Global Health Research Unit

FUNDED BY NIHR | National Institute for Health Research

Bioinformatics Training using *de novo* assembly as an exemplar

The bioinformatics training will comprise of the 3 phases.

1. Training on Galaxy to use command line software via a web interface. The objective is to
   a. become familiar with the bioinformatics steps for a particular process
   b. understand what software is used and how it is parameterized or configured
2. Reproduce the steps performed in Galaxy on the command line in order to become trained in how command line tools are used
3. To run batches of sequences through a pipeline that reproduces all the individual command line tools used from step 2.

Genome Assembly

For this purpose we will initially start with the process of assembling contigs from raw reads. Contigs is a term that means contiguous DNA and refers to the consensus sequence that is formed when sequence reads (usually from fastq files) are ‘stitched together’ to form large regions from the genome. With short reads, repetitive sequences usually prevent complete closed genomes from being produced but instead the end result is usually smaller pieces of contiguous DNA that make up the most of the genome.

Genome Assembly Tutorial

Galaxy is an open source, web-based platform for accessible, reproducible, and transparent computational biomedical research. It allows users without programming experience to easily specify parameters and run individual tools. It also captures run information so that any user can repeat and understand a complete computational analysis.

Learning Objectives

By following this tutorial you will be able to:

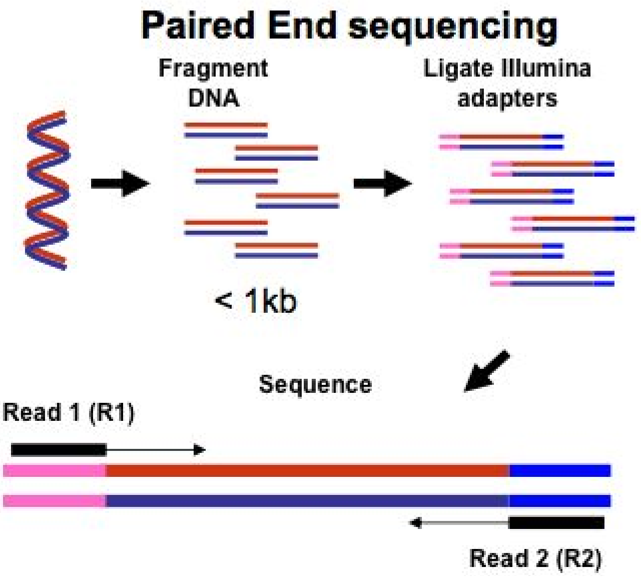
- Be able to explain the principles of *de novo* assembly
- Login to a Galaxy server.
- Upload data to a Galaxy server from a file on your local computer
- Access and run the software tools applying them to data you have uploaded
- Assemble a genome from a pair of fastq files

Stages in the Assembly Process

- Raw fastq QC assessment
- Fastq trimming
- Trimmed fastq QC assessment
- Assembly
- Assembly QC

Introduction to Next Generation Sequencing

Next generation or high throughput sequencing involves massively parallel sequencing of small fragments of DNA that have been generated from an original nucleic acid source. This can be genomic DNA, PCR amplicons or cDNA generated from RNA. In this tutorial you will be working with sequence data that has been generated from whole genome DNA extracts. In this case the genomic DNA is fragmented and then Illumina adapter sequences ligated so that common sequencing primers can be used prime from the fragments and generate sequence. The fragmentation and ligation can either be done in one step (e.g with the enzymatic NextEra system that uses a process they call tagmentation) or in two steps such as with the TruSeq process where the DNA is sheared physically and the adapters ligated subsequently. The Illumina sequencing technology works best when the fragment size (also known as insert size) is 300 - 500bp. With paired end sequencing both ends of the fragment are sequenced but these reads (one in the forward and one in the reverse direction) may not meet. If the insert size is too small the reads can be, and for assembly, should be merged.
A nice video of this process can be found here.



At the end of the sequencing process when paired end sequencing is used each sample sequenced will produce a pair of fastq files corresponding to one file with all the read 1s from the fragments and in the second file all the read 2s. The order of the reads in these files is the same so that the R1 and R2 for each fragment are synchronised. However since the fragments are generated randomly in the case of library preparation from whole genomes they are not in any order in relation to the genome. This can only be achieved by giving each read context through *de novo* assembly or aligning (usually known as mapping) to a close reference genome.

Fastq Format
A FASTQ file normally has four lines per sequence fragment.

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
Line 2 contains the sequence as letters that represent the nucleotides.
Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
Line 4 encodes the quality values for the sequence in Line 2, and contains the same number of symbols as letters in the sequence.

So for example FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The quality metrics are encoded as shown below (taken from https://en.wikipedia.org/wiki/FASTQ_format#Encoding). Most sequence data is now in Phred+33 format where quality ranges from 0 to 40 and is encoded as ! to J.

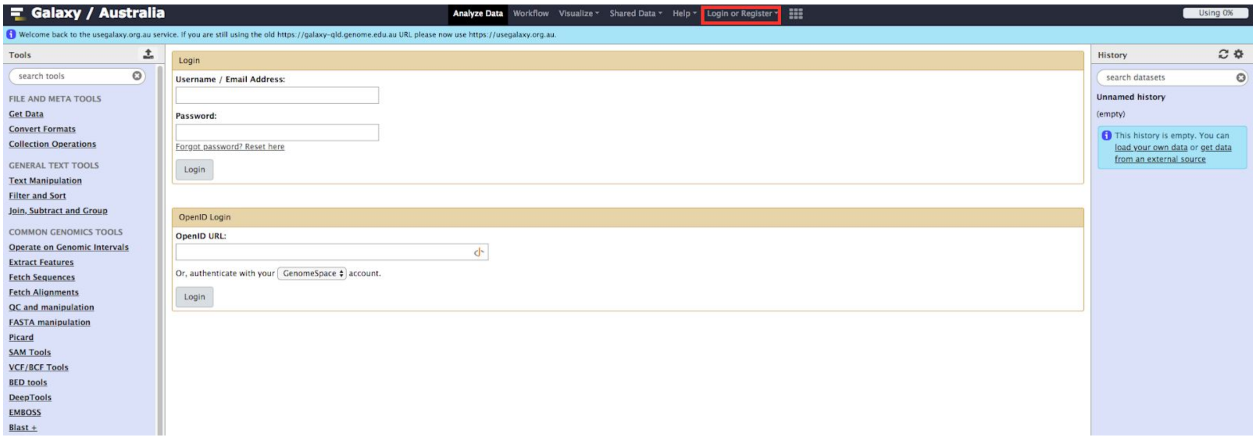
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
  0.2......................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

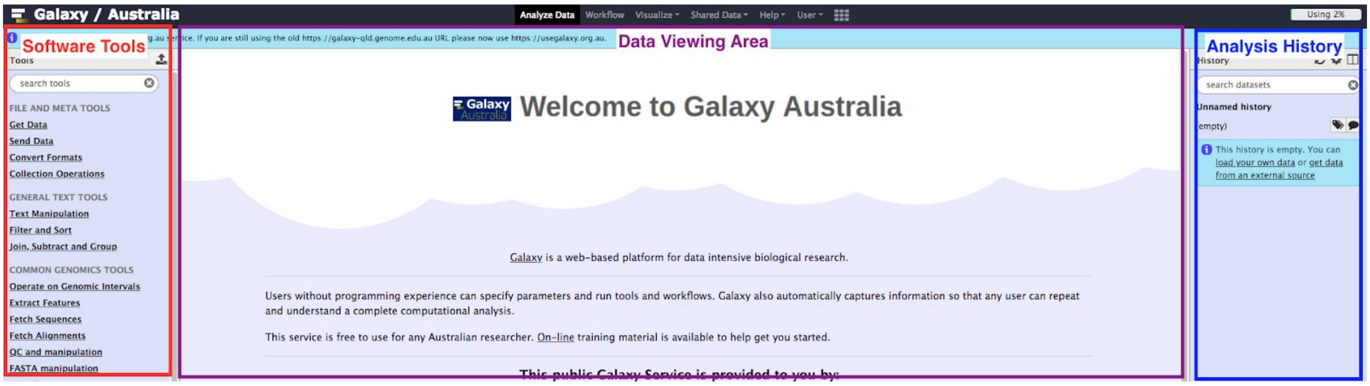
Using Galaxy to assemble a Genome

Logging onto Galaxy, navigating the system and uploaded data

1. Go to the Galaxy Australia website at https://usegalaxy.org.au/
   If you haven't registered, do so via the Register link and then Login


Welcome back to the usegalaxy.org.au service. If you are still using the old https://galaxy-qld.genome.edu.au URL please now use https://usegalaxy.org.au.

2. There are three main regions in the Galaxy window



Welcome to Galaxy Australia

Galaxy is a web-based platform for data intensive biological research.

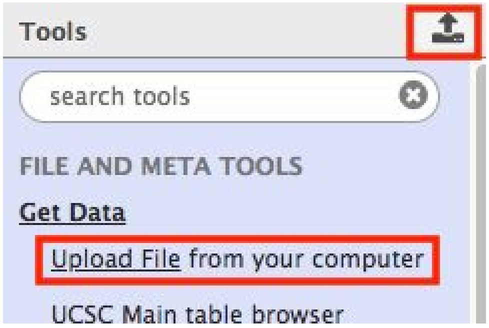
Users without programming experience can specify parameters and run tools and workflows. Galaxy also automatically captures information so that any user can repeat and understand a complete computational analysis.

This service is free to use for any Australian researcher. On-line training material is available to help get you started.

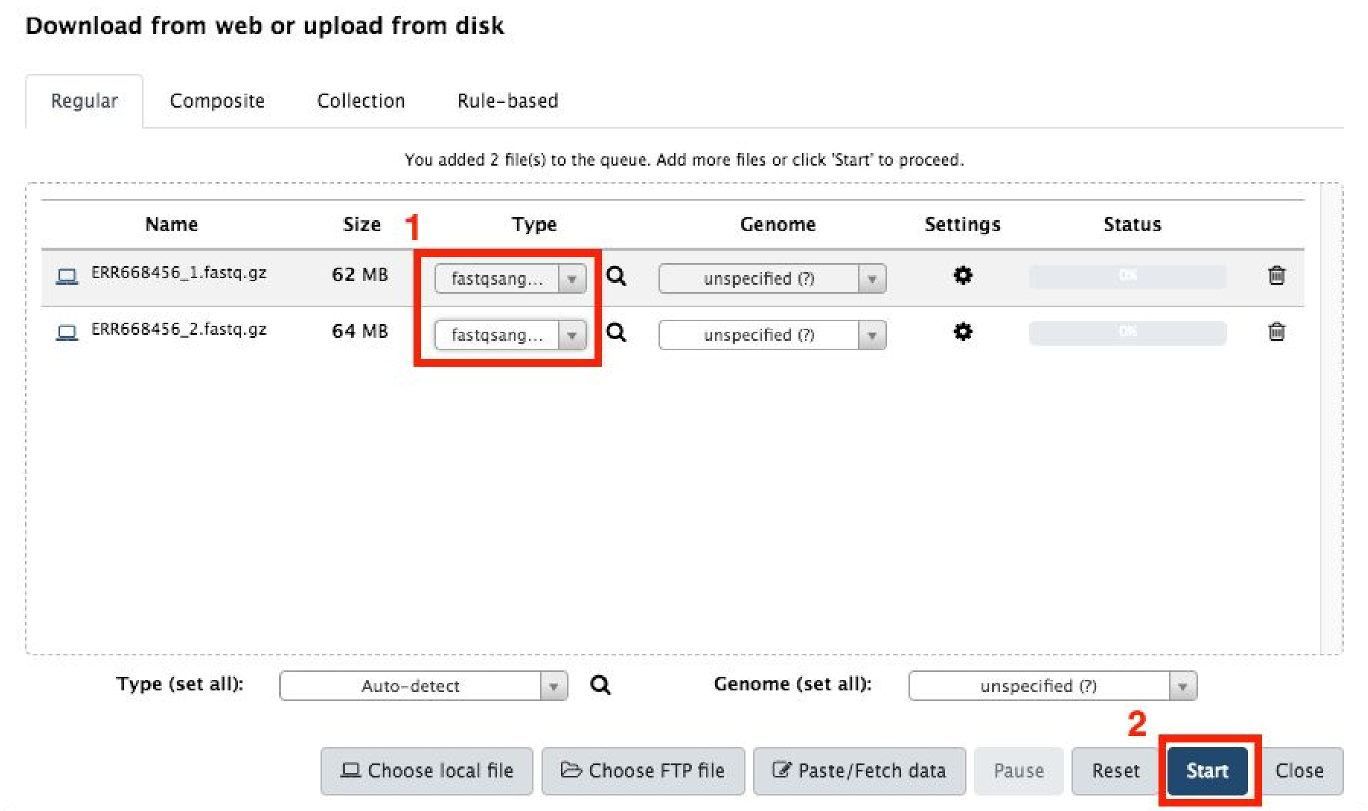
This public Galaxy Service is provided to you by:

Software tools are selected by searching by name on the 'Software Tools' region on the left side of the window. Clicking on them brings up the options for running the software in the central 'Data viewing area' and these software tools can be applied to data seen in the 'Analysis History' region on the right.

3. When you first login you will have no data to analyse, so let's enter some data. For the purposes of this tutorial we will use a pair of fastq.gz (gzipped fastq) files from small *Salmonella typhi*.
These can be downloaded from the following 2 links:
https://drive.google.com/file/d/14sGYl1hnkJwCKQmSzwKgh_t_Cda0pST_
https://drive.google.com/file/d/14pv42xRDKyDZNtR_VNz_XFV9fjyoBxMl
Download these 2 files to a directory on your computer
4. On the left hand side click on Get Data->Upload File or on the upload file icon



5. Drag the 2 fastq files you downloaded onto the pop up window that appears. In the Type dropdown for each file make sure you select ‘fastqsanger.gz’ and then click on the blue ‘Start’ button



Download from web or upload from disk

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
ERR668456_1.fastq.gz	62 MB	fastqsang...	unspecified (?)		0%
ERR668456_2.fastq.gz	64 MB	fastqsang...	unspecified (?)		0%

Type (set all): Auto-detect

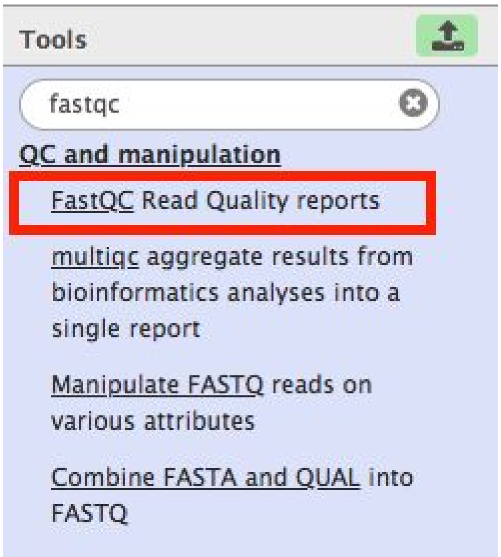
Genome (set all): unspecified (?)

Depending on your internet speed it may now take a few minute to upload the data. The files will then appear in the History on the right hand section of the screen in a small box that is first grey whilst it is waiting to be processed, then yellow as it is processing and finally green when the file has been registered in the galaxy system. In the screenshot below you can see 1 data file that has finished uploading and being processed and another that has uploaded and is now being processed so that it can be accessed on Galaxy.

Raw fastq QC assessment

1. Now that you have data in the galaxy server you can run software on that data. Behind the scenes it is using CLI (command line interface) UNIX software but galaxy provides a web browser interface to those software. With the 2 fastq files we have uploaded we will first assess the quality using the

fastqc tool. Type fastqc into the search tools box on the left


Tools

fastqc

QC and manipulation

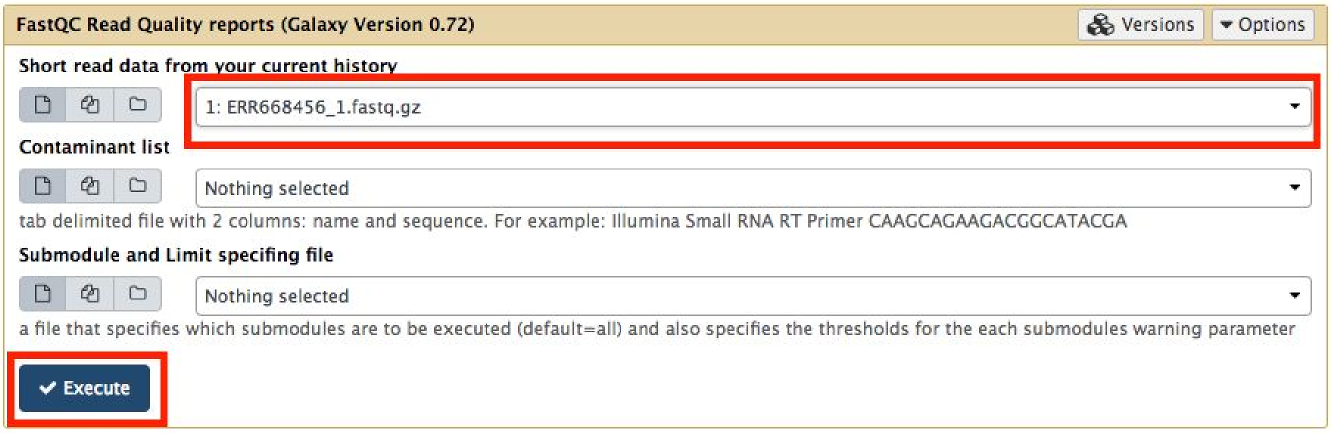
FastQC Read Quality reports

multiqc aggregate results from bioinformatics analyses into a single report

Manipulate FASTQ reads on various attributes

Combine FASTA and QUAL into FASTQ

2. Click on the FastQC link. This will bring up a screen in the central data viewing area where you can select parameters for the software. In this case you just want to run fastqc with its standard parameters but first make sure you select the _1 fastq file (Read 1):


FastQC Read Quality reports (Galaxy Version 0.72)

Short read data from your current history: 1: ERR668456_1.fastq.gz

Contaminant list: Nothing selected

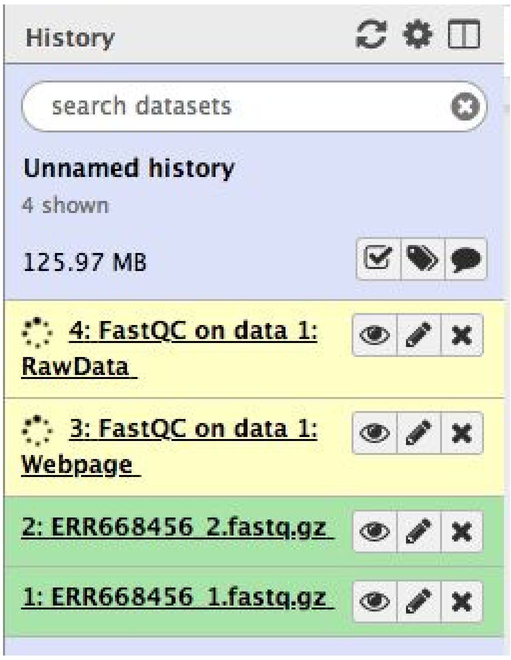
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file: Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Then click the blue Execute button.

On the right hand side you will see the two output files for the fastqc software start to be processed as fastqc is run using Galaxy



History

search datasets

Unnamed history
4 shown

125.97 MB

4: FastQC on data 1: RawData

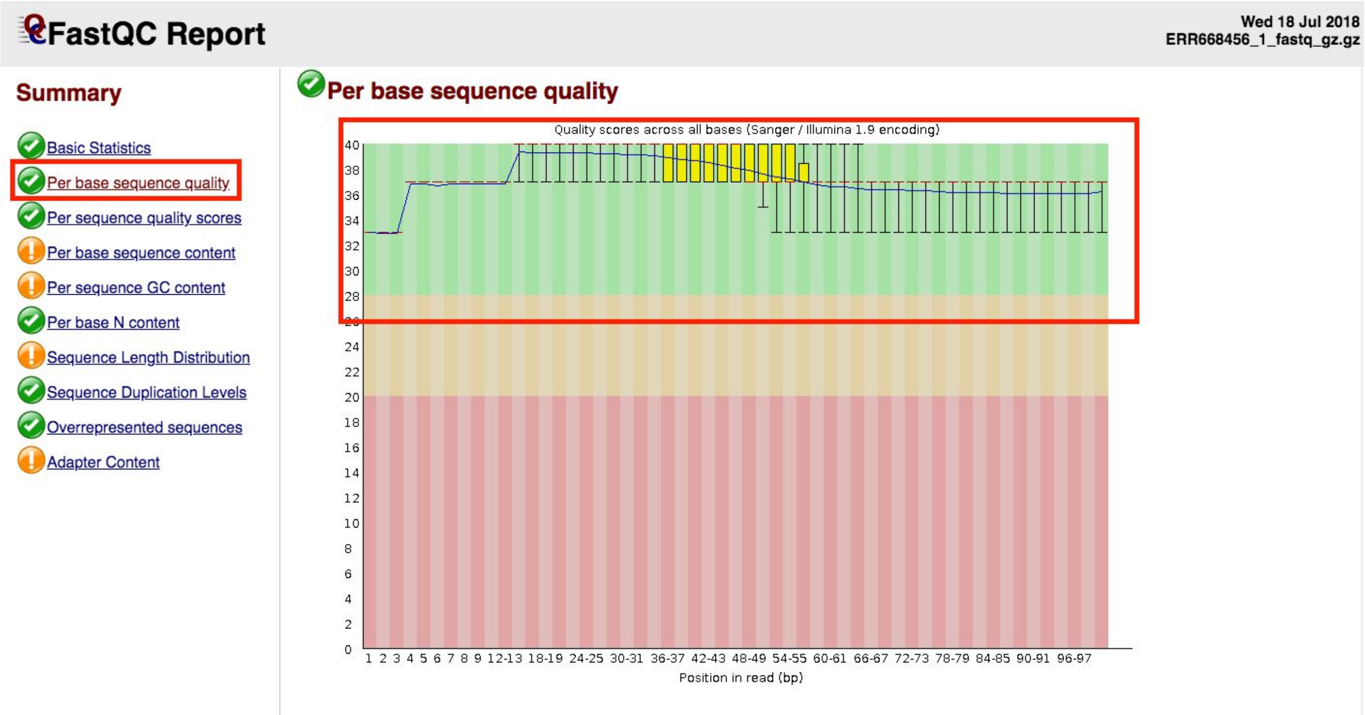
3: FastQC on data 1: Webpage

2: ERR668456_2.fastq.gz

1: ERR668456_1.fastq.gz

When fastqc completes the boxes will turn green. Click on the eye icon 👁 in the box that says 'FastQC .... Web page'

3. In the central data view you will now see the output that has been created when running fastqc



FastQC Report

Wed 18 Jul 2018
ERR668456_1_fastq_gz.gz

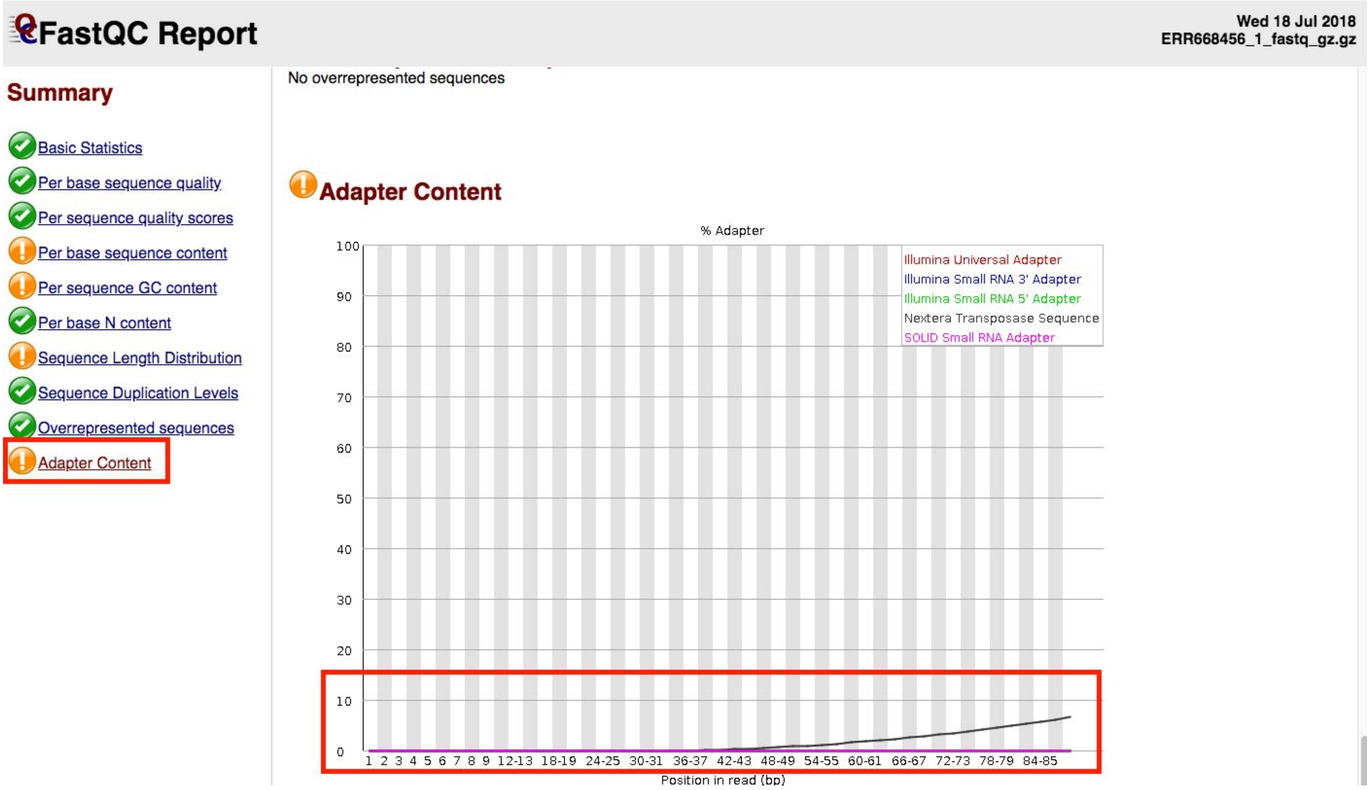
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Adapter Content

✓ Per base sequence quality

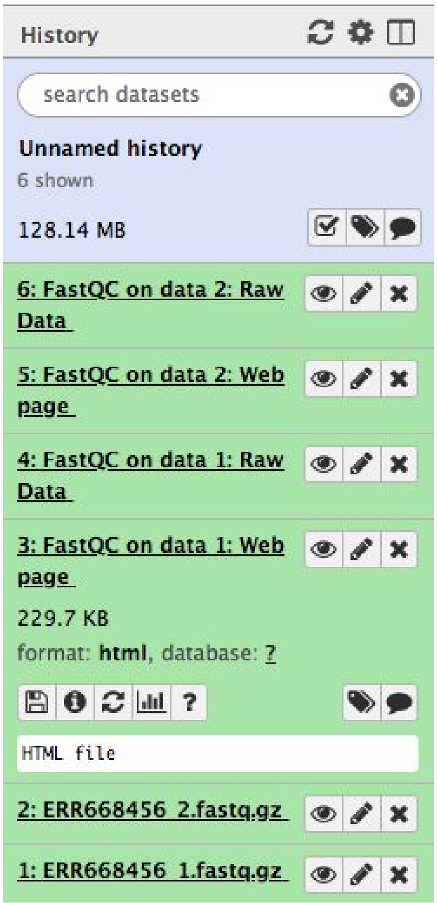
If you click on 'Per base sequence quality' you will see that the quality of this fastq file is very high. Anything above 25 is usually considered good quality.

However if you click on the adapter content link you will see that there is some contamination of the Nextera Transposase sequence suggesting that the sequencing inserts may be small in some of the sequence data. In this case the sequencing has read all the way through from one side of the insert to the other into the adapter sequences on the other side of the sequence insert. Since this is not sequence from the originating DNA this should be removed.



No overrepresented sequences

Before going further run fastqc on the second fastq file. Have a look at the report from this and note if there are any significant differences from the read 1 data. At the end of this your history should look something like this:


History

search datasets

Unnamed history
6 shown

128.14 MB

6: FastQC on data 2: Raw Data

5: FastQC on data 2: Web page

4: FastQC on data 1: Raw Data

3: FastQC on data 1: Web page

229.7 KB

format: **html**, database: ?

HTML file

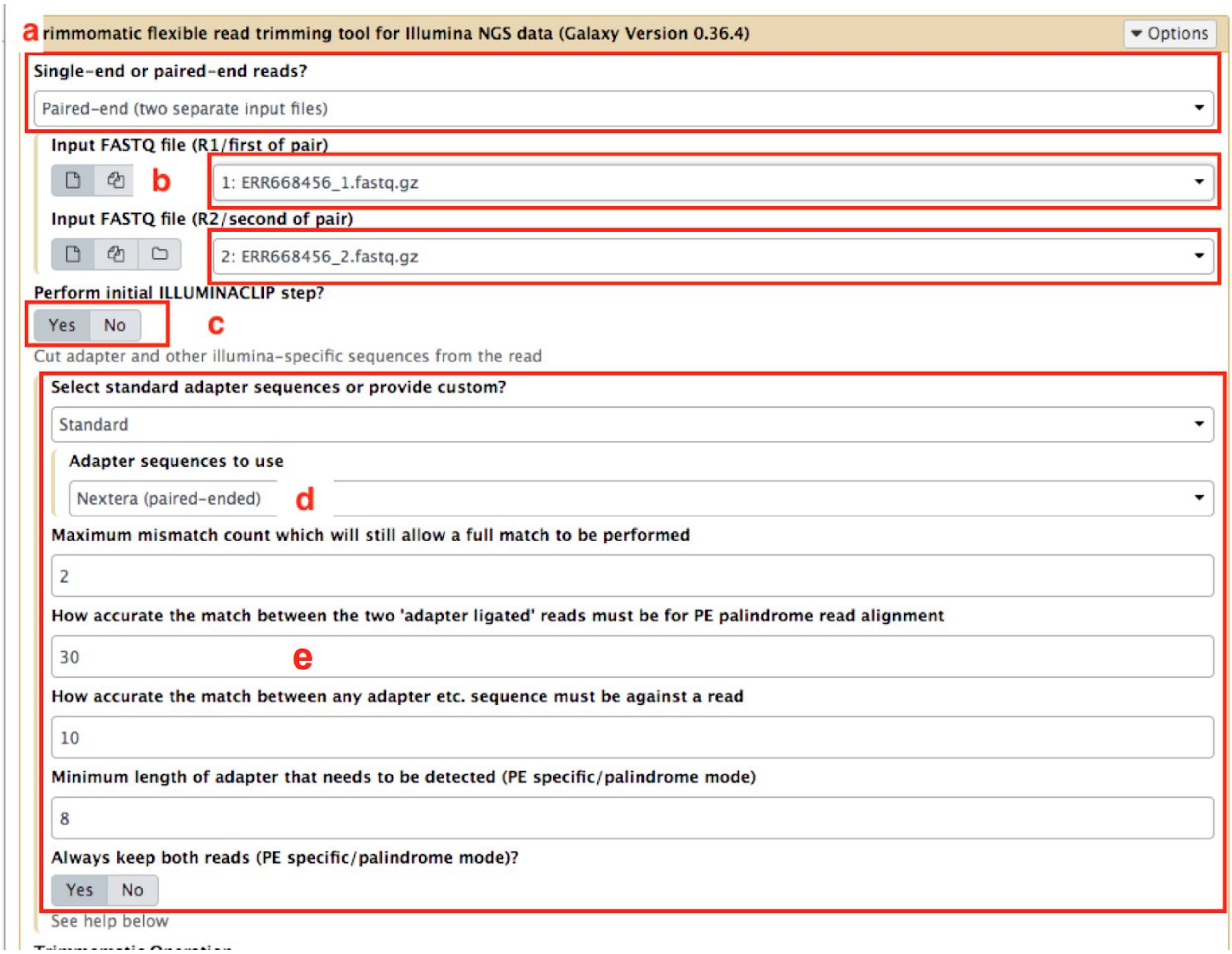
2: ERR668456_2.fastq.gz

1: ERR668456_1.fastq.gz

Fastq trimming

1. Because of the adapter contamination we will use a program called Trimmomatic to remove any low quality data and the adapters. Search for this by typing 'Trimmomatic' into the search box on the left. Click on the link that says 'Trimmomatic flexible read trimming tool for Illumina NGS'

   In the central area the software should now be configured with a few parameters
   a. First select Paired end
   b. Then ensure that you have both the read 1 and read 2 files selected (_1 and _2)
   c. Then click yes to perform the ILLUMINACLIP step and
   d. choose the Nextera (paired-ended) adapter sequences
   e. The other parameters should be left as 2, 30, 10 and 8 respectively



rimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.4)

Single-end or paired-end reads?
Paired-end (two separate input files)

Input FASTQ file (R1/first of pair)
1: ERR668456_1.fastq.gz

Input FASTQ file (R2/second of pair)
2: ERR668456_2.fastq.gz

Perform initial ILLUMINACLIP step?
Yes No
Cut adapter and other illumina-specific sequences from the read

Select standard adapter sequences or provide custom?
Standard

Adapter sequences to use
Nextera (paired-ended)

Maximum mismatch count which will still allow a full match to be performed
2

How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment
30

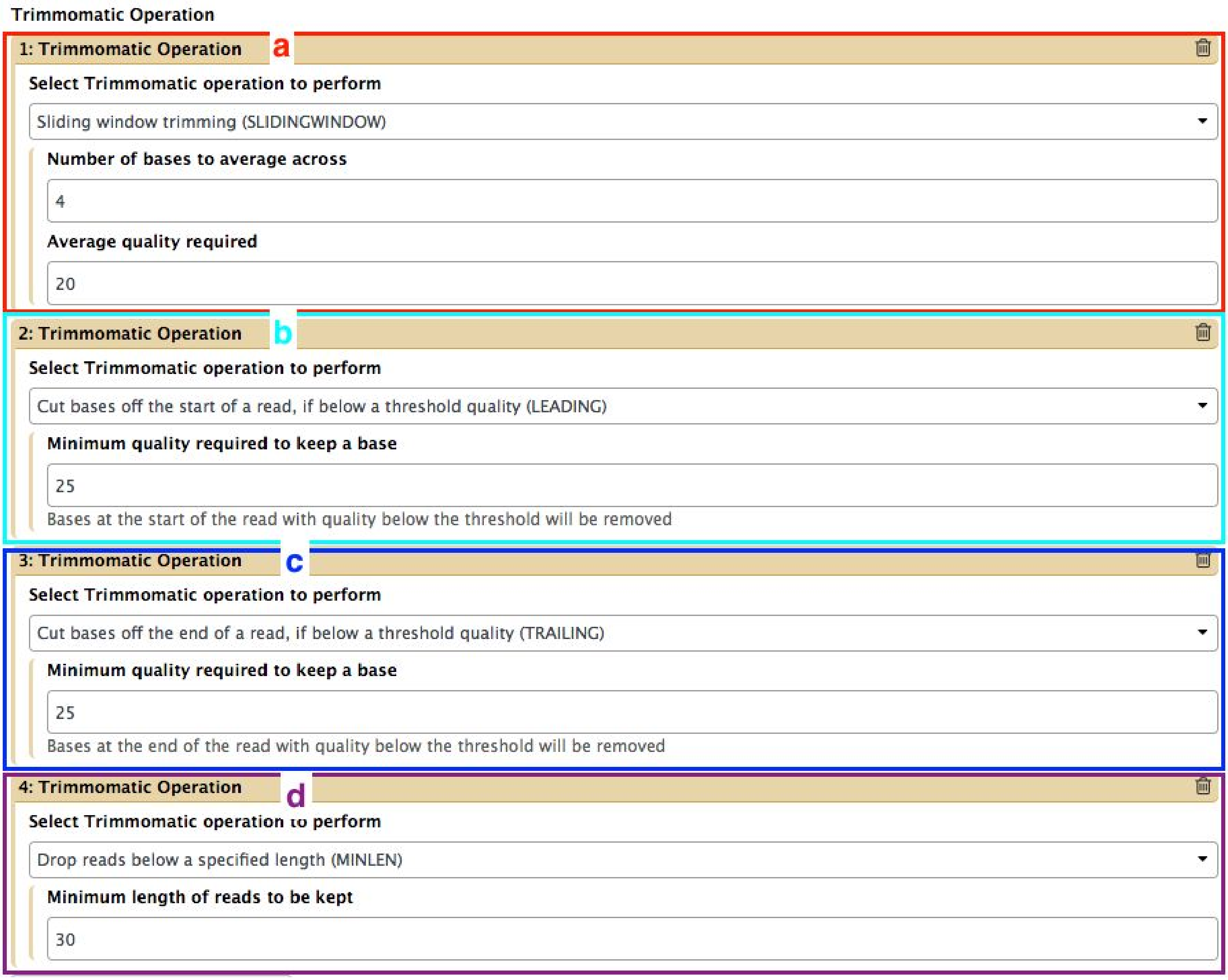
How accurate the match between any adapter etc. sequence must be against a read
10

Minimum length of adapter that needs to be detected (PE specific/palindrome mode)
8

Always keep both reads (PE specific/palindrome mode)?
Yes No
See help below

By specifying this parameter we are telling Trimmomatic to remove Illumina adapters

2. Next you should add some more parameters for trimming. These will be
    a. SLIDINGWINDOW (remove areas where the average quality is less than 20 across 4 base)
    b. LEADING (cut off low quality bases at the 5’ end below 25)
    c. TRAILING (cut off low quality bases at the 3’ end below 25)
    d. MINLEN (remove read pairs if either is less than 30 bases after trimming)



Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Cut bases off the start of a read, if below a threshold quality (LEADING)

Minimum quality required to keep a base

25

Bases at the start of the read with quality below the threshold will be removed

3: Trimmomatic Operation

Select Trimmomatic operation to perform

Cut bases off the end of a read, if below a threshold quality (TRAILING)

Minimum quality required to keep a base

25

Bases at the end of the read with quality below the threshold will be removed

4: Trimmomatic Operation

Select Trimmomatic operation to perform

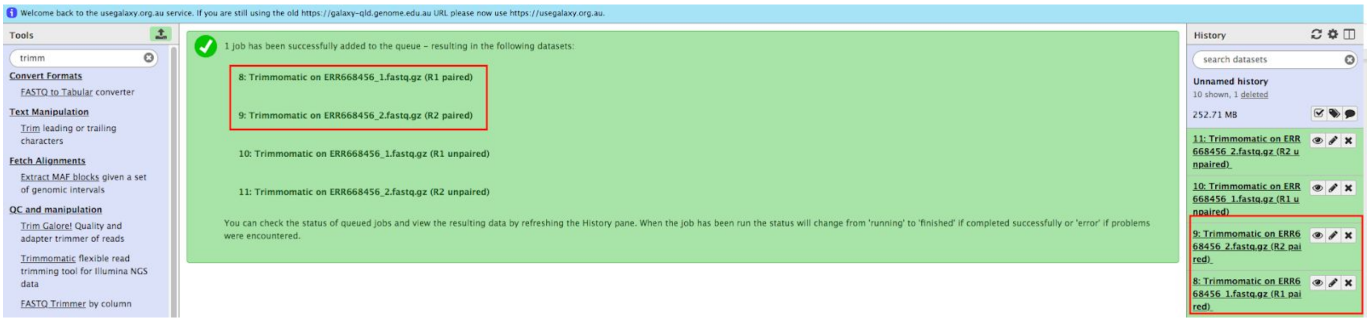
Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

30

These extra settings are added using the **+ Insert Trimmomatic Operation** button. Once you have entered all of these click the blue 'Execute button'

11. When trimmomatic has finished you will see 4 files have been produced it is the R1_paired and R2_paired files that we will use for assembly. These are the fastq files that have been trimmed for quality and adapter sequences. They are called _paired since these are the paired reads that have been kept. For downstream processes we do not want any unpaired reads.



Welcome back to the usegalaxy.org.au service. If you are still using the old https://galaxy-qld.genome.edu.au URL please now use https://usegalaxy.org.au.

Tools
trimm
Convert Formats
FASTQ to Tabular converter
Text Manipulation
Trim leading or trailing characters
Fetch Alignments
Extract MAF blocks given a set of genomic intervals
QC and manipulation
Trim Galore! Quality and adapter trimmer of reads
Trimmomatic flexible read trimming tool for Illumina NGS data
FASTQ Trimmer by column

✓ 1 job has been successfully added to the queue – resulting in the following datasets:

8: Trimmomatic on ERR668456_1.fastq.gz (R1 paired)

9: Trimmomatic on ERR668456_2.fastq.gz (R2 paired)

10: Trimmomatic on ERR668456_1.fastq.gz (R1 unpaired)

11: Trimmomatic on ERR668456_2.fastq.gz (R2 unpaired)

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

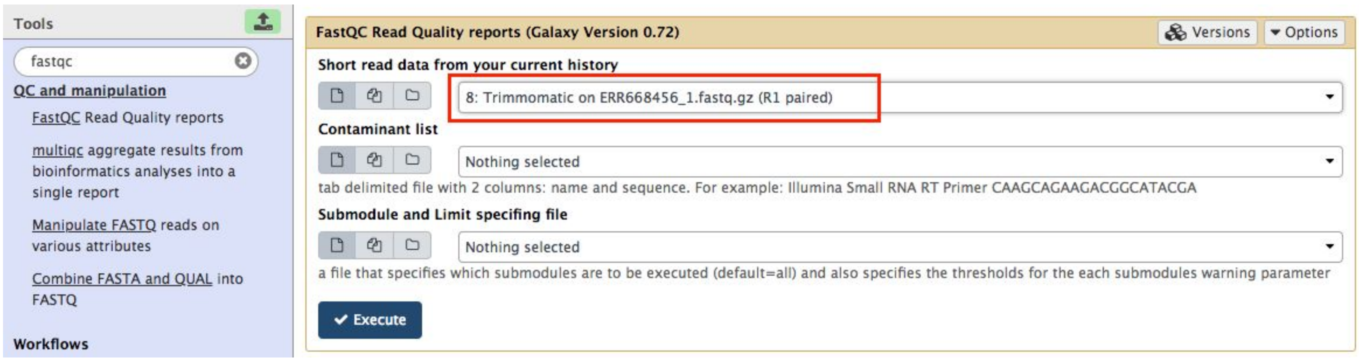
Unnamed history
10 shown, 1 deleted
252.71 MB

11: Trimmomatic on ERR668456_2.fastq.gz (R2 unpaired)
10: Trimmomatic on ERR668456_1.fastq.gz (R1 unpaired)
9: Trimmomatic on ERR668456_2.fastq.gz (R2 paired)
8: Trimmomatic on ERR668456_1.fastq.gz (R1 paired)

Before assembling them however, you should first check on the quality.

Trimmed fastq QC assessment

1. Run fastqc again but this time select first the trimmed R1 paired file and then the trimmed R2 paired file.



Tools
fastqc
QC and manipulation
FastQC Read Quality reports
multiqc aggregate results from bioinformatics analyses into a single report
Manipulate FASTQ reads on various attributes
Combine FASTA and QUAL into FASTQ
Workflows
Versions
Options
Execute

FastQC Read Quality reports (Galaxy Version 0.72)

Short read data from your current history: 8: Trimmomatic on ERR668456_1.fastq.gz (R1 paired)

Contaminant list: Nothing selected

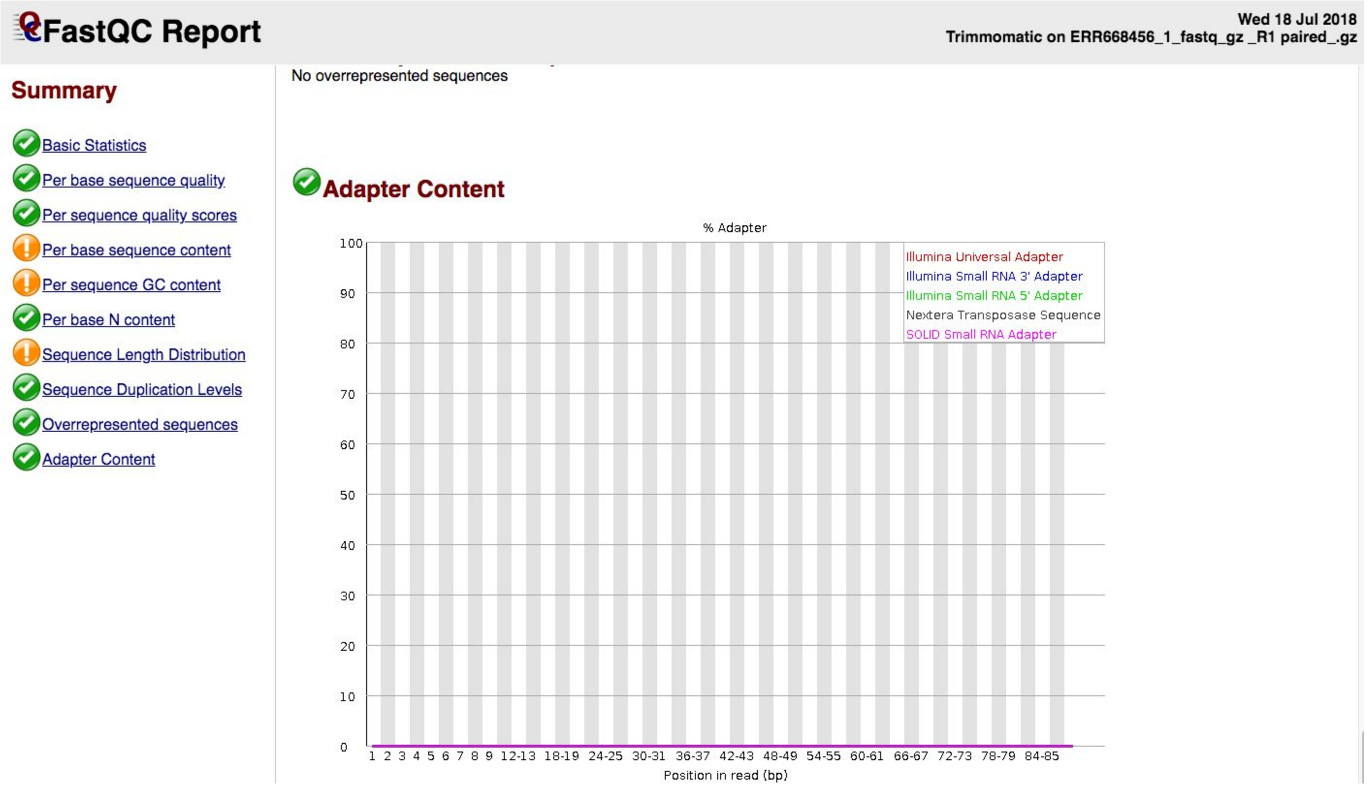
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifing file: Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Make sure you do this for both files.

2. Look at the output and you will see that the adapters have been removed



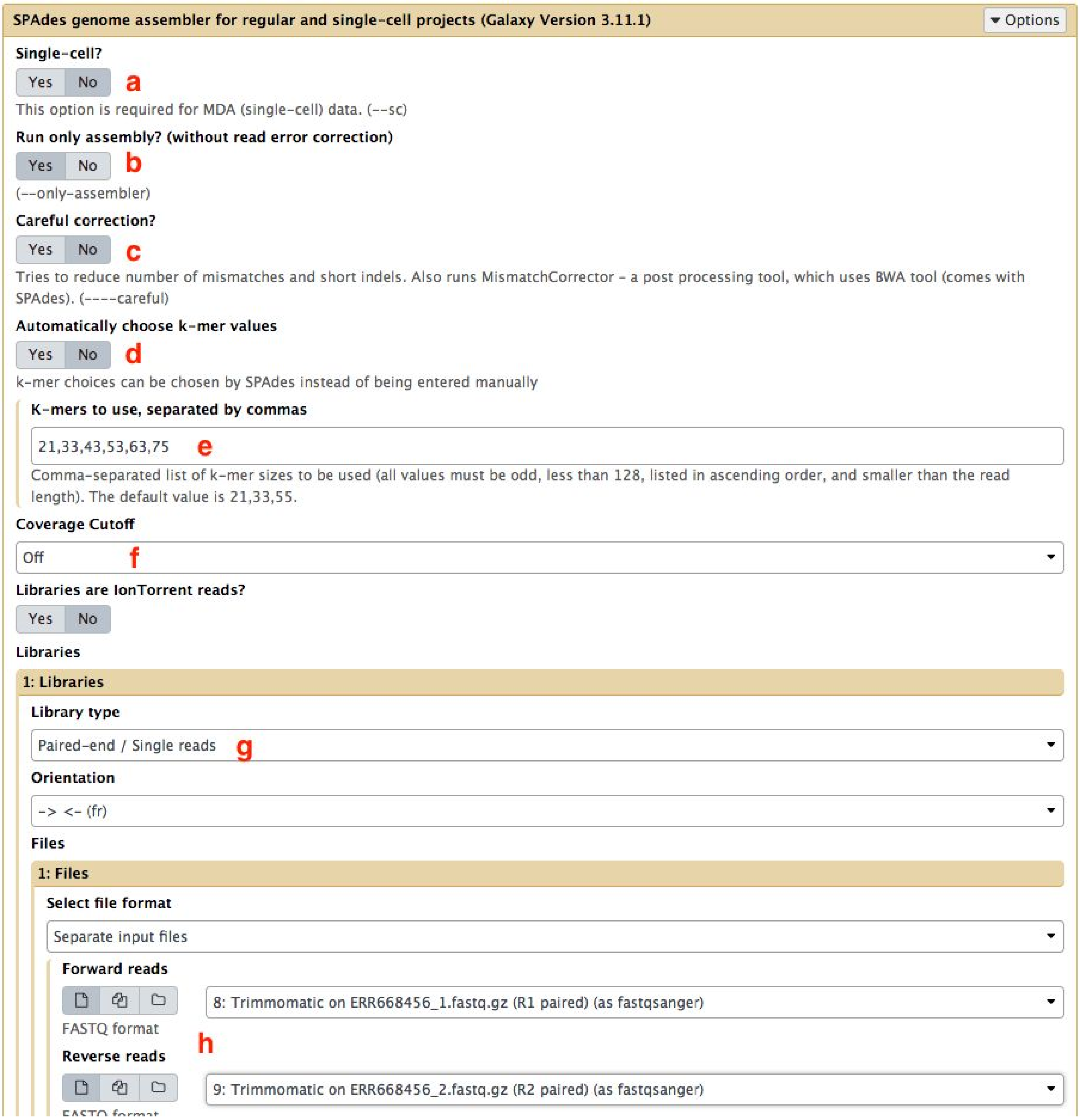
FastQC Report
Wed 18 Jul 2018
Trimmomatic on ERR668456_1_fastq_gz _R1 paired_.gz
Summary
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
No overrepresented sequences
Adapter Content
% Adapter
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter

Assembly

1. Finally we come to assembly. Search for SPAdes in the tool search box.
   Enter the following options
   a. Single-cell: No
   b. Run only assembly: Yes
   c. Careful correction: No
   d. Automatically choose k-mer values: No
   e. K-mers to use: 21,33,43,53,63,75
   f. Coverage Cutoff: Off

g. Library type: Paired-end/ Single Reads
h. select the R1 paired and R2 paired files from the previous Trimmomatic step

Click on the blue Execute button to run SPAdes



SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.11.1)

Single-cell?

Yes No

This option is required for MDA (single-cell) data. (--sc)

Run only assembly? (without read error correction)

Yes No

(--only-assembler)

Careful correction?

Yes No

Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector - a post processing tool, which uses BWA tool (comes with SPAdes). (----careful)

Automatically choose k-mer values

Yes No

k-mer choices can be chosen by SPAdes instead of being entered manually

K-mers to use, separated by commas

21,33,43,53,63,75

Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.

Coverage Cutoff

Off

Libraries are IonTorrent reads?

Yes No

Libraries

1: Libraries

Library type

Paired-end / Single reads

Orientation

-> <- (fr)

Files

1: Files

Select file format

Separate input files

Forward reads

8: Trimmomatic on ERR668456_1.fastq.gz (R1 paired) (as fastqsanger)

FASTQ format

Reverse reads

9: Trimmomatic on ERR668456_2.fastq.gz (R2 paired) (as fastqsanger)

FASTQ format

Please note that because Galaxy is a shared resource it may sometime take a while for the SPAdes assembly to start (boxes turn yellow). Patience may be required :)

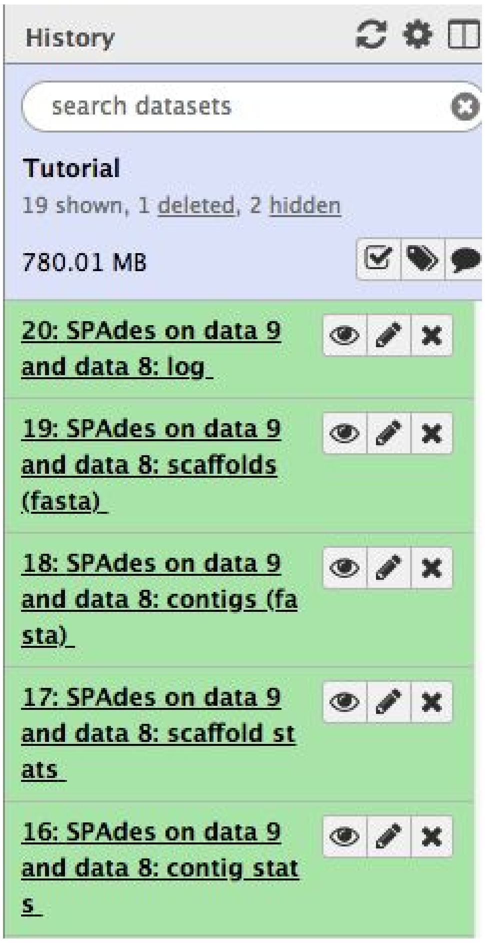
While you are waiting

You may want to read

- the original SPAdes publication: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342519
- the SPAdes manual: http://spades.bioinf.spbau.ru/release3.11.1/manual.html

- And if you are really keen this primer about bacterial WGS http://cmr.asm.org/content/29/4/881.full.pdf

2. When SPAdes is finished your history should look something like this



History

search datasets

Tutorial

19 shown, 1 deleted, 2 hidden

780.01 MB

20: SPAdes on data 9 and data 8: log

19: SPAdes on data 9 and data 8: scaffolds (fasta)

18: SPAdes on data 9 and data 8: contigs (fasta)

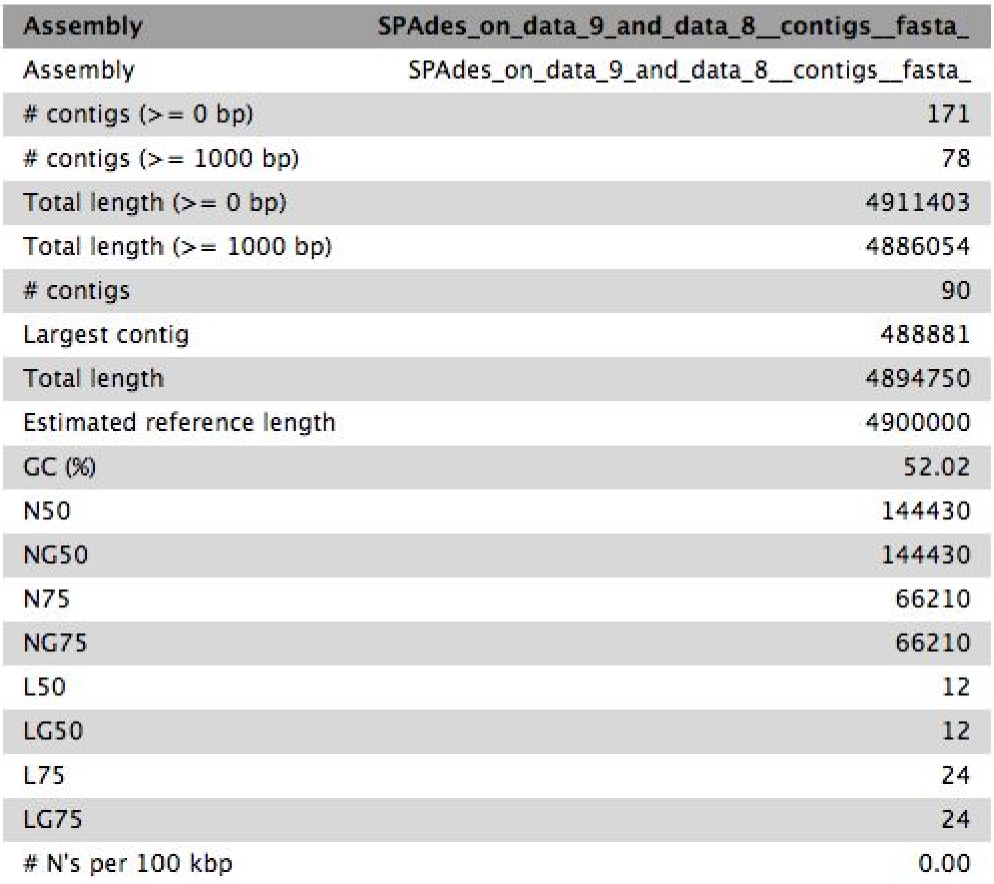
17: SPAdes on data 9 and data 8: scaffold stats

16: SPAdes on data 9 and data 8: contig stats

The outputs you will see are
scaffolds.fasta contains resulting scaffolds in fasta format (contigs joined by paired end read information but where the gaps are padded with Ns, see this article for some more details https://genome.jgi.doe.gov/help/scaffolds.jsf)
contigs.fasta contains resulting contigs in fasta format
scaffolds stats length and coverage information about each resulting scaffold
contigs stats length and coverage information about each resulting contig

Assembly QC assessment

3. Now you will assess the quality of the assembly using Quast. Type Quast into the tool search bar. Select the SPAdes **contigs** file and specify a reference genome size of 4900000 (approximate size of *S.typhi*). Click on execute to run the Quast quality assessment. The report.tsv output will look like this



Assembly	SPAdes_on_data_9_and_data_8__contigs__fasta_
Assembly	SPAdes_on_data_9_and_data_8__contigs__fasta_
# contigs (>= 0 bp)	171
# contigs (>= 1000 bp)	78
Total length (>= 0 bp)	4911403
Total length (>= 1000 bp)	4886054
# contigs	90
Largest contig	488881
Total length	4894750
Estimated reference length	4900000
GC (%)	52.02
N50	144430
NG50	144430
N75	66210
NG75	66210
L50	12
LG50	12
L75	24
LG75	24
# N's per 100 kbp	0.00

The critical figures to look at are

a. # (number) of contigs: smaller the better
b. Total length: Should be approximately the size of the genome expected for the species
c. N50/NG50: Larger the better (see this article for an explanation)

Conclusions

Well done you have completed a genome assembly and simple QC. The next step will be to dive into the command line and do the same from there.

Global Health Research Unit

FUNDED BY NIHR | National Institute for Health Research

Version Control Table

Title	Genome Assembly Tutorial			
Description	A document describing how to perform *de novo* genome assembly for a single bacterial genome starting from a pair of Fastq files generated with short read sequencing using the Galaxy web platform			
Created By	Anthony Underwood			
Date Created	27th July 2018			
Maintained By	Anthony Underwood			
Version Number	**Modified By**	**Modifications Made**	**Date Modified**	**Status**
1.0	Anthony Underwood	First version	27th July 2018	First Live Version