names for pass\_1 and pass\_2. What information is the same, and what is different?

The lengths and the read names are the same between both. On the other hand, the symbols and the base pair sequences are different.

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

It differentiates the forward and reverse sequences.

Alignment Time

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

SRR5324768.bam

SRR5324768.bam.bai

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

Pipes allow for multiple lines of code to be run back-to-back without having to stop in-between. Although there could be accuracy problems by doing so, it allows the program/packages to run the code faster than running the lines individually.

1. How many reads were in the fastq files?

There were 250,803 reads

1. How many reads aligned concordantly?

There were 170,147 reads that aligned concordantly exactly 1 time (67.84%)

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

Concordantly means that the reads are the same direction, while discordantly refers to reads that do not have matching 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?

It refers to the forward Strand

1. What do the commas mean?

It refers to the reverse Strand

1. What does uppercase mean?

It represents base pairs of the forward sequence that do not match the reference sequence

1. What does lowercase mean?

It represents base pairs of the reverse sequence that do not match the reference sequence

1. What does an asterisk mean?

Refers to a missing base pair, which could be due to a deletion.

1. What do colors mean?

The colors act as a quality score for that represented base

1. What does the underline mean?

The underline could be referring to an “orphan read”. After trimming paired-end reads, one pair might be trimmed while the other is not. The untrimmed read is the orphan read.

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

GT refers to the organism’s genotype. Since the organism is haploid, the GT number represents this with “1”.

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

GT should be “2” because humans are diploids.

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

AD stands for “Allele Depth”, which is apparently always zero due to there being a very few number of reads.

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?

It refers to 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?

It can be useful for detecting single base differences. For example, I can see this file type being useful for insertions, deletions, and/or SNPs. An annotated reference genome could be useful here to determine or find specific mutations that you might be looking for as well.