Name:

Comparative Genomics

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.
2. How many sequences are in the genome assembly for this bacterium?
3. What cellular structures contain the genomic information (hint: look at the sequence names)?
4. Why do we have to index a genome before mapping?
5. The output from bowtie is a .bam file? What is a .bam file?

Get sequence reads

1. Use ls to view the contents of the fastq/ directory and copy/paste the output in your shell here.
2. What are the read lengths?
3. What do the 4 lines for each read in a fastq file indicate?
4. Look at the read names for pass\_1 and pass\_2. What information is the same, and what is different?
5. How do you explain the differences in the read names between the two files?

Alignment Time

1. Use ls to view the contents of the alignment/ directory and copy/paste the output in your shell here.
2. This set of commands involves the use of pipes. What is the utility of this?
3. How many reads were in the fastq files?
4. How many reads aligned concordantly?
5. What is the meaning of 'concordantly' and 'discordantly'?

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?
2. What do the commas mean?
3. What does uppercase mean?
4. What does lowercase mean?
5. What does an asterisk mean?
6. What do colors mean?
7. What does the underline mean?

Variant calls with GATK

1. Use ls to view the contents of the variants/ directory and copy/paste the output in your shell here.
2. Open the .vcf file using ls. 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)?
3. Look in the sample-level information (columns 9 and 10): why is GT always 1?
4. What would you expect the possibilities for GT to be if this were a human genome?
5. Why is AD always 0?
6. What is the range for DP (just scroll up and down and give a reasonable ballpark answer)?
7. What does DP mean?
8. What do you think this .vcf file be useful for in the future, if it was for your project?