Team Exercises and Solutions

There are three exercises, which can be done in any order.

The goal of each exercise is to run QC tests to figure out why the datasets cannot be assembled into a single genome. If the cause cannot be determined with certainty, then suggest hypotheses consistent with the evidence. The final step is to identify how one can estimate the genome size of the sample organism.

There are five points for each exercise.
1 point for identifying the issue
1 point for identifying some of the evidence
1 point for identifying all of the evidence
1 point for describing how to identify the genome size
1 point for the correct estimation.

Exercise 1

This is a public dataset http://www.ebi.ac.uk/ena/data/view/PRJNA60717.

Data:

SRR492065_{1,2}.fastq.gz

Questions:

What is the issue with this dataset?

It is metagenomic

What is the underlying evidence?

FastQC per sequence GC content shows two peaks indicating multiple organisms
KAT gcp plot shows three distinct blobs indicating multiple organisms
Kraken identifies three main species of organisms in the data set.

What steps should be taken to identify the genome size?

Estimate k-mer coverage
Convert to read coverage estimate
Multiply total data quantity by proportion of main genome
Divide remaining data quantity by read coverage estimate

What is the estimated size of the genome?

K-mer peak lies at about 120x coverage
Read length is 100
k-mer size is 27
*Therefore read coverage N = M*L/(L-k+1) = 120*100/(100-27+1) = 162*
*Total data is read length * 2 * num. reads = 100*2*5354356 = 1,070,871,200*
*Enterococcus is about 52% of the reads, therefore 1,070,871,200*0.52=556853024*

Divide data amount by read coverage = 556853024/162 = 3,437,364 (3.4Mb)
Therefore the estimated genome size is around 3.4Mb

Exercise 2

This is a public dataset http://www.ebi.ac.uk/ena/data/view/PRJNA60811.

Public Dataset ():

SRR948594_{1,2}.fastq.gz
SRR948595_{1,2}.fastq.gz

Questions:

What is the issue with this dataset?

Only SRR948594 is WGS. SRR948595 is RNA-seq data.

What is the underlying evidence?

There is no evidence that determines the type of library in the files. It's only listed on ENA.
Kat comp indicates a difference between the two libraries, but it is impossible to conclude the cause of the issue.

What steps should be taken to identify the genome size?

Using only SRR948594 Estimate k-mer coverage
Convert to read coverage estimate
Divide total data quantity by read coverage estimate

What is the estimated size of the genome?

The mean k-mer coverage is 200x.
*Therefore the read coverage is 200*100/(100-27+1) = 270*
*The total data quantity is 100*2*18886541 = 3777308200*
This means the estimated genome size is 13,990,030 (14Mb)

Exercise 3

This is a bacteria of unknown origin.

Dataset:

Bacteria_001_{1,2}.fastq.gz
Bacteria_002_{1,2}.fastq.gz

Questions:

What is the issue with this dataset?

Bacteria_001 is one strain, Bacteria_002 is another strain of the same bacteria (i.e. two individuals)

What is the underlying evidence?

Kat comp spectra_mx shows minor proportions of the genome unique to both datasets. The mean of the exclusive content peak is the same as the peak of the shared content, which means the exclusive content shares the same depth of coverage as the rest of the genome. This suggests the difference is caused by genetic variation rather than technical biases.

What steps should be taken to identify the genome size?

Using either dataset Estimate k-mer coverage
Convert to read coverage estimate using median read length
Divide total data quantity by read coverage estimate

What is the estimated size of the genome?

K-mer coverage is 26x
Median read length 151
*Mean read coverage = 26 * 151 / (151-27+1) = 31*
Total data quantity is 225,890,464
Estimated genome size is 225,890,464/31 = 7,286,789 (7.3Mb)