Your name: Kamil Kawiecki

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.

```
README.md            ncbi_dataset
Thermus_thermophilus_TTHNAR1.fa   ncbi_dataset.zip
```

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

8

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

Plasmids

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

Indices allow the aligner to narrow down the potential origin of a query sequence within the genome, saving both time and money.

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

The binary alignment map consists of the raw comprehensive genomic sequence data represented as compressed binary.

Get sequence reads

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

```
SRR5324768_1.fastq.gz     SRR5324768_2.fastq.gz
```

7. What are the read lengths?

101 nt

8. 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

9. Look at the read names for pass_1 and pass_2. What information is the same, and what is different?

@SRR5324768.1.1 1 length=101
AGGAGGGTATATGGACCGAAGGAGGTTTTTCCGGATTCTAGGCGCGGGGGGGCTTTTTGG
GGTCCTCAAGGCCAAGGCCCAAGGGCTTCCCTGGACGGAGG
+SRR5324768.1.1 1 length=101
@@CFFFFAFHHHHJIJJJJJIIGIJ?
DHIIIIJJIIJJJGIFIGIIH>@BBB&&+>C(05&&)0<>>>@@A(8<A<<28@8A88?##################
@SRR5324768.2.1 2 length=101
GAAGGAGGTTTTTCCGGATTCTAGGCGCGGGGGGGGGTTTTGGGCTCCTCCAGGCCAAG
GCCCCAAGGGCTCCCCCGGCGGAGGGGGGCGTCCCCCCCCCC
+SRR5324768.2.1 2 length=101
@BCFFFFFDFHHHIJJJIJJJJJJJJ###
############################
@SRR5324768.3.1 3 length=101
TATATGGACCGAAGGAGGTTTTTCCGGATTCTAGGCGCGGGGGGGGGTTTTTGGGCGCTTCA
AGGACAAGGGACAAAGGCTTTCCCGGGTGGAGGGGTGCAT
@SRR5324768.1.2 1 length=101
GGTGGGGGCGAAGGTCTCCTCCGTCCAGGGAAGCCCTTGGGCCTTGGCCTTGAGGAGGC
CCAAAAGCCCCCCCGCGCCTAGAATCCGGAAAAACCTCCTTC
+SRR5324768.1.2 1 length=101
@@@DFFFFHHHHHJFHIGIEHIJJJJJJJIHHHFFFFFEEDEDDDDDDDDCDDDDDDDDDDDDD
DCBDBDBDDDDDDDDDD>ACDDDDDDBD@BDCDDDDD
@SRR5324768.2.2 2 length=101
AGGGGCGCCCCTAAGGTCTTGGTGGGGGCGAAGGTCTCCTCCGTCCAGGGAAGCCCTTG
GGCCTTGGCCTTGAGGAGGCCCAAAAGCCCCCCCGCGCCTAG
+SRR5324768.2.2 2 length=101
CCCFFFFFHHHHHJJJGHIJIIGHJJJJJDDDDDCDDDDDDDDDDDDDDDBBDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD@@B99?
@SRR5324768.3.2 3 length=101
TACTCGGCGAGGGGCGCCCCTAAGGTCTTGGTGGGGGCGAAGGTCTCCTCCGTCCAGGG
AAGCCCTTGGGCCTTGGCCTTGAGGAGGCCCAAAAGACCCCC

The read lengths are all the same and all the sequences are represented by the same name. The differences between the files are found after the read names.

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

The middle number denotes which file the sequence belongs to and the first and third number represents the specific sequence number within its file.

Alignment Time

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

```
alignment   fastq       genome
```

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

Pipes form a temporary software connection; in this case between two programs: samtools and bowtie2.

13. How many reads were in the fastq files?

250,803 total reads

14. How many reads aligned concordantly?

```
170147 (67.84%) aligned concordantly exactly 1 time
  6571 (2.62%) aligned concordantly >1 times
```

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

Concordant pairs match pair expectations with the expected distances between mates, while discordant pairs don’t.

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

16. What do the dots mean?

Forward bases matching the reference base

17. What do the commas mean?

Reverse bases matching the reference base

18. What does uppercase mean?

Forward base is a mismatch to the reference base

19. What does lowercase mean?

Reverse base is a mismatch to the reference base

20. What does an asterisk mean?

Deletion of the reference base

21. What do colors mean?

It denotes the complete confidence of a mutation of the alignment compared to the reference genome.

22. What does the underline mean?

Contains the most likely contains for the msa.

Variant calls with GATK

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

24. 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)?

REF refers to the reference nucleotide at that position and the ALT represents any possible alternative allele also at that position. At the 574 position the ALT could any of the 4 nucleotides and at 578 the reference genome nucleotide is not certain.

25. 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

Genotype is always 1 because the allele value is equal to the first allele listed in ALT.

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

For the human genome the diploid nature of the chromosome could contain values of 0/1, 1 I 0, or 1/2, etc.

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

Unfiltered allele depth; the number of reads that support each of the reported alleles. Essentially, an AD value of 0 quantifies an uninformative read meaning that the difference between the most likely allele and the second most likely allele is not significant.

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

I had seen a range of approximately 8 - 40.

29. What does DP mean?

Filtered depth which gives you the number of filtered reads that support each of the reported alleles.

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

It provides a raw sequence file that has a comprehensive, yet linear sequence data that could be used to represent the confidence in the information that is being presented.