- Useful links/tutorials: Information on mapping Fastq manipulation & SAM/BAM

1) You are provided with 2 datasets (one contains forward reads, one contains reverse reads). Create a new history, unzip and upload the data to your history and load them into FastQC and MultiQC. Comment on the results.
    a) How similar are the results for the two datasets?
    b) What kind of problems there are in the datasets (share screenshots of the issues you find)?
    c) Share a link to your history.

2) Create a new history. Repeat the steps (FastQC and MultiQC) in part 1 for the following datasets (see **Assignment3_part2.pdf** to help you with finding the data and the BWA-MEM step):

- ZR751 paired-end RNA-seq subsampled (end 1)
- ZR751 paired-end RNA-seq subsampled (end 2)

a) How similar are the results for the two datasets?

b) What kind of problems there are in the datasets (share screenshots of the issues you find)?

c) Use 'Map with BWA-MEM' tool to map the data on the human genome.

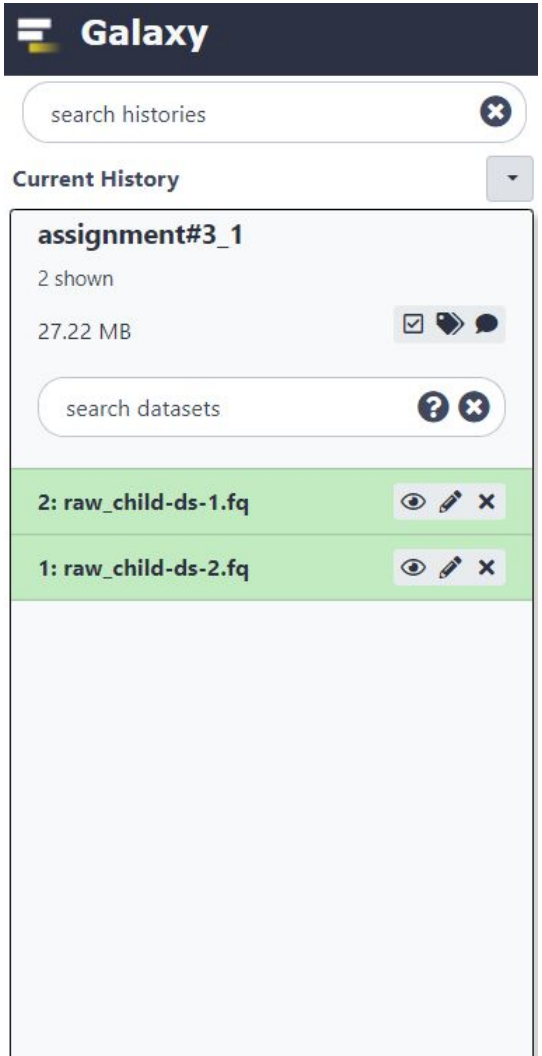
d) Share a link to your history.

3) Create a new history. Choose 6 Illumina iDEA datasets (like how you did in part 2). This time, use fastp instead of FastQC to quality-check your data. Share screenshots

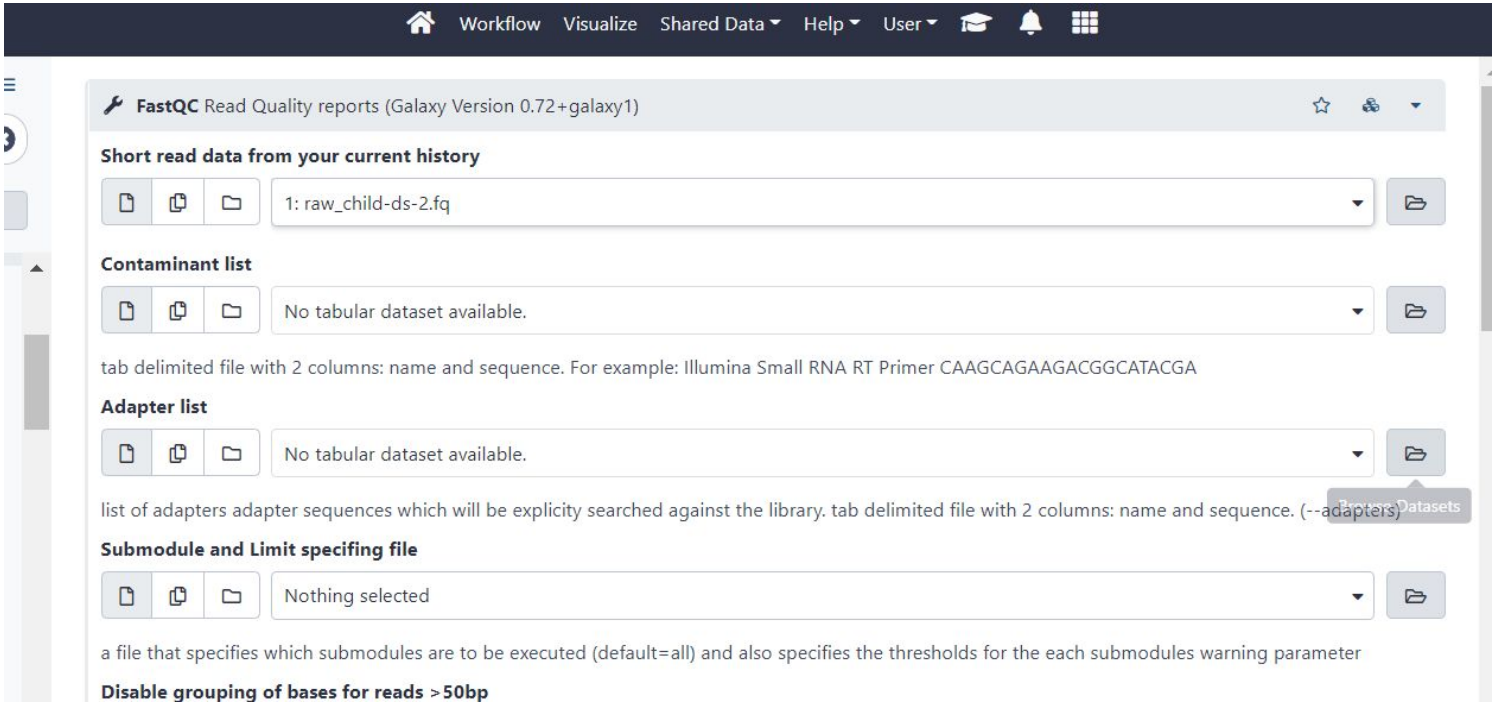
of the graphs you get and explain what you see (You don't have to share a screenshot for 6 different datasets if they are all similar. Focus on the problems and differences you observe.). **Share a link to your history.**

4) Create a new history. Complete this tutorial. Add screenshots to the steps you take. Shortly explain what you have done. **Share a link to your history.**

https://usegalaxy.org/u/leman_nur_nehri/h/assignment31



FastQC for data#1



Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

- **1: raw_child-ds-2.fq**

It produces 2 outputs:

- **3: FastQC on data 1: Webpage**
- **4: FastQC on data 1: RawData**

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

fastQC execution for data#2

Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

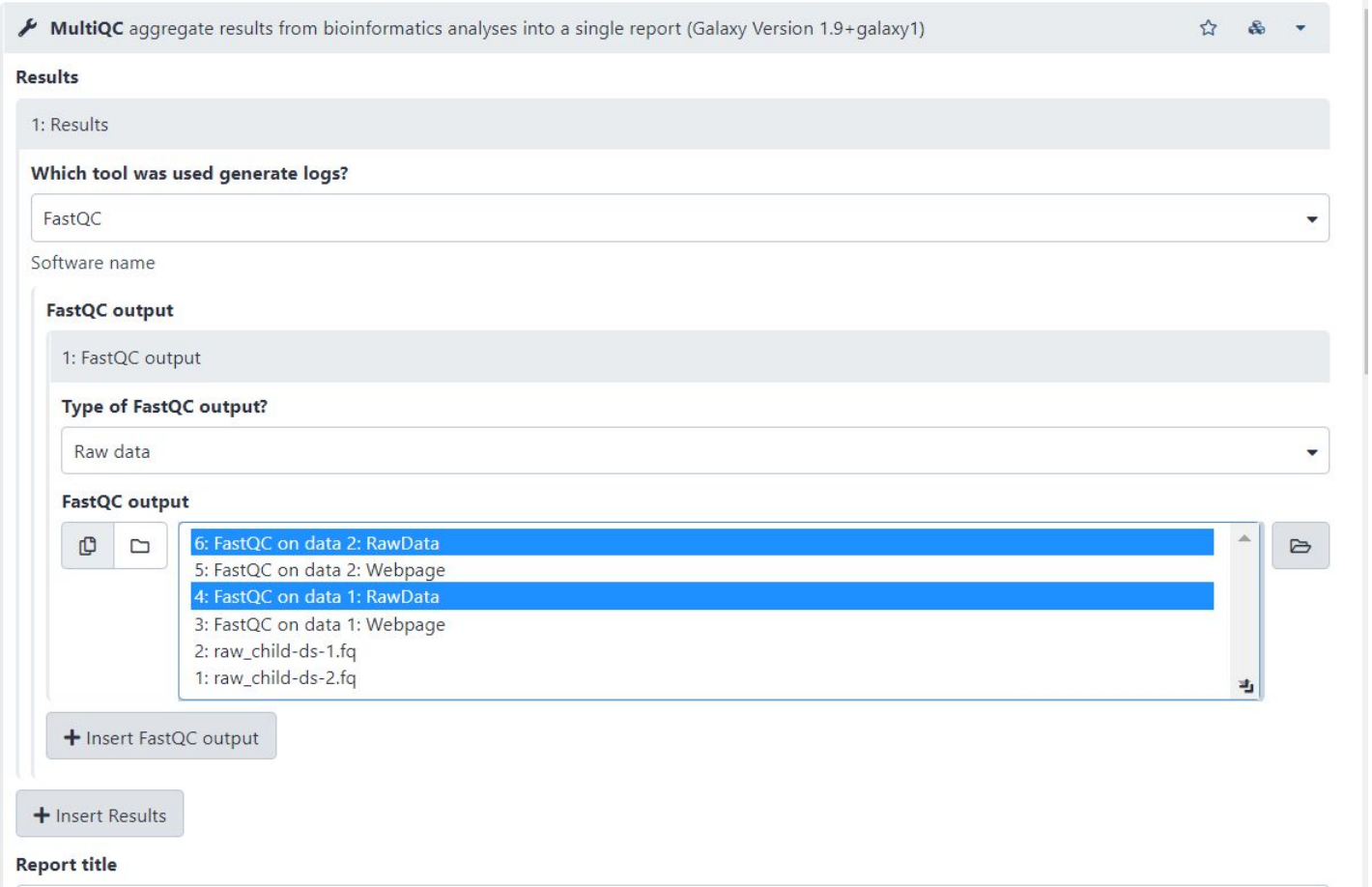
- **2: raw_child-ds-1.fq**

It produces 2 outputs:

- **5: FastQC on data 2: Webpage**
- **6: FastQC on data 2: RawData**

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

multiQC


MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.9+galaxy1)
Results
1: Results
Which tool was used generate logs?
FastQC
Software name
FastQC output
1: FastQC output
Type of FastQC output?
Raw data
FastQC output
6: FastQC on data 2: RawData
5: FastQC on data 2: Webpage
4: FastQC on data 1: RawData
3: FastQC on data 1: Webpage
2: raw_child-ds-1.fq
1: raw_child-ds-2.fq
Insert FastQC output
Insert Results
Report title

Executed **MultiQC** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- **4: FastQC on data 1: RawData**
- **6: FastQC on data 2: RawData**

It produces this output:

- **8: MultiQC on data 6 and data 4: Webpage**

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

RESULTS

FastQC works on only single read sets

MultiQC aggregates the FastQC results in one report

multiQC

duplicate reads for each seq.

average GC content for each seq.

average seq. length

failed modules in fastQC report

total seq.s as million

General Statistics

Copy table | Configure Columns | Plot | Showing 2/2 rows and 5/5 columns.

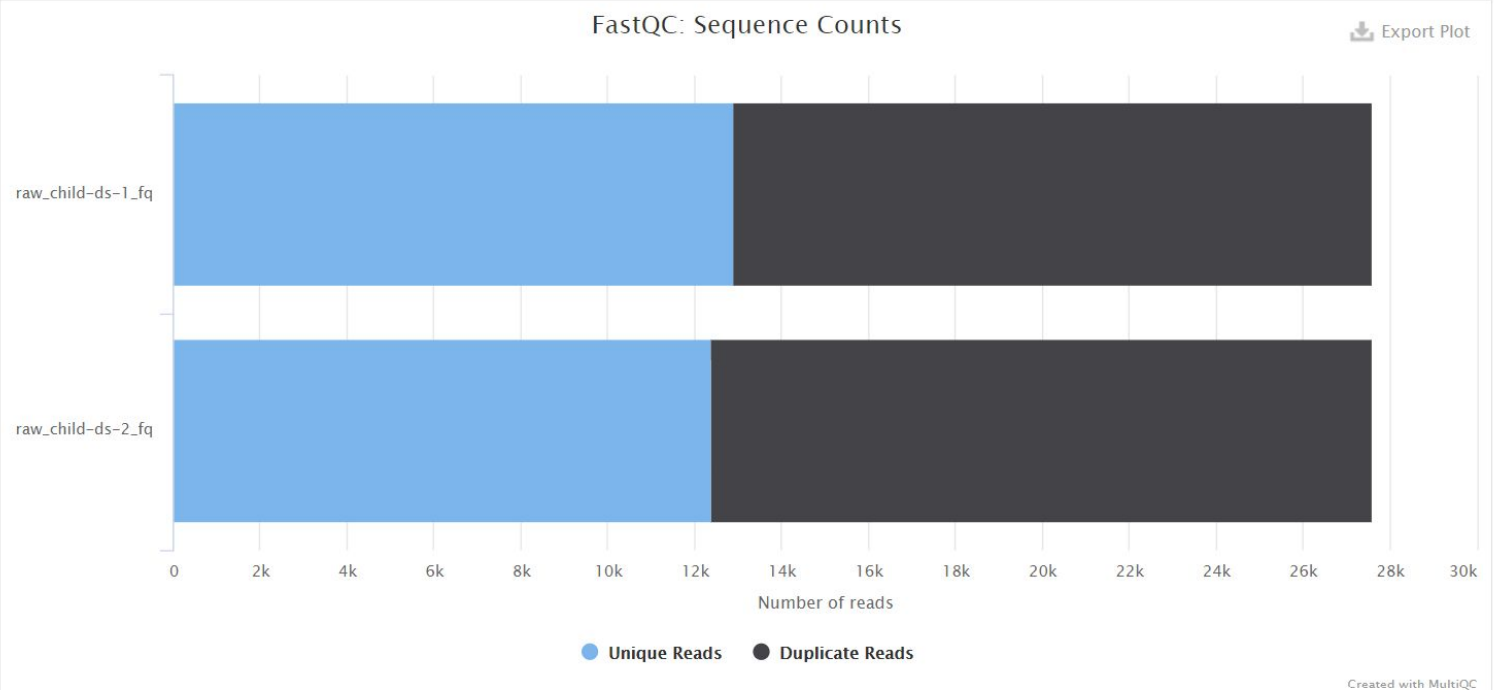
Sample Name	% Dups	% GC	Length	% Failed	M Seqs
raw_child-ds-1_fq	53.2%	44%	233 bp	18%	0.0
raw_child-ds-2_fq	55.1%	44%	233 bp	27%	0.0

FastQC

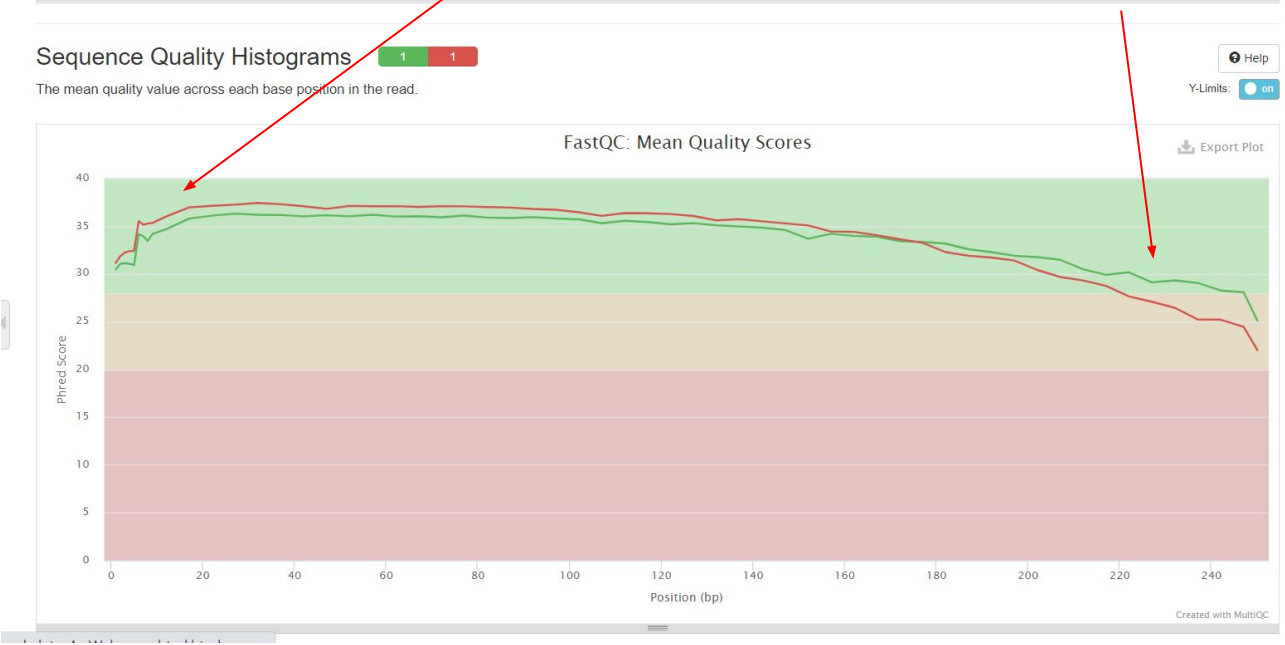
the degree of duplication for every sequence

Sequence counts for each sample. Duplicate read counts are an estimate only.

Number of reads | Percentages



overview of the qualities as mean (general): in the beginning, quality scores are high for sequences, and lower in the end of the sequences, it is expected, there is known reasons for that, this is an expected error



overview of the qualities across all bases

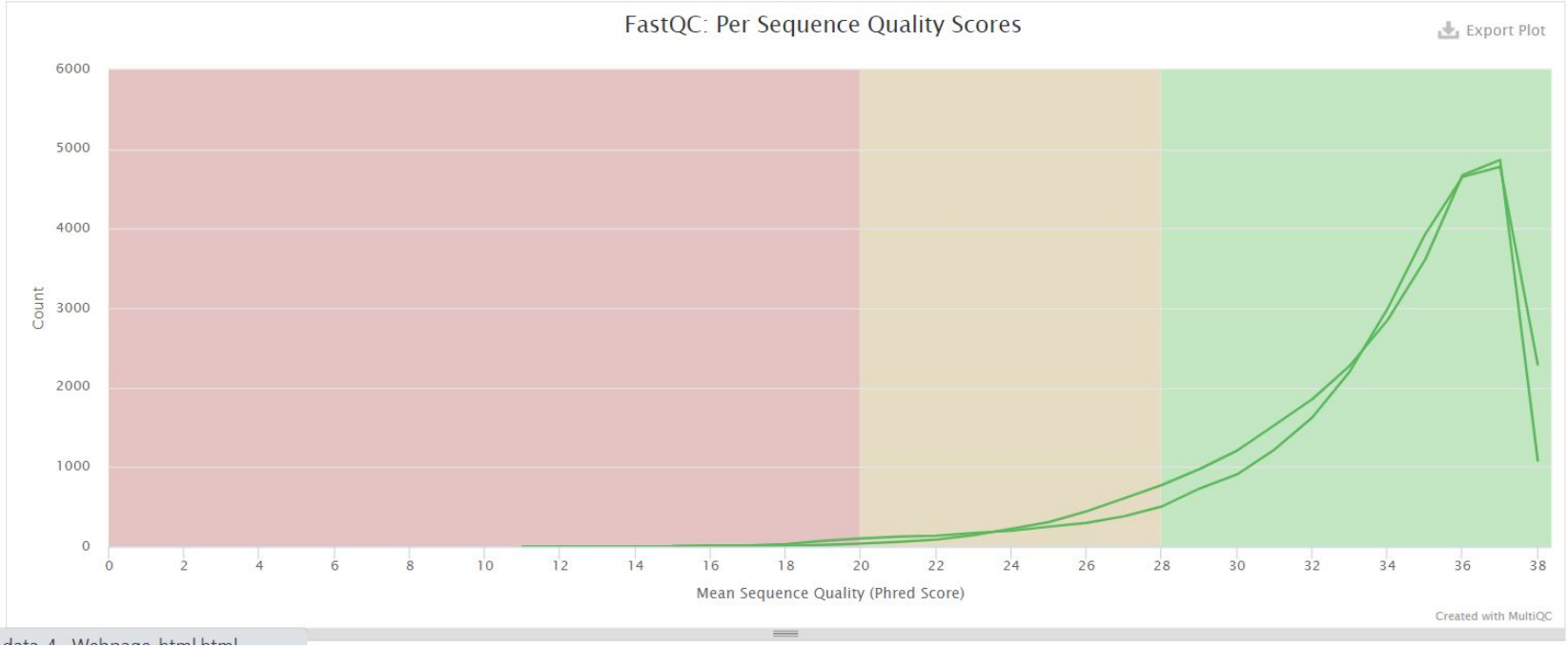
Per Sequence Quality Scores

2

Help

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

Y-Limits: on



proportion of each base position

Per Base Sequence Content

2

Help

The proportion of each base position for which each of the four normal DNA bases has been called.

Click a sample row to see a line plot for that dataset.

Rollover for sample name

Position: - %T: - %C: - %A: - %G: -

Export Plot

GC content for sequences

Per Sequence GC Content

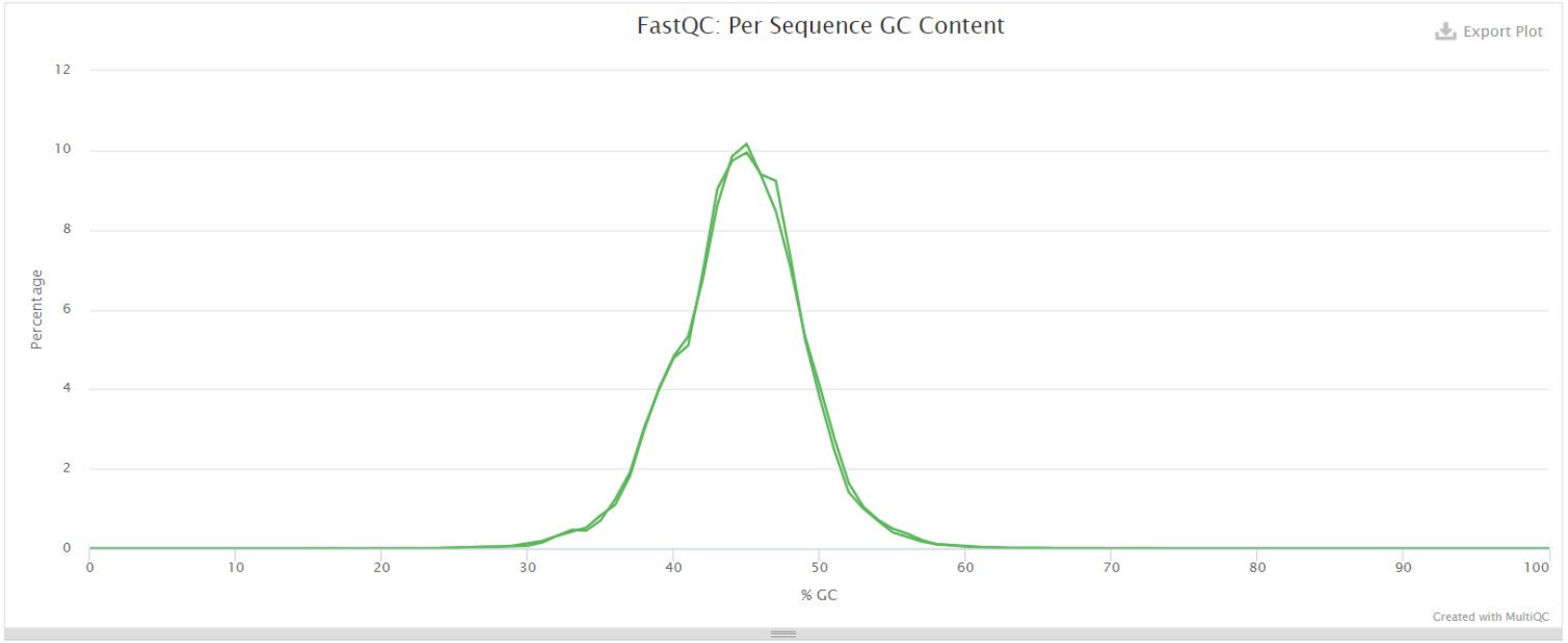
2

Help

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Y-Limits: on

Percentages | Counts



N content for bases

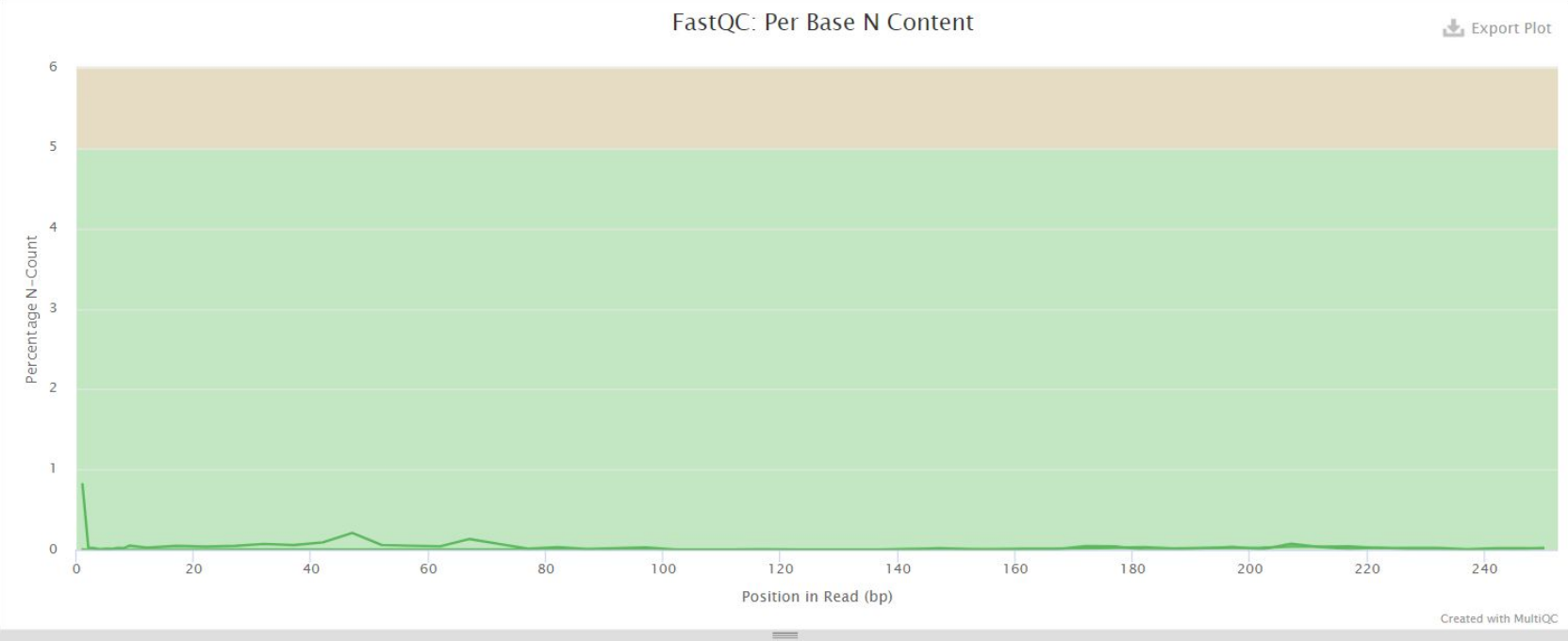
Per Base N Content

2

Help

The percentage of base calls at each position for which an `N` was called.

Y-Limits: on



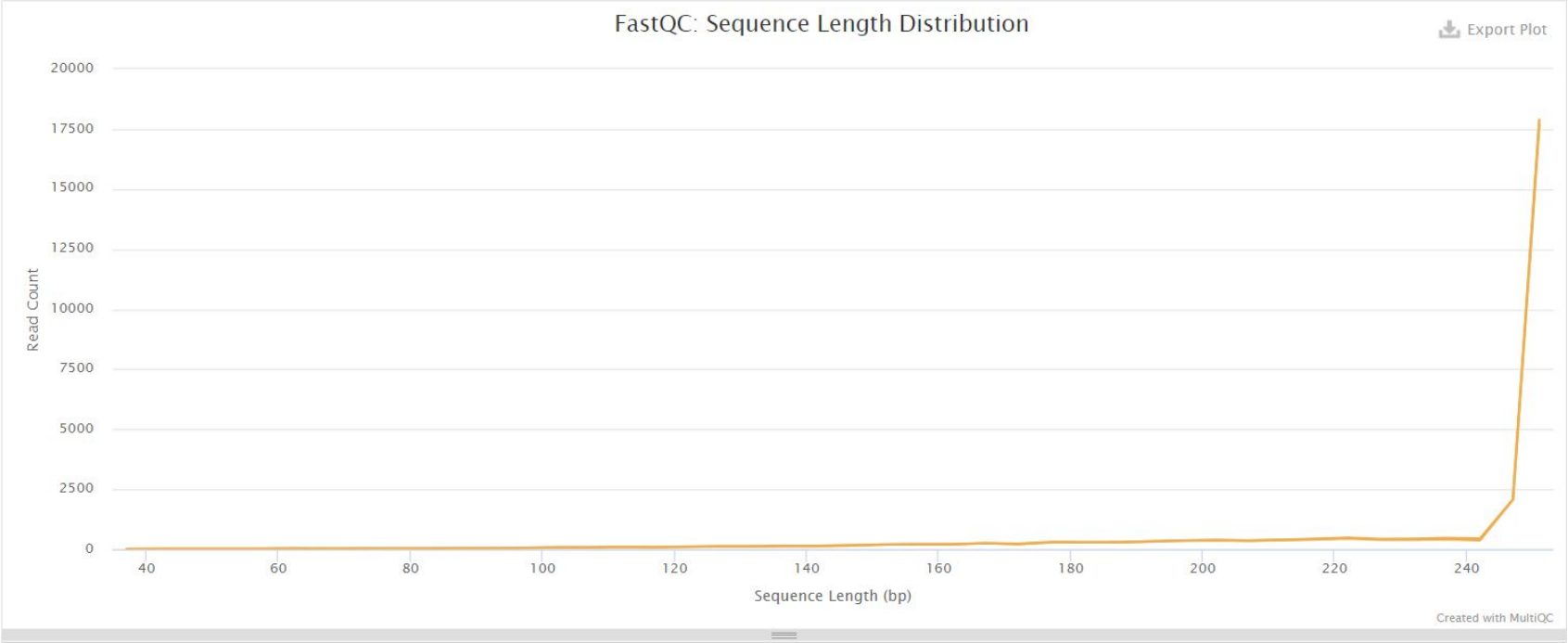
sequence lenghts

Sequence Length Distribution

2

The distribution of fragment sizes (read lengths) found. See the FastQC help

Y-Limits: on



Toolbox

duplication levels for each sequence

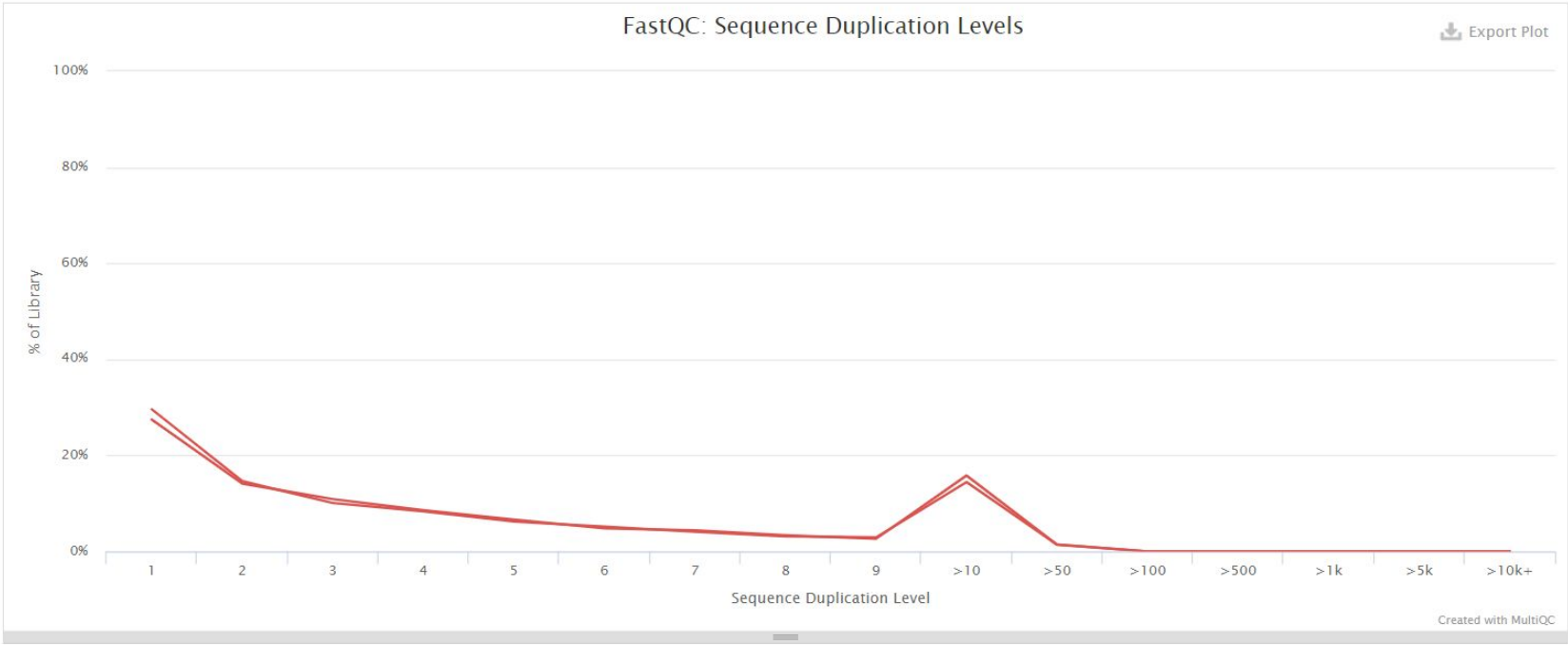
Sequence Duplication Levels

2

Help

The relative level of duplication found for every sequence.

Y-Limits: on

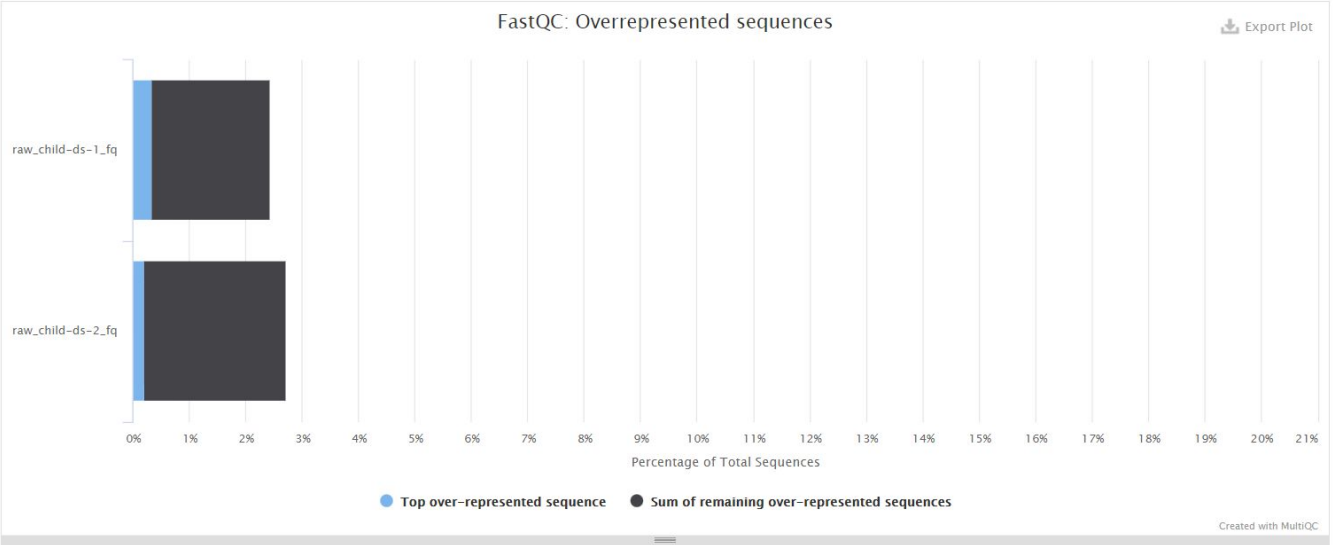


sequences which appear more than the expected: ERROR: the overrepresented sequences are the unexpected ones and therefore it can be counted as errors

Overrepresented sequences 2

Help

The total amount of overrepresented sequences found in each library.



Toolbox

reads of known adapters in the seqences

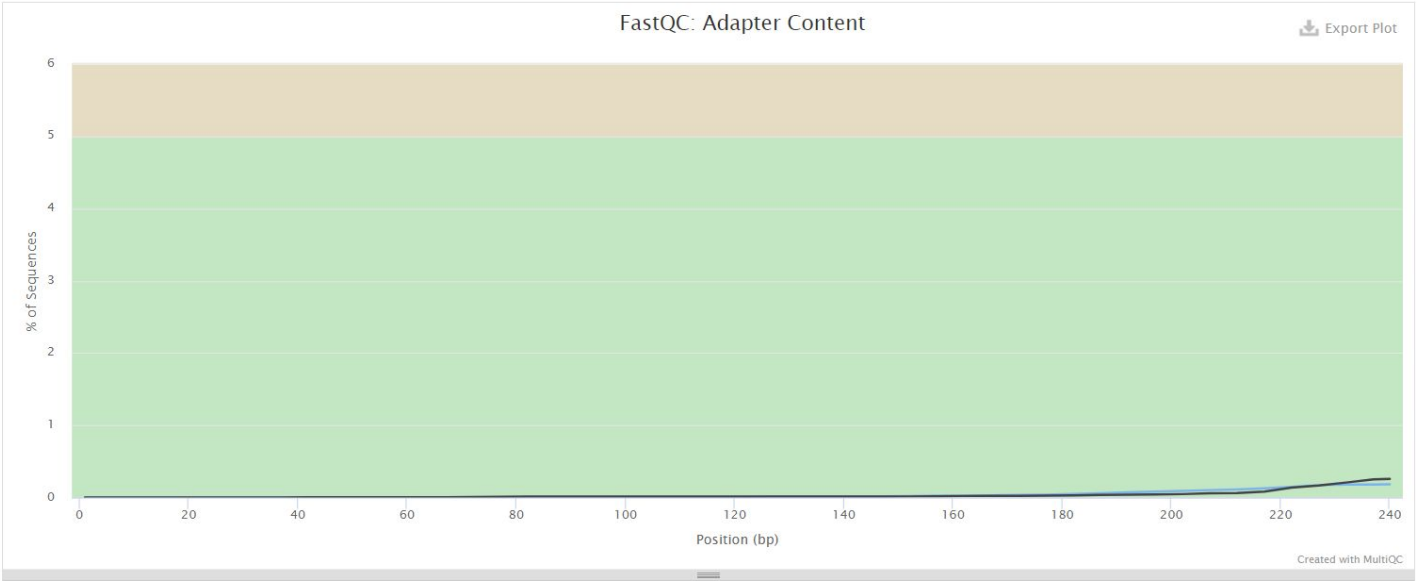
Adapter Content

2

Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Y-Limits: on

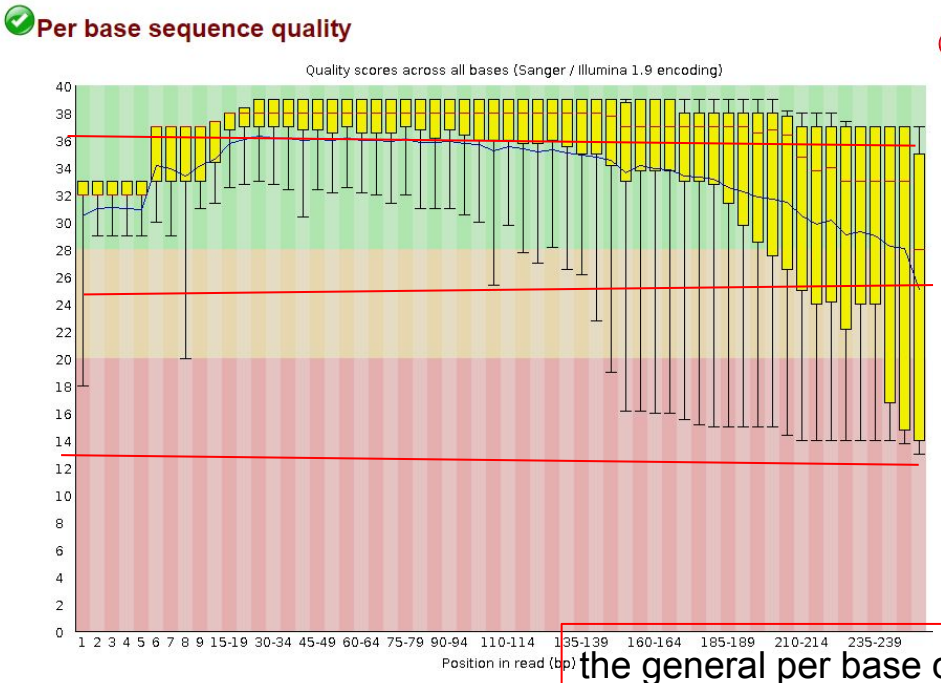


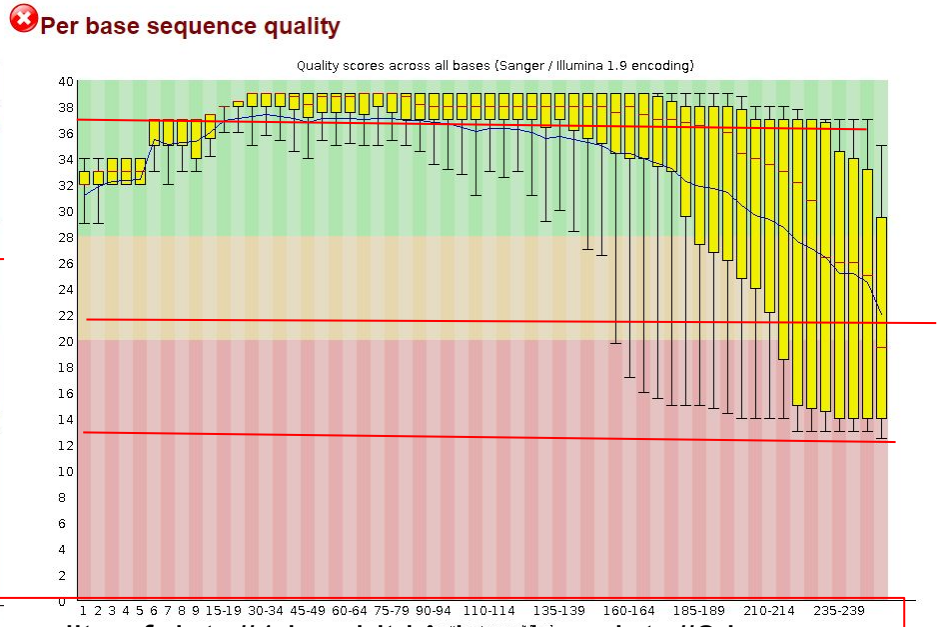
Toolbox

the only difference between the 2 seq.s is per base sequence quality: the 2.fq data quality is bad comparing to the data 1.fq.

general summary of the statistics of the analyses: except the the per base seq. quality, all other statistics are similar to each other for the two dataset.

FASTQC results: left images for data#1, right images for data#2

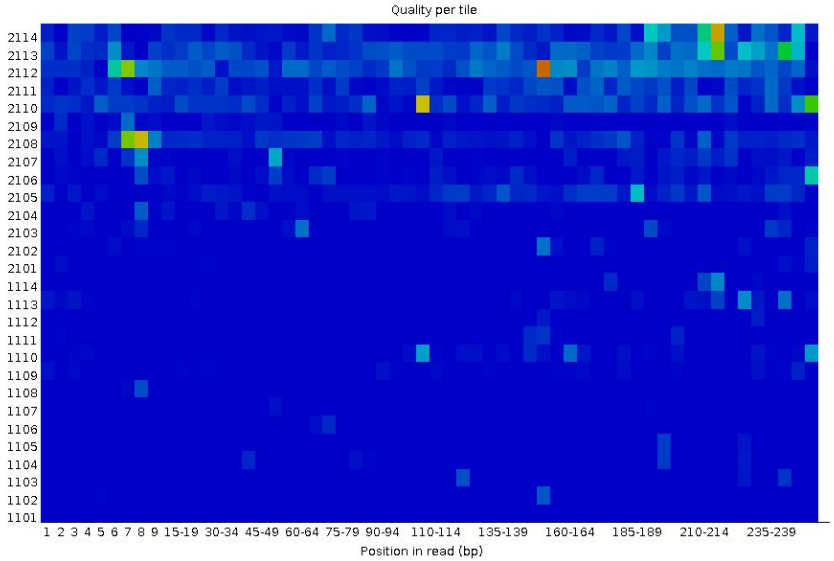




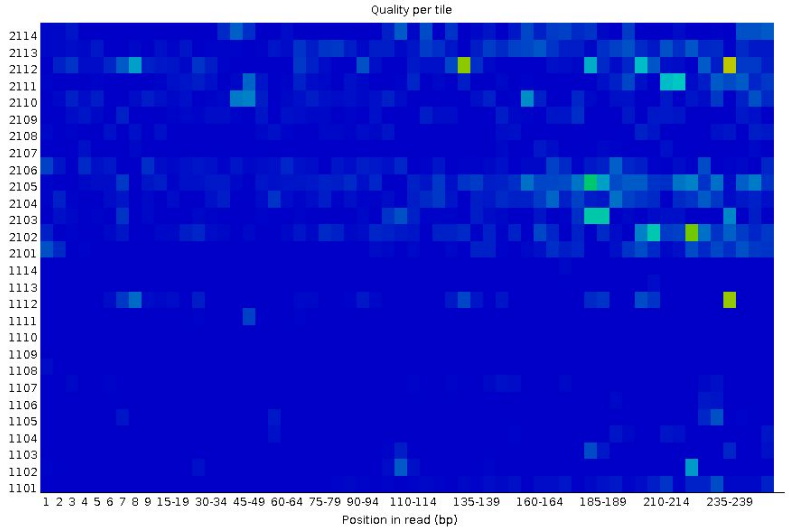
the general per base quality of data#1 is a bit higher than data#2 by comparing the average per base quantities and standard deviations of the qualities, but in general they can be counted as similar

deviation from the average quality for each tile, hotter colors are bad qualities, cooler colors are better qualities==average. in general, the qualities are similar but the qualities are more better in second data set comparison to the first for considering per tile

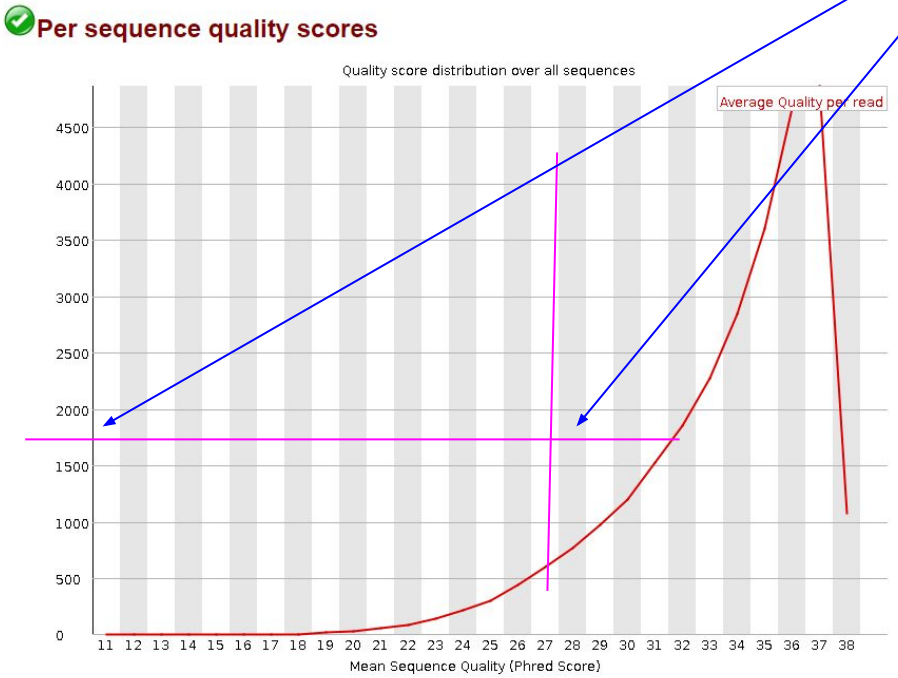
Per tile sequence quality

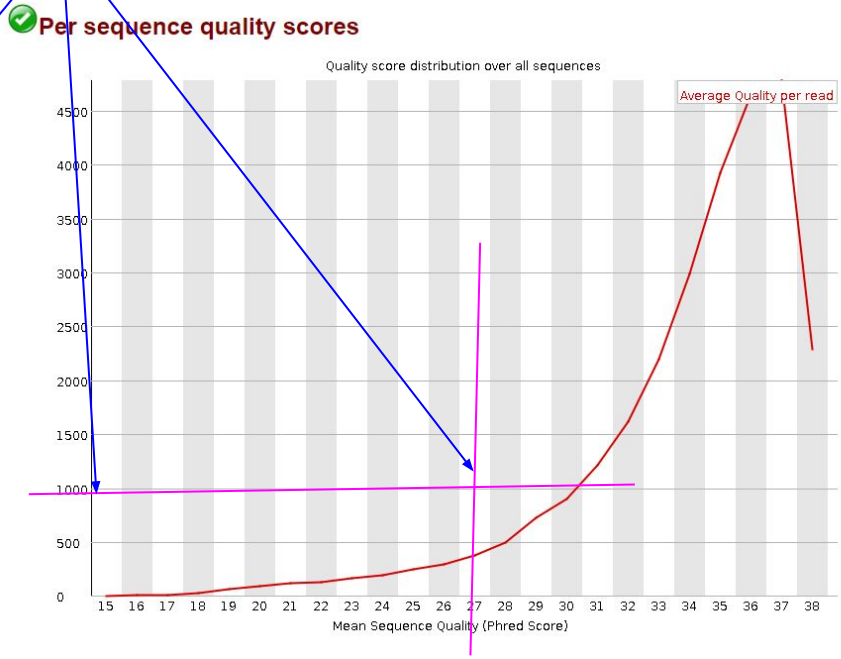


Per tile sequence quality

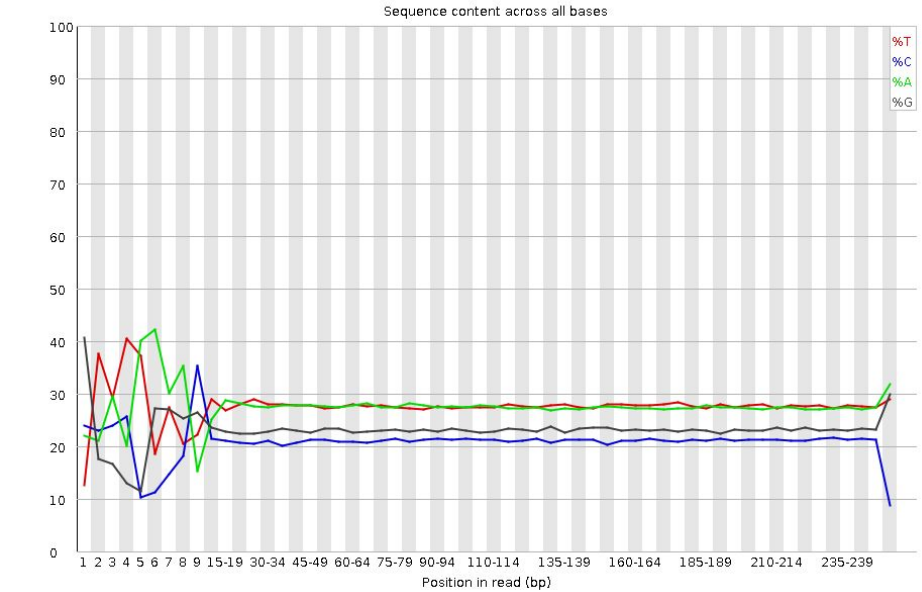


it shows the qualities of the subsets. if a significant portion of the seq. shows a higher score, it indicates a sistematic error. In these graphs, the observed mean quality is not below 27, (%0.2 erro rate), but it is close to the 27. Therefore even there are errors for both sequences, the errors are not so high. The more error can be seen in the data #1

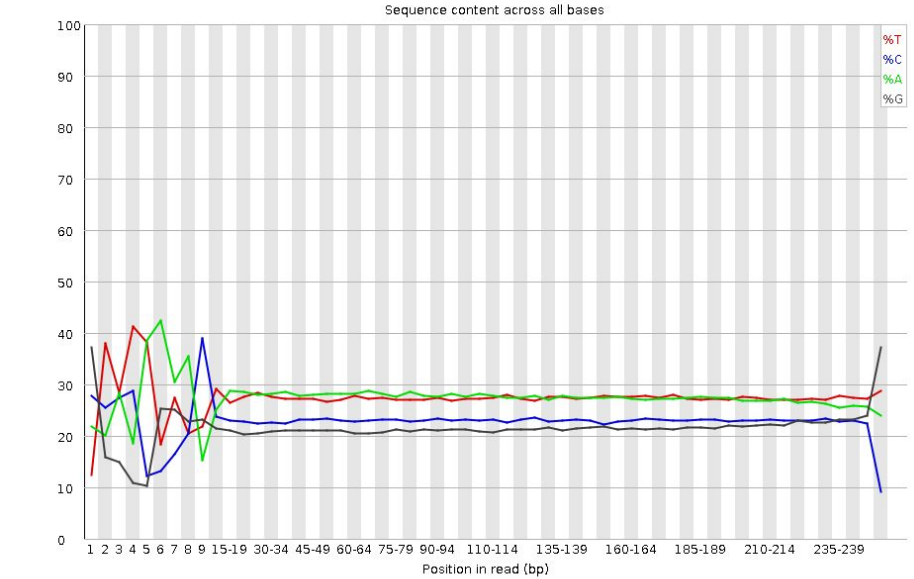




Per base sequence content

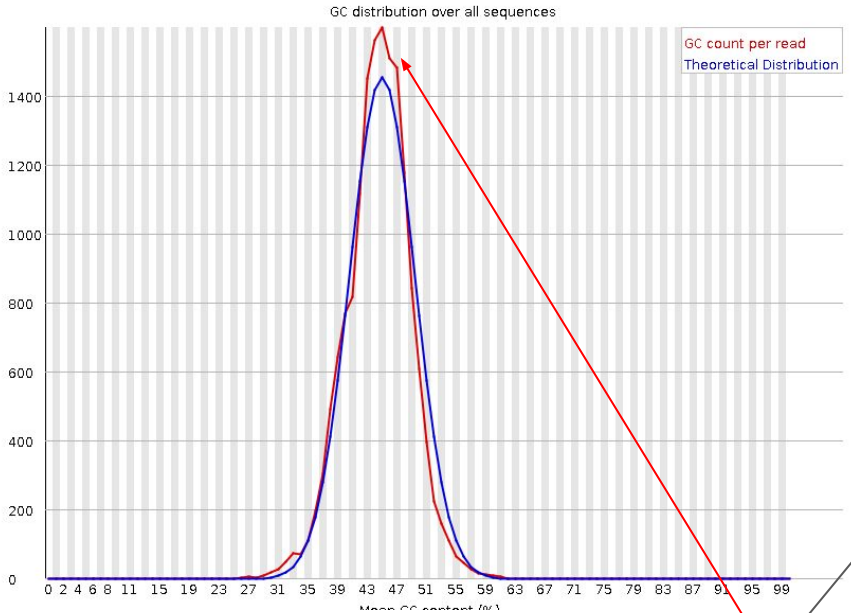


Per base sequence content

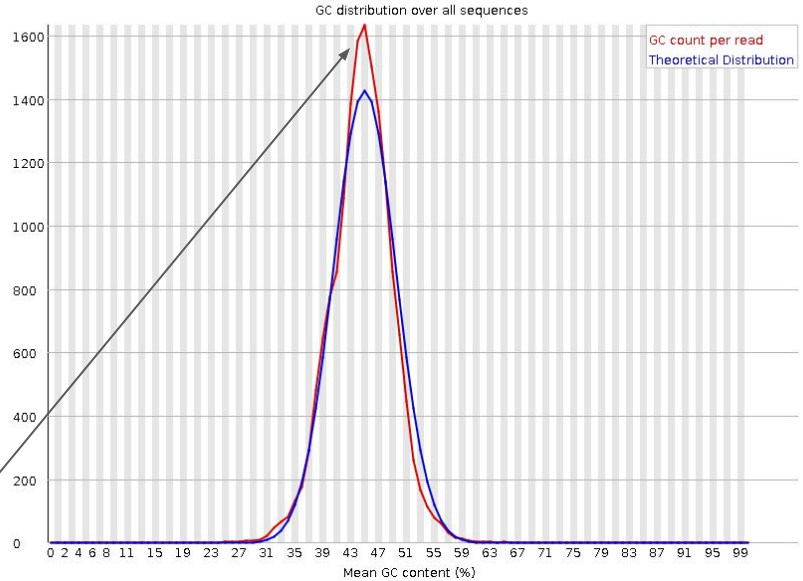


the results are similar for per base sequence content for two data sets

Per sequence GC content

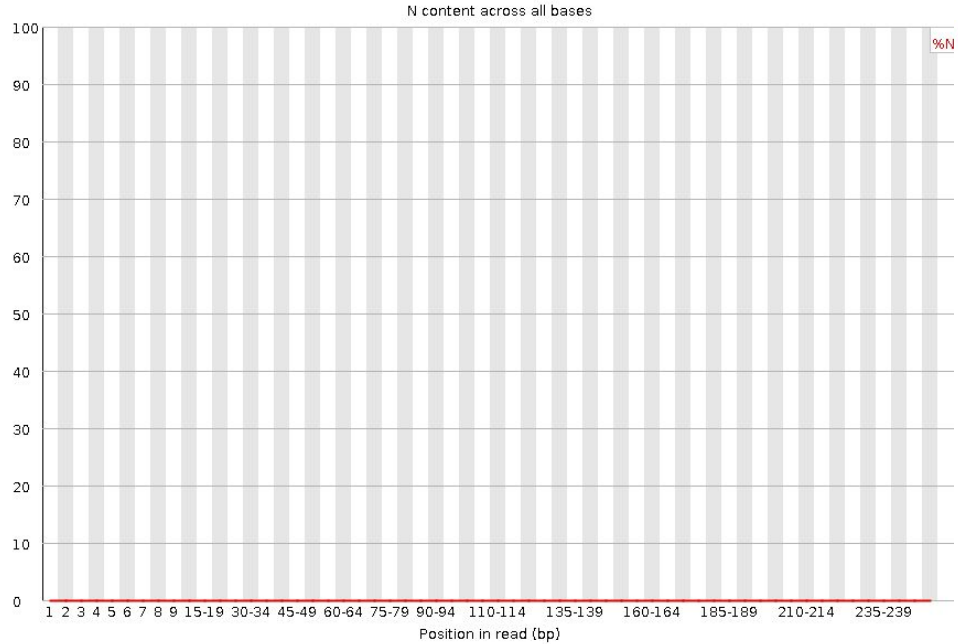


Per sequence GC content

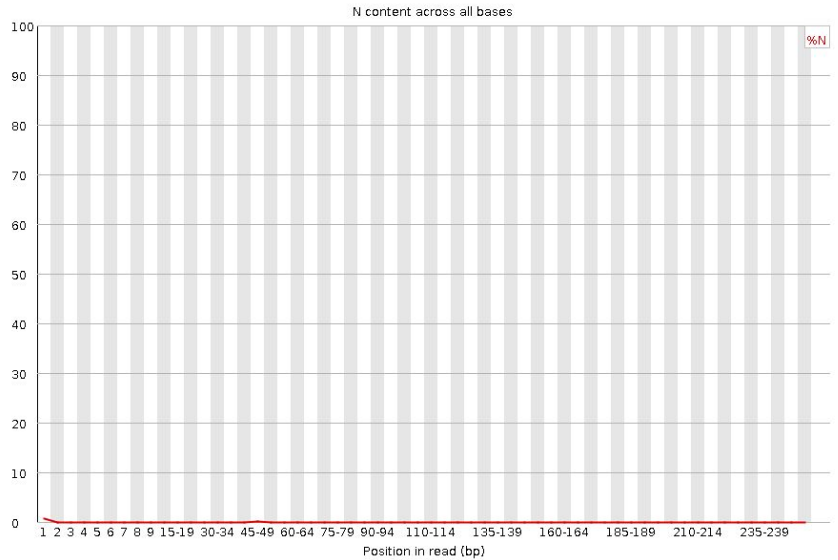


the results are similar for GC content for two data sets with a small difference

Per base N content

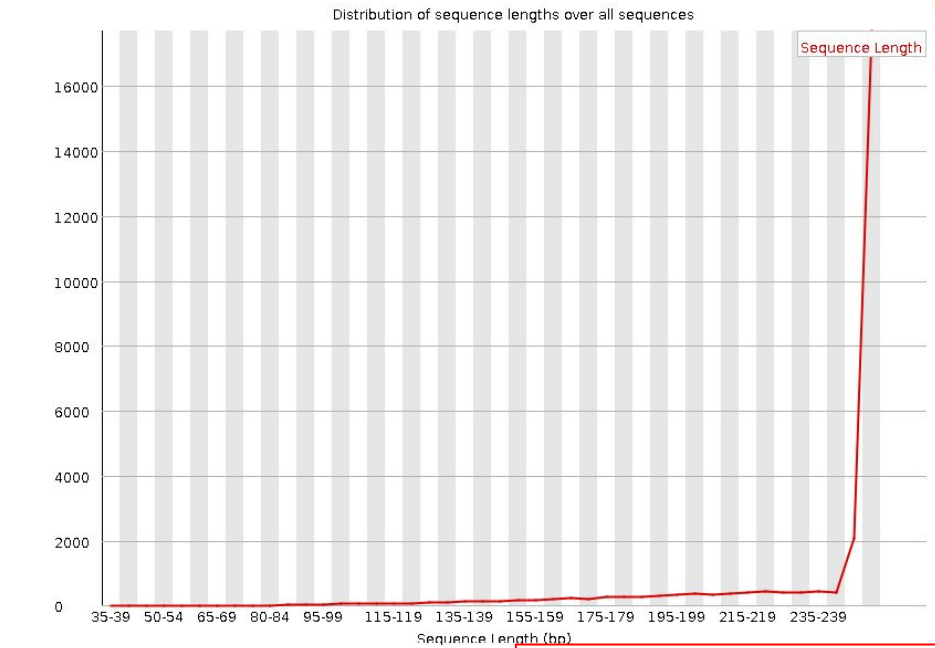


Per base N content

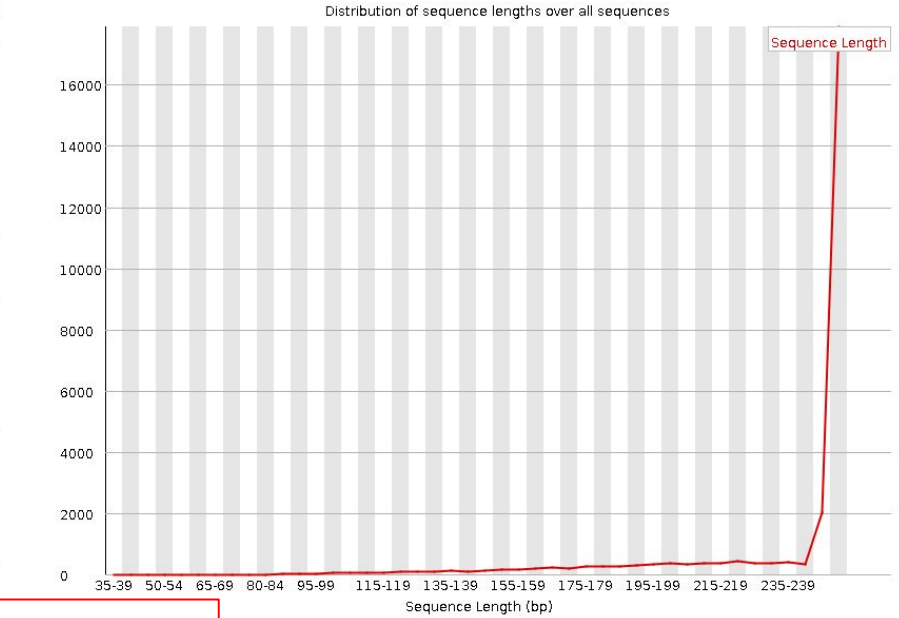


the results are similar for N content for two data sets

Sequence Length Distribution

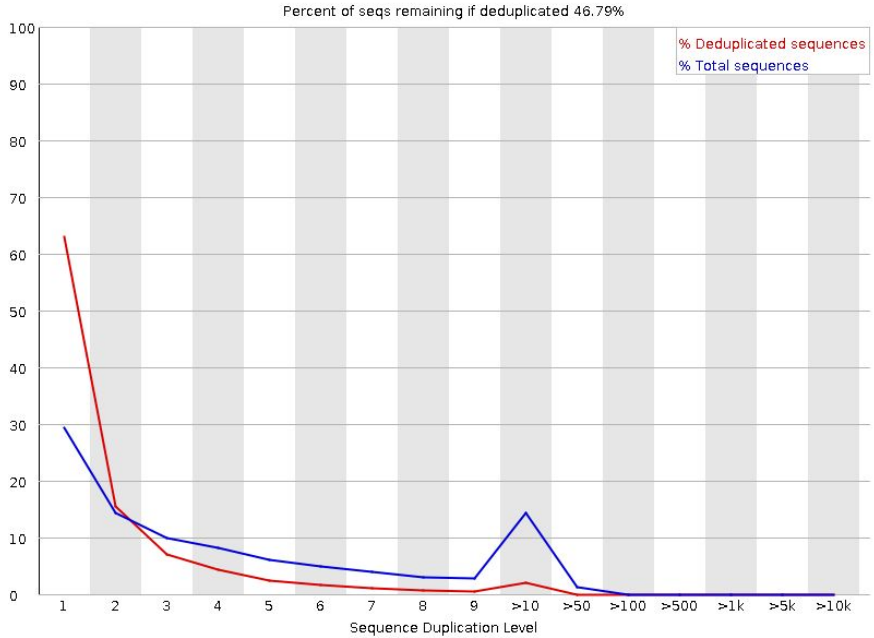


Sequence Length Distribution

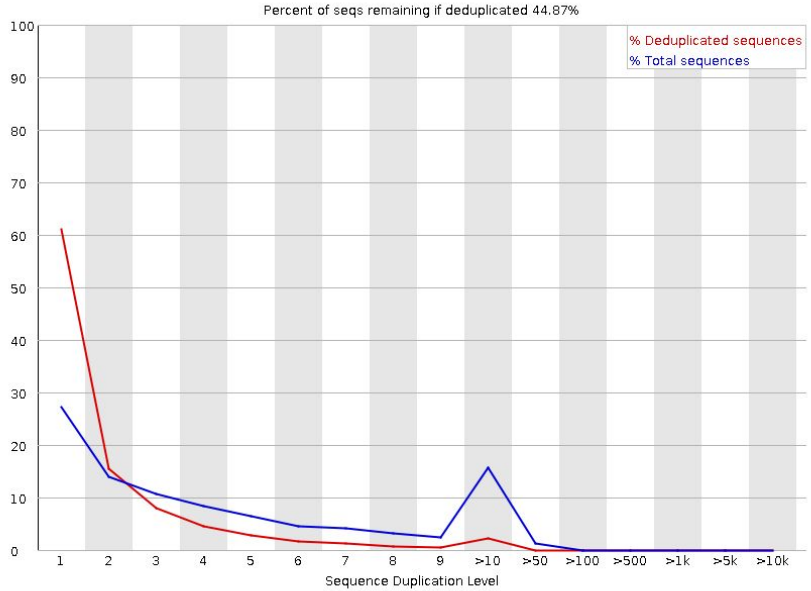


the results are similar for sequence length for two data sets

Sequence Duplication Levels



Sequence Duplication Levels



the results are similar for duplication levels for two data sets

the results are similar for overrepresented sequences for two data sets but in the data set # 2 there are more overrepresented sequences can be found

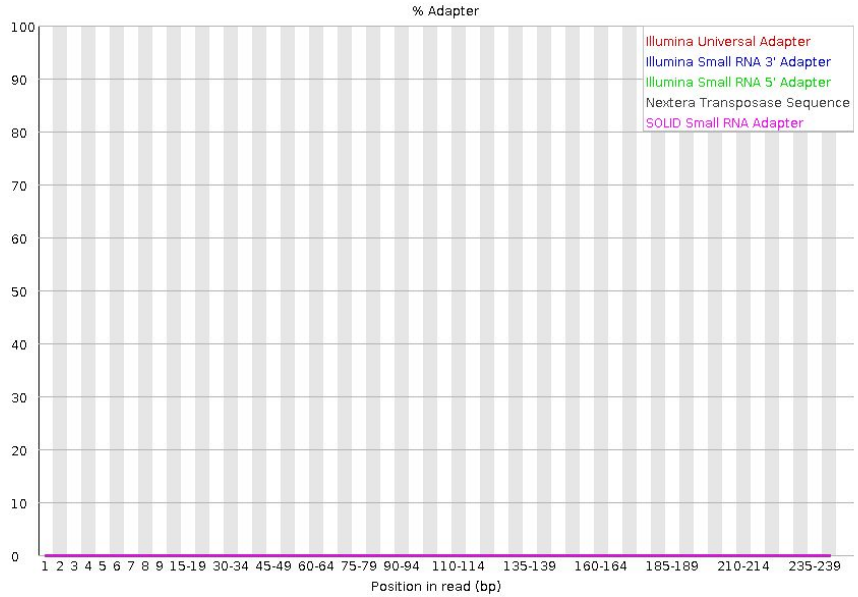
Overrepresented sequences

Sequence	Count	Percentage	Possible
GTACAAGGAAGGGGTAGGCTATGTGTTTTGTCAGGGGGTTGAGAATGAGT	93	0.33688328624212127	No Hit
GTATTATACCATAGCCGCCTAGTTTCAAGAGTACTGCGGCAAGTACTATT	65	0.23545606027675142	No Hit
TCCTAATACTAACTACCTGACTCCTACCCCTCACAATCATGGCAAGCCAA	62	0.22458885749474752	No Hit
GTGTAAGCTAGTCATATTAAGTTGTTGGCTCAGGAGTTTGATAGTTCTTG	59	0.21372165471274363	No Hit
GCATTAGGAATGCCATTGCGATTAGAATGGGTACAATGAGGAGTAGGAGG	59	0.21372165471274363	No Hit
GGATAGTACAAGGAAGGGGTAGGCTATGTGTTTTGTCAGGGGGTTGAGAA	53	0.19198724914873577	No Hit
ATACTAATCTCCCTACAAATCTCCTTAATTATAACATTCACAGCCACAGA	50	0.18112004636673187	No Hit
CATAGGGATAGTACAAGGAAGGGGTAGGCTATGTGTTTTGTCAGGGGGTT	39	0.14127363616605085	No Hit
TCTCTATACTAATCTCCCTACAAATCTCCTTAATTATAACATTCACAGCC	37	0.13402883431138157	No Hit
GGCTTTAGGGAGTCATAAGTGGAGTCCGTAAAGAGGTATCTTTACTATAA	34	0.12316163152937767	No Hit
GTACTATTGACCCAGCGATGGGGGCTTCGACATGGGCTTTAGGGAGTCAT	32	0.11591682967470841	No Hit
GTCATAAGTGGAGTCCGTAAAGAGGTATCTTTACTATAAAAGCTATTGTG	31	0.11229442874737376	No Hit
GTACATGCTAAGACTTCACCAGTCAAAGCGAACTACTATACTCAATTGAT	30	0.10867202782003911	No Hit
ATGCTAGGGTGAGTGGTAGGAAGTTTTTTCATAGGAGGTGTATGAGTTGG	28	0.10142722596536984	No Hit

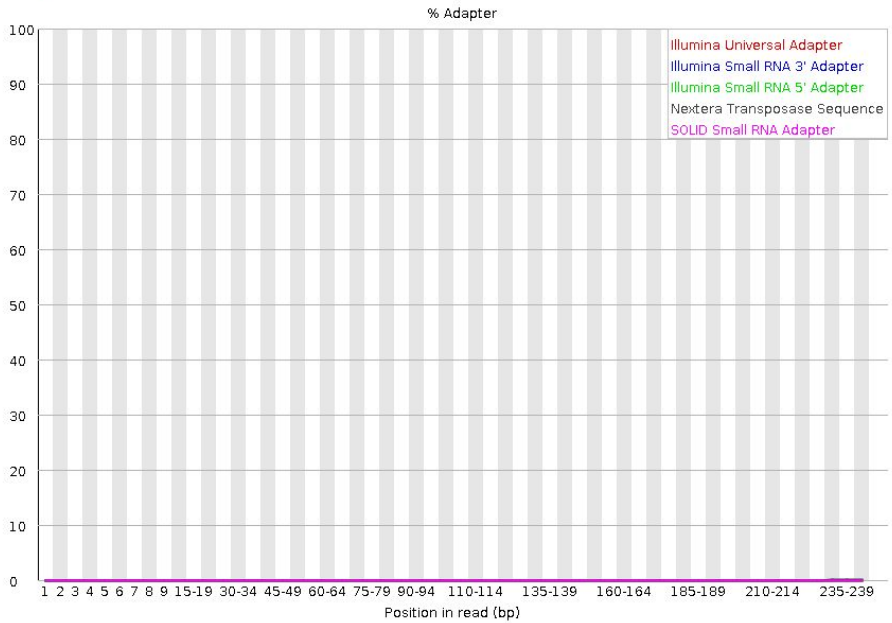
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ATACTAATCTCCCTACAAATCTCCTTAATTATAACATTCACAGCCACAGA	59	0.21372165471274363	No Hit
GCATTAGGAATGCCATTGCGATTAGAATGGGTACAATGAGGAGTAGGAGG	57	0.20647685285807432	No Hit
TCTCTATACTAATCTCCCTACAAATCTCCTTAATTATAACATTCACAGCC	55	0.19923205100340508	No Hit
CCCTAGTAGGCTCCCTTCCCCTACTCATCGCACTAATTTACACTCACAAC	55	0.19923205100340508	No Hit
GTACAAGGAAGGGGTAGGCTATGTGTTTTGTCAGGGGGTTGAGAATGAGT	52	0.18836484822140115	No Hit
GTACATGCTAAGACTTCACCAGTCAAAGCGAACTACTATACTCAATTGAT	52	0.18836484822140115	No Hit
TCCTAATACTAACTACCTGACTCCTACCCCTCACAATCATGGCAAGCCAA	52	0.18836484822140115	No Hit
GTATTATACCATAGCCGCCTAGTTTCAAGAGTACTGCGGCAAGTACTATT	42	0.15214083894805477	No Hit
TTCTACTACTCACTCTCACTGCCCAAGAACTATCAAACTCCTGAGCCAAC	38	0.13765123523871622	No Hit
CTCCTACCCCTCACAATCATGGCAAGCCAACGCCACTTATCCAGTGAACC	38	0.13765123523871622	No Hit
TCCTTGTACTATCCCTATGAGGCATAATTATAACAAGCTCCATCTGCCTA	36	0.13040643338404695	No Hit
CTCTATACTAATCTCCCTACAAATCTCCTTAATTATAACATTCACAGCCA	35	0.1267840324567123	No Hit
AATTTACACTCACAACACCCTAGGCTCACTAAACATTCTACTACTCACTC	33	0.11953923060204304	No Hit
TGCTAAGACTTCACCAGTCAAAGCGAACTACTATACTCAATTGATCCAAT	31	0.11229442874737376	No Hit
ACCCTAGGCTCACTAAACATTCTACTACTCACTCTCACTGCCCAAGAACT	30	0.10867202782003911	No Hit
CTACTATACTCAATTGATCCAATAACTTGACCAACGGAACAAGTTACCCT	29	0.10504962689270449	No Hit
CCACAGAACTAATCATATTTTATATCTTCTTCGAAACCACACTTATCCCC	29	0.10504962689270449	No Hit
CCCTAGCATTACTTATATGATATGTCTCCATACCCATTACAATCTCCAGC	28	0.10142722596536984	No Hit

Adapter Content

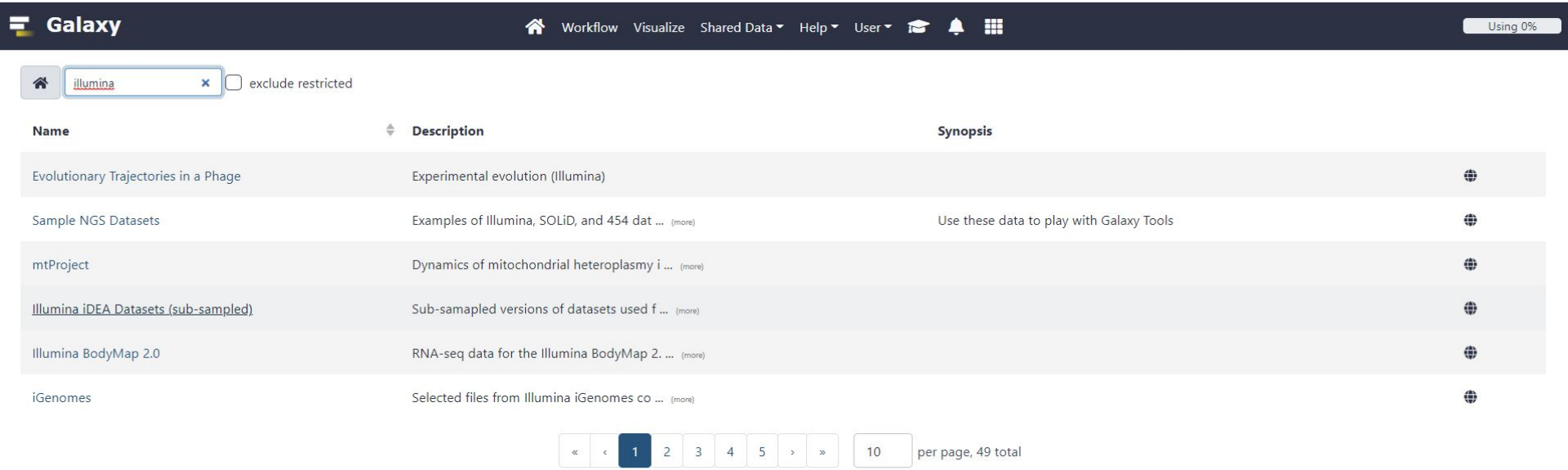


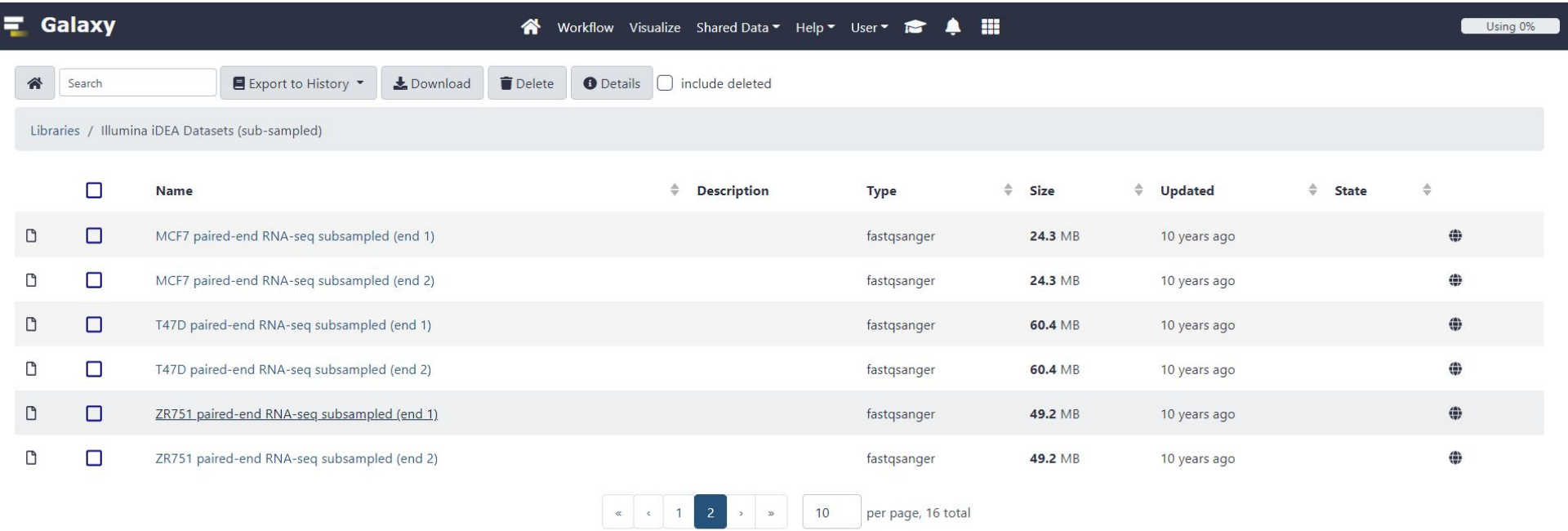
Adapter Content



the results are similar for adapter content for two data sets

Part 2: getting the data



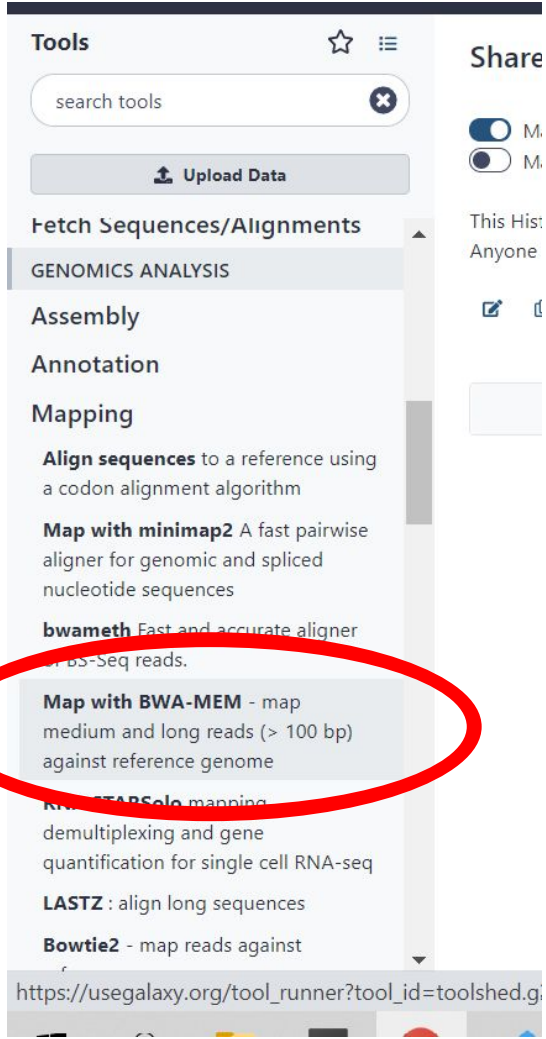

Galaxy
Workflow
Visualize
Shared Data
Help
User
Using 0%
Search
Export to History
Download
Delete
Details
include deleted
Libraries / Illumina iDEA Datasets (sub-sampled)
Name
Description
Type
Size
Updated
State
MCF7 paired-end RNA-seq subsampled (end 1)
fastqsanger
24.3 MB
10 years ago
MCF7 paired-end RNA-seq subsampled (end 2)
fastqsanger
24.3 MB
10 years ago
T47D paired-end RNA-seq subsampled (end 1)
fastqsanger
60.4 MB
10 years ago
T47D paired-end RNA-seq subsampled (end 2)
fastqsanger
60.4 MB
10 years ago
ZR751 paired-end RNA-seq subsampled (end 1)
fastqsanger
49.2 MB
10 years ago
ZR751 paired-end RNA-seq subsampled (end 2)
fastqsanger
49.2 MB
10 years ago
« ‹ 1 2 › »
10
per page, 16 total

Libraries / Illumina iDEA Datasets (sub-sampled)

	Name	Description	Type	Size	Updated	State
	MCF7 paired-end RNA-seq subsampled (end 1)		fastqsanger	**24.3** MB	10 years ago	
	MCF7 paired-end RNA-seq subsampled (end 2)		fastqsanger	**24.3** MB	10 years ago	
	T47D paired-end RNA-seq subsampled (end 1)		fastqsanger	**60.4** MB	10 years ago	
	T47D paired-end RNA-seq subsampled (end 2)		fastqsanger	**60.4** MB	10 years ago	
	ZR751 paired-end RNA-seq subsampled (end 1)		fastqsanger	**49.2** MB	10 years ago	
	ZR751 paired-end RNA-seq subsampled (end 2)		fastqsanger	**49.2** MB	10 years ago	

« ‹ 1 2 › » 10 per page, 16 total

BWA-MEM algorithm=for high-quality queries(faster && more accurate)
History information: https://usegalaxy.org/u/leman_nur_nehri/h/part2assisgnment3



Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.2)

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

Human (Homo sapiens) (b38): hg38

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

2: ZR751 paired-end RNA-seq subsampled (end 1)

Specify dataset with forward reads

Select second set of reads

1: ZR751 paired-end RNA-seq subsampled (end 2)

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

while "250,10" will not. See below for details.

Set read groups information?

Set read groups (SAM/BAM specification)

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value

Platform/technology used to produce the reads (PL)

ILLUMINA

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value

Executed **Map with BWA-MEM** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- **2: ZR751 paired-end RNA-seq subsampled (end 1)**
- **1: ZR751 paired-end RNA-seq subsampled (end 2)**

It produces this output:

- **3: Map with BWA-MEM on data 1 and data 2 (mapped reads in BAM format)**

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

Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

- **2: ZR751 paired-end RNA-seq subsampled (end 1)**

It produces 2 outputs:

- **4: FastQC on data 2: Webpage**
- **5: FastQC on data 2: RawData**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change

Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

- **1: ZR751 paired-end RNA-seq subsampled (end 2)**

It produces 2 outputs:

- **6: FastQC on data 1: Webpage**
- **7: FastQC on data 1: RawData**

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

MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.9+galaxy1)

Results

1: Results

Which tool was used generate logs?

FastQC

Software name

FastQC output

1: FastQC output

Type of FastQC output?

Raw data

FastQC output

7: FastQC on data 1: RawData
6: FastQC on data 1: Webpage
5: FastQC on data 2: RawData
4: FastQC on data 2: Webpage
2: ZR751 paired-end RNA-seq subsampled (end 1)
1: ZR751 paired-end RNA-seq subsampled (end 2)

+ Insert FastQC output

+ Insert Results

Executed **MultiQC** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- **5: FastQC on data 2: RawData**
- **7: FastQC on data 1: RawData**

It produces this output:

- **14: MultiQC on data 7 and data 5: Webpage**

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

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.2)

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

Human (Homo sapiens) (b38): hg38

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

2: ZR751 paired-end RNA-seq subsampled (end 1)

Specify dataset with forward reads

Select second set of reads

1: ZR751 paired-end RNA-seq subsampled (end 2)

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Executed **Map with BWA-MEM** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- **2: ZR751 paired-end RNA-seq subsampled (end 1)**
- **1: ZR751 paired-end RNA-seq subsampled (end 2)**

It produces this output:

- **10: Map with BWA-MEM on data 1 and data 2 (mapped reads in BAM format)**

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

results: multiqc

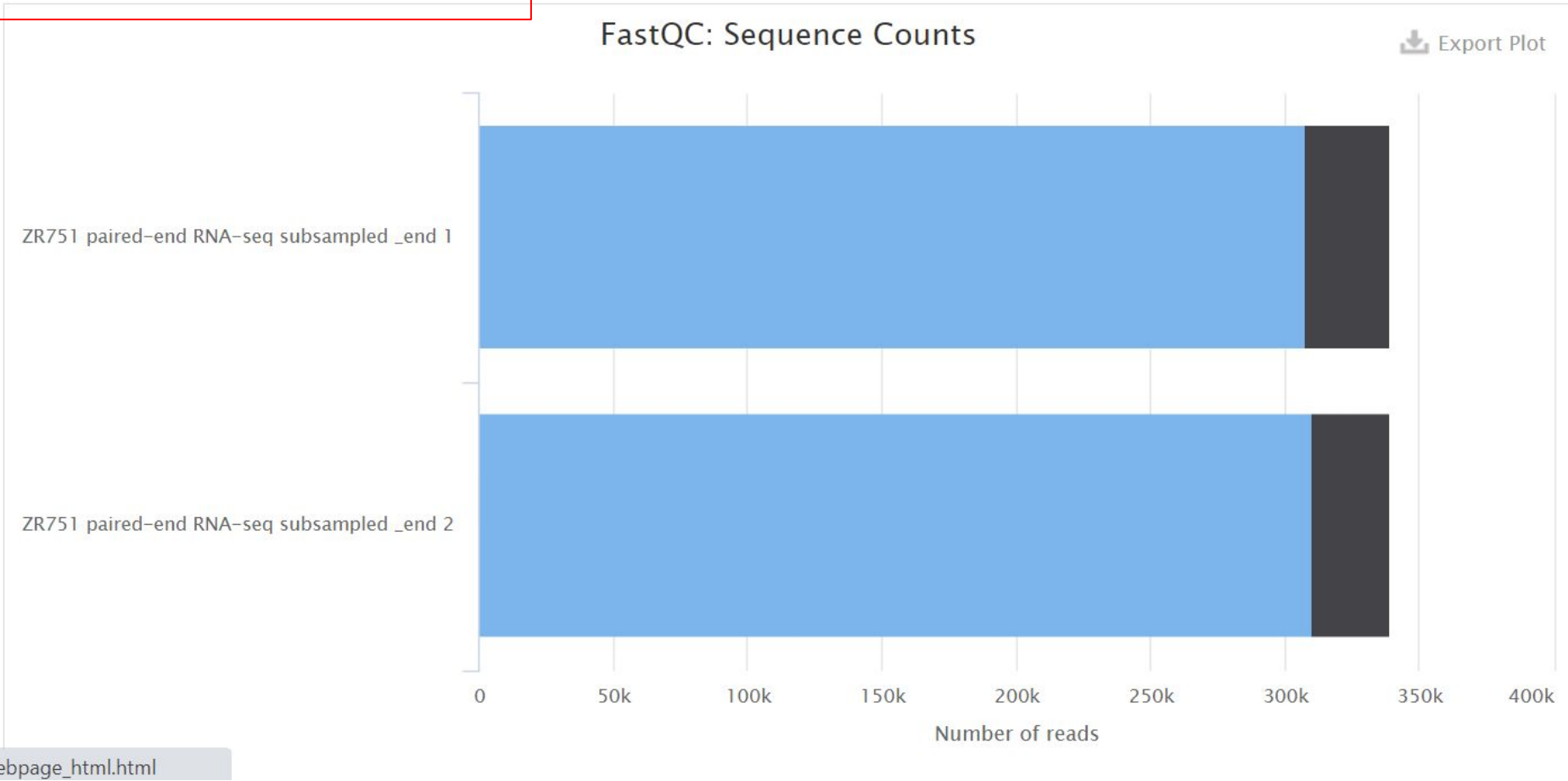
nearly same sequences, only duplication levels are different

General Statistics

Copy table | Configure Columns | Plot | Showing 2/2 rows and 5/5 columns.

Sample Name	% Dups	% GC	Length	% Failed	M Seqs
ZR751 paired-end RNA-seq subsampled _end 1	9.3%	50%	50 bp	0%	0.3
ZR751 paired-end RNA-seq subsampled _end 2	8.5%	50%	50 bp	0%	0.3

same for sequence counts



Webpage_html.html

Sequence Quality Histograms

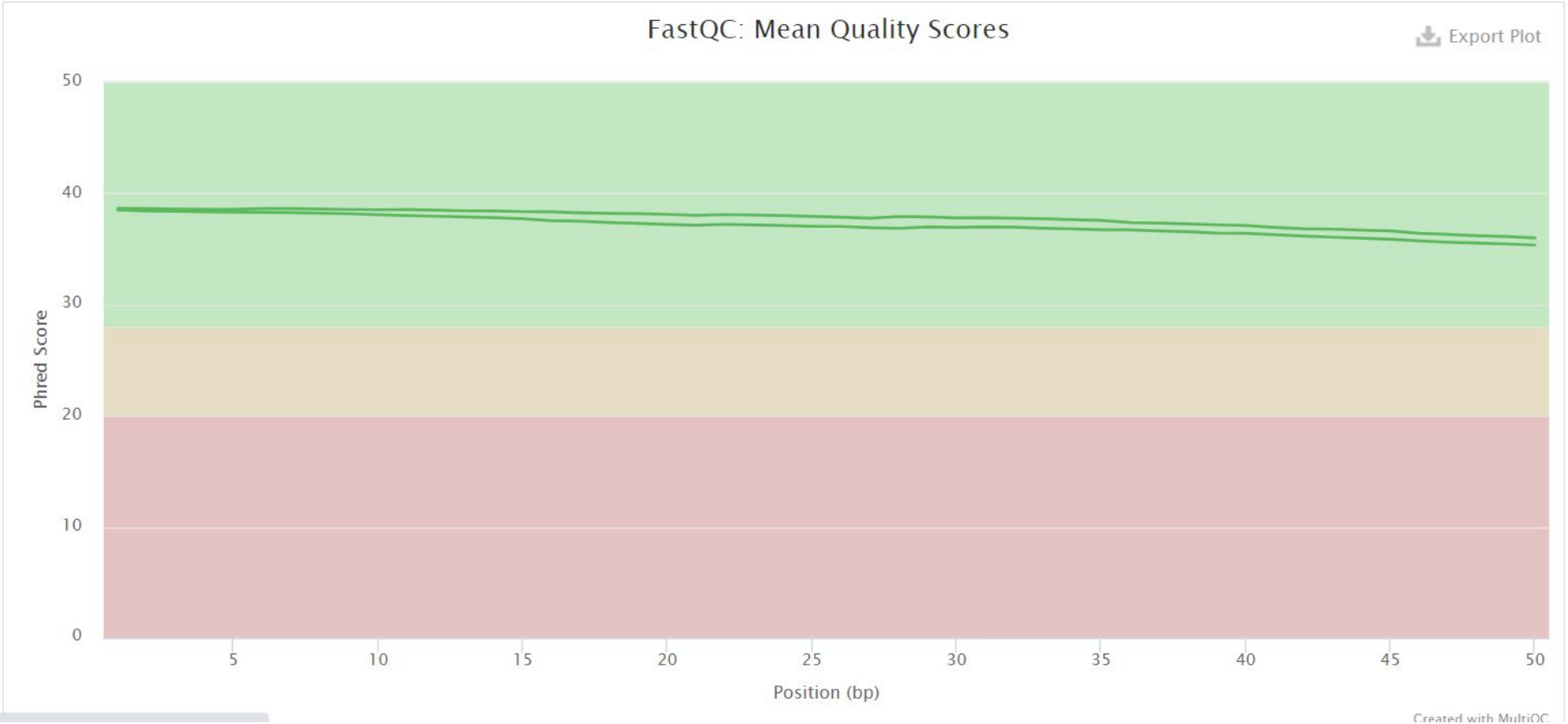
2

The mean quality value across each base position in the read.

nearly same for mean quality scores

Help

Y-Limits: on



Per Sequence Quality Scores

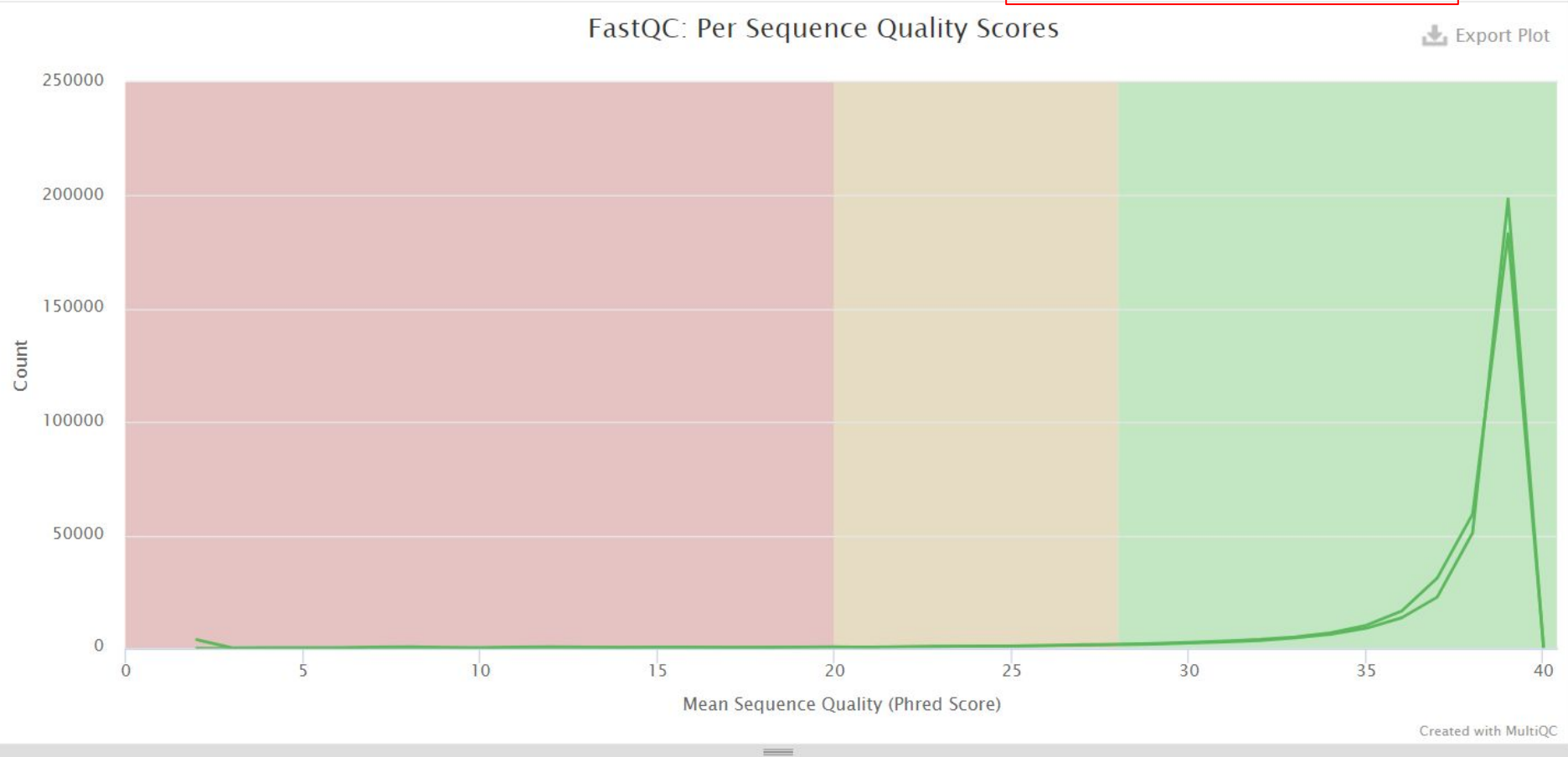
2

The number of reads with average quality scores. Shows if a subset of reads has poor quality.

nearly same for per sequence quality score

Help

Y-Limits: on



nearly same for per base seq. content

Per Base Sequence Content

2

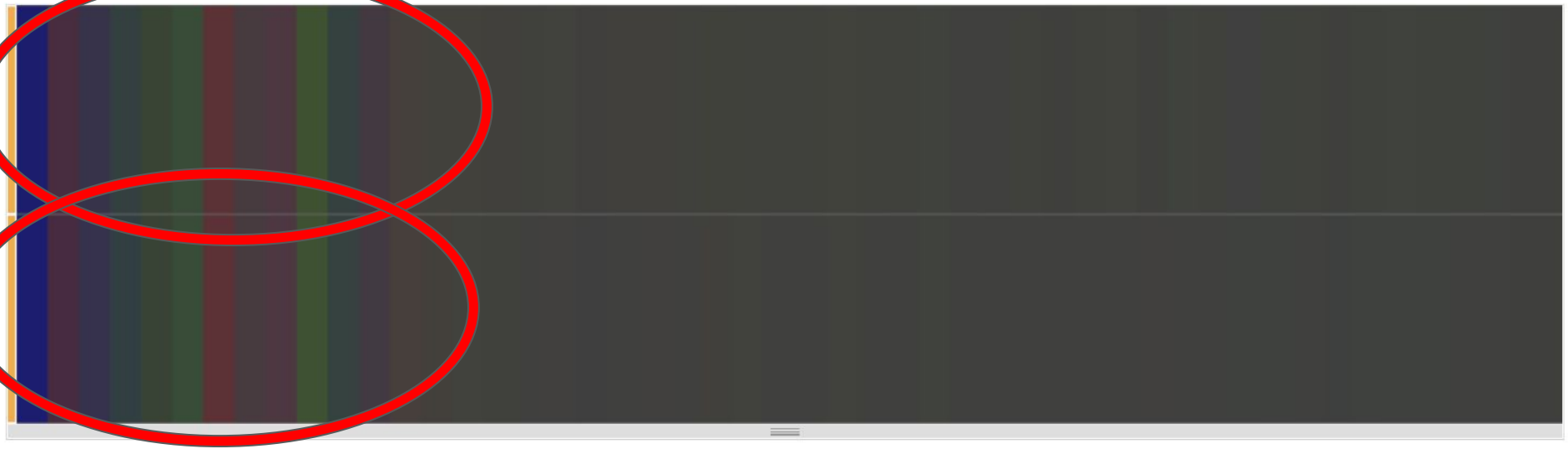
Help

The proportion of each base position for which each of the four normal DNA bases has been called.

Click a sample row to see a line plot for that dataset.

Rollover for sample name

Position: - %T: - %C: - %A: - %G: - Export Plot



nearly same for per sequence GC content

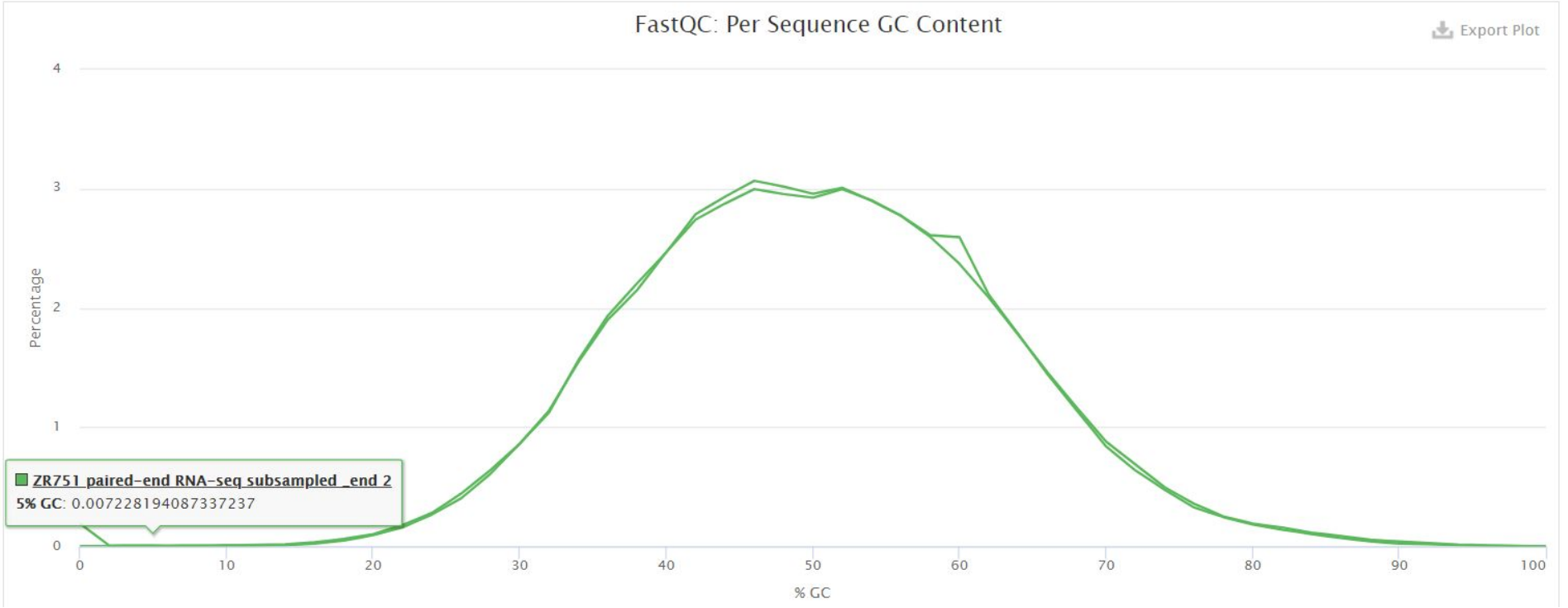
Per Sequence GC Content 2

Help

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Y Limits: on

Percentages | Counts



closer per base N contents but data#2 has more N content than data #1

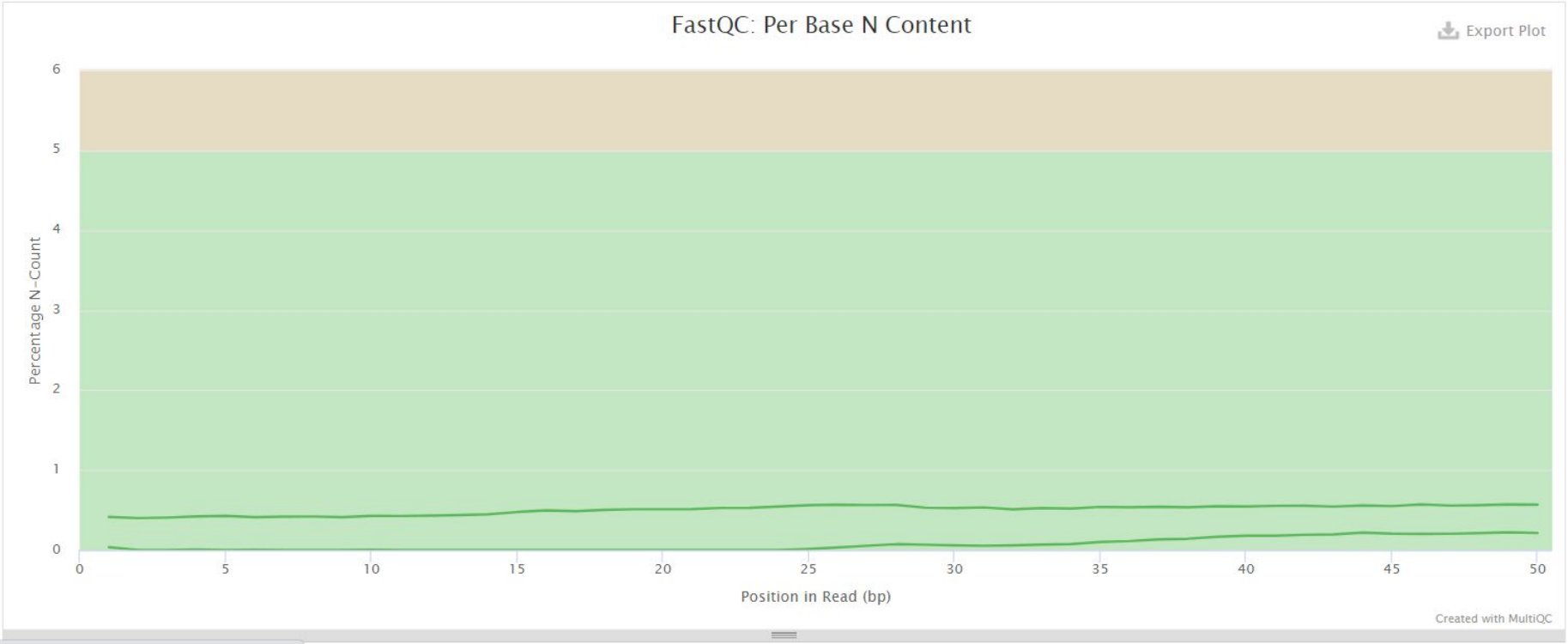
Per Base N Content

2

The percentage of base calls at each position for which an `N` was called.

Help

Y-Limits on



a question mark, how can be possible?

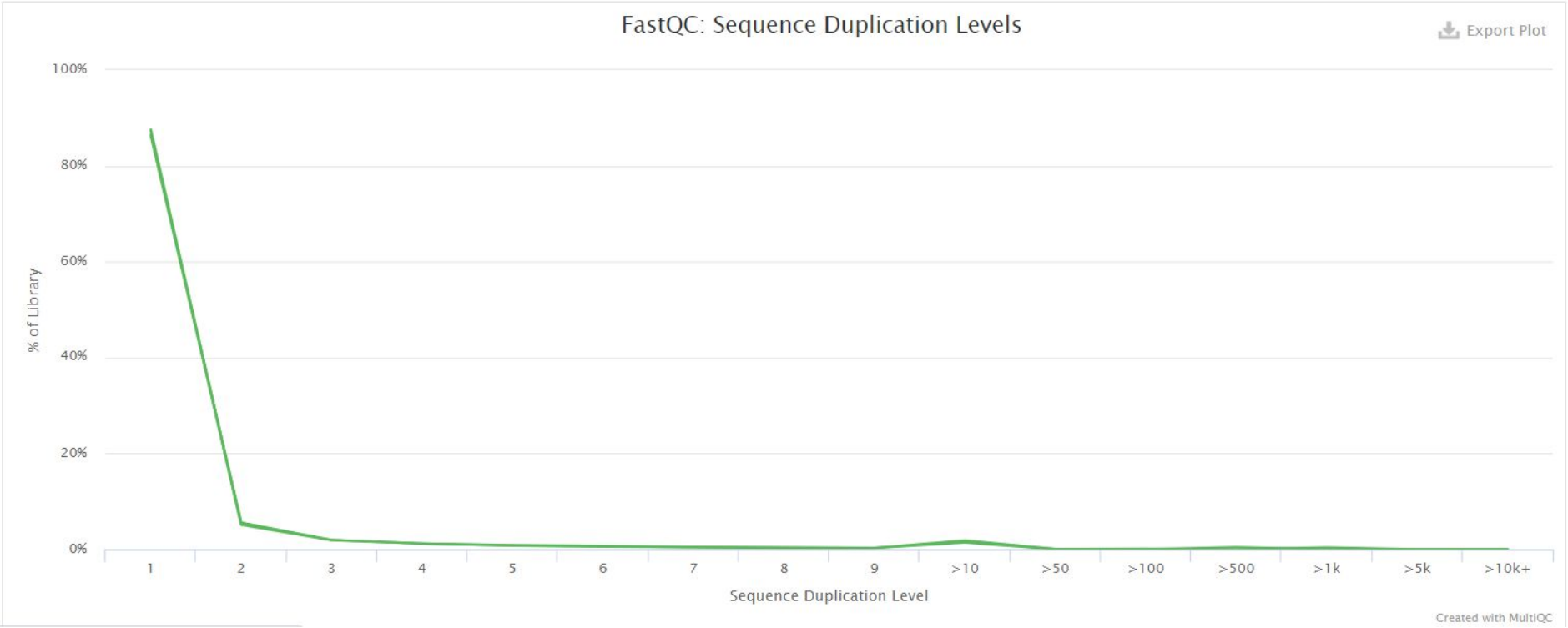
same ratios for sequence duplication levels

Sequence Duplication Levels 2

Help

The relative level of duplication found for every sequence.

Y-Limits: on



already all seq.s are known

Overrepresented sequences 2

The total amount of overrepresented sequences found in each library.

2 samples had less than 1% of reads made up of overrepresented sequences

data#1 is a bit higher than data #2 for adapter content, but they are very similar.

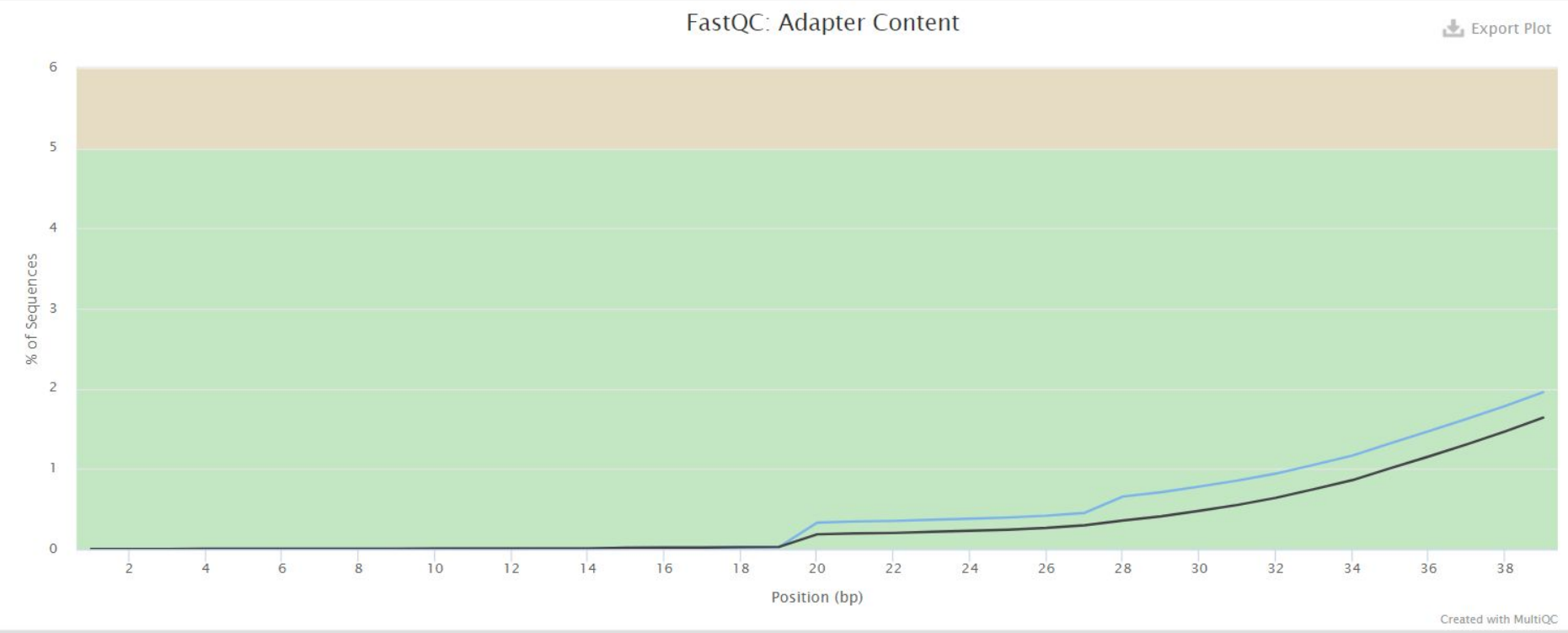
Adapter Content

2

Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Y-Limits: on



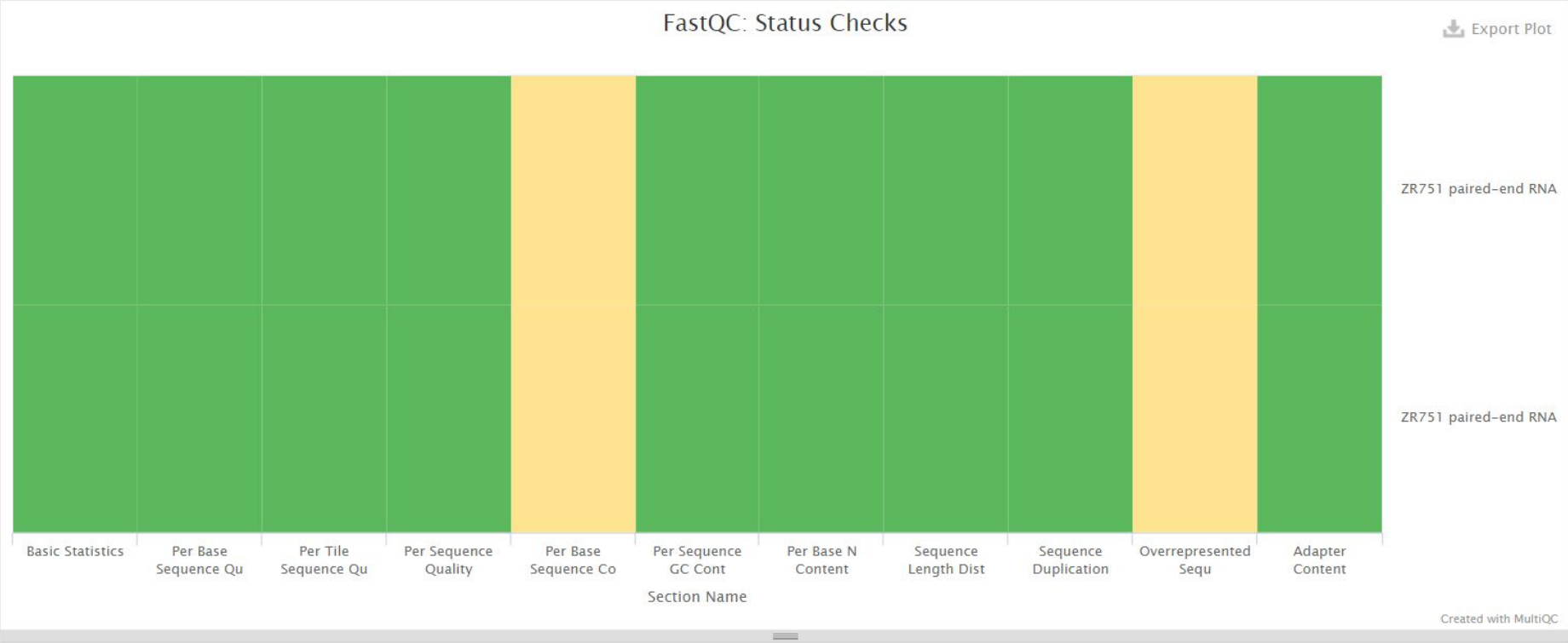
similar results for all the statistics, no difference

Status Checks

Help

Status for each FastQC section showing whether results seem entirely normal (green), slightly abnormal (orange) or very unusual (red).

Sort by highlight



left and right for : fastqc results

Basic Statistics

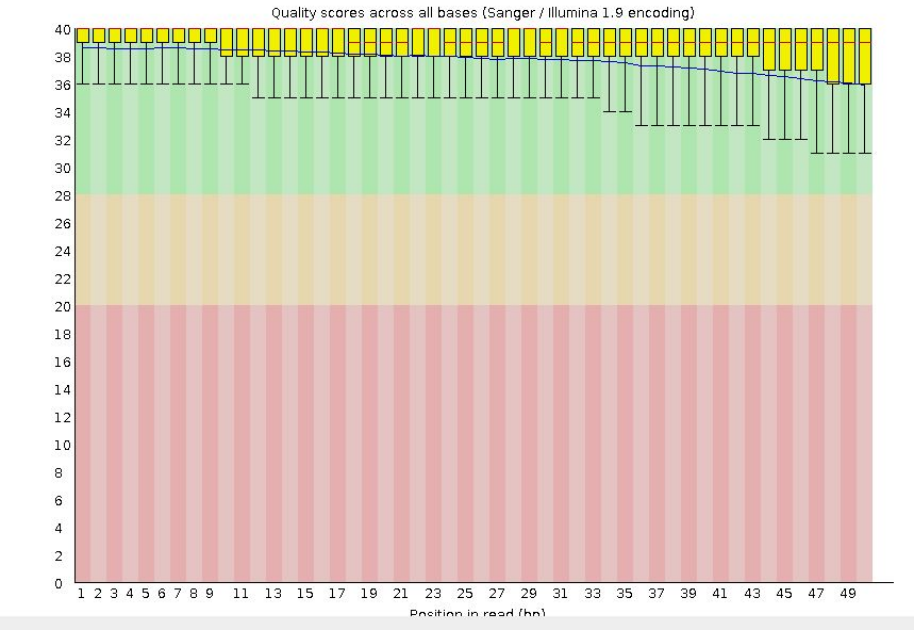
Measure	Value
Filename	ZR751 paired-end RNA-seq subsampled _end 1_
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	339276
Sequences flagged as poor quality	0
Sequence length	50
%GC	50

Basic Statistics

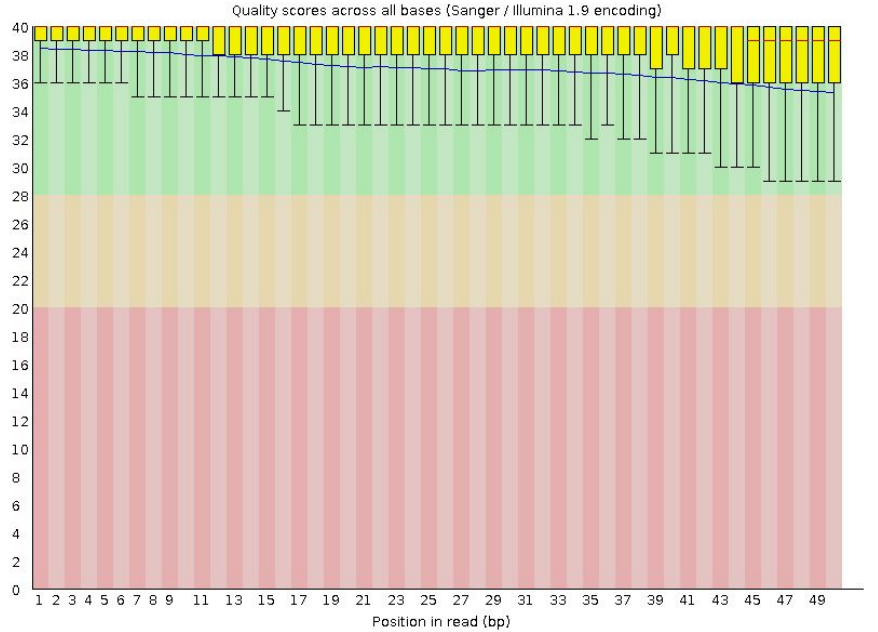
Measure	Value
Filename	ZR751 paired-end RNA-seq subsampled _end 2_
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	339276
Sequences flagged as poor quality	0
Sequence length	50
%GC	50

two seq.s are similar

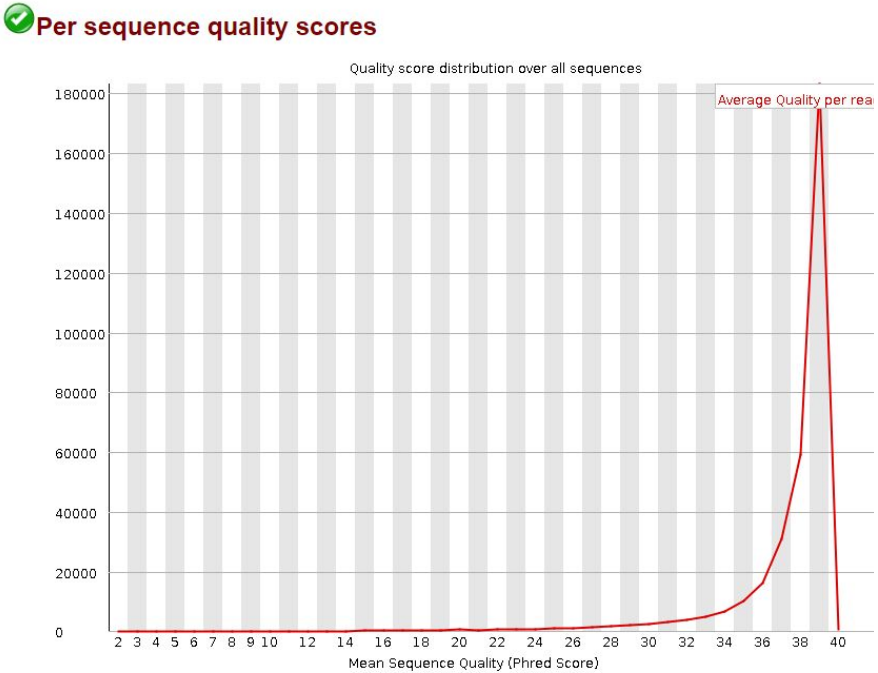
Per base sequence quality

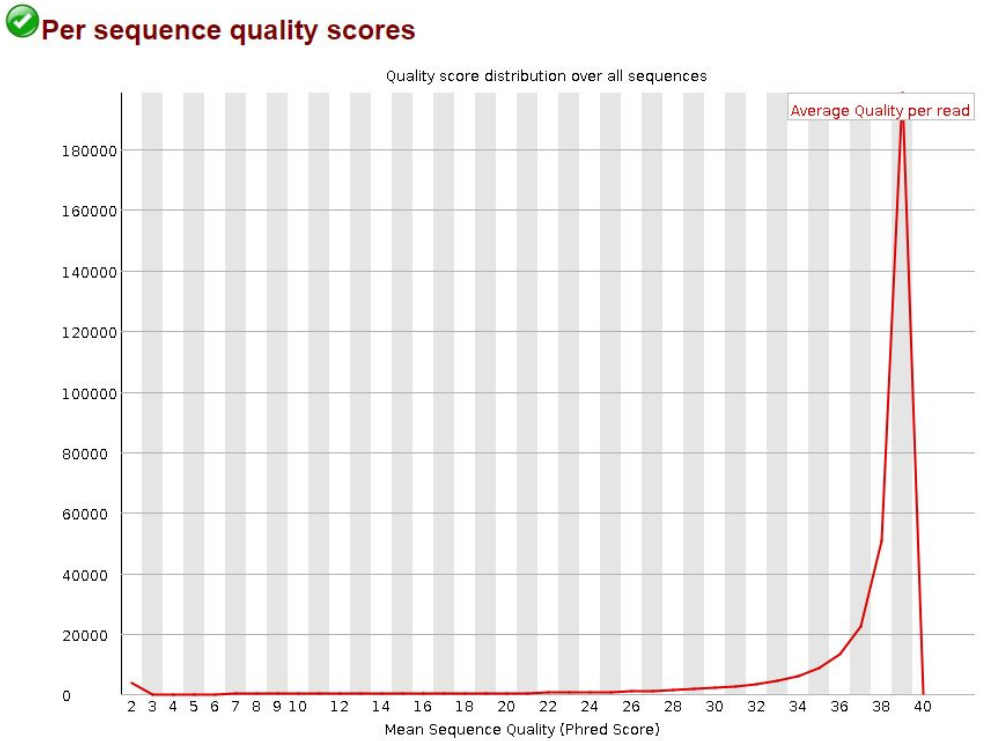


Per base sequence quality



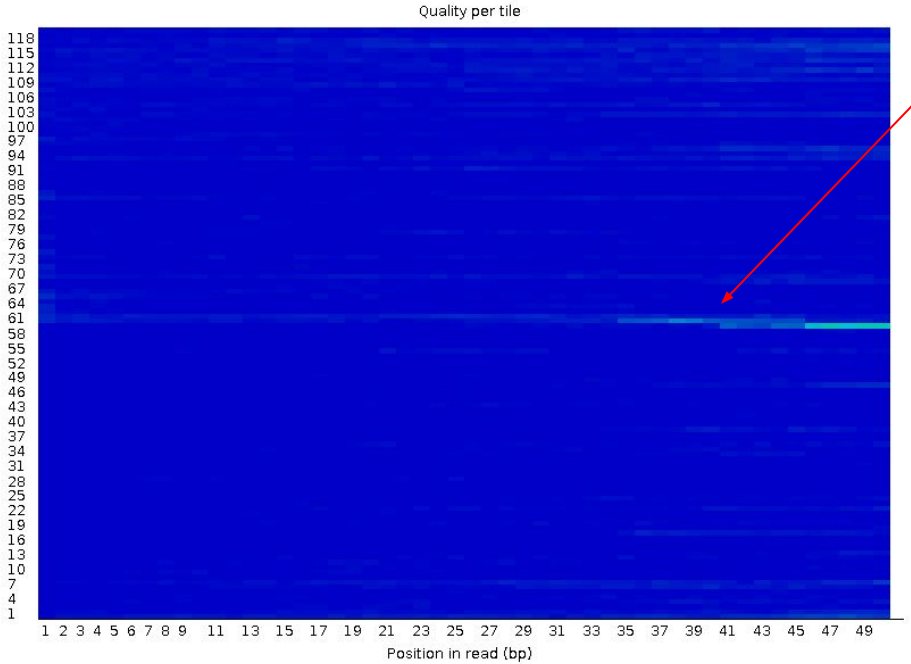
it shows the qualities of the subsets. if a significant portion of the seq. shows a higher score, it indicates a systematic error. in these graphs, the observed mean quality is below 27, (%0.2 error rate), therefore even there are errors for both sequences, the errors are small, can be ignored.



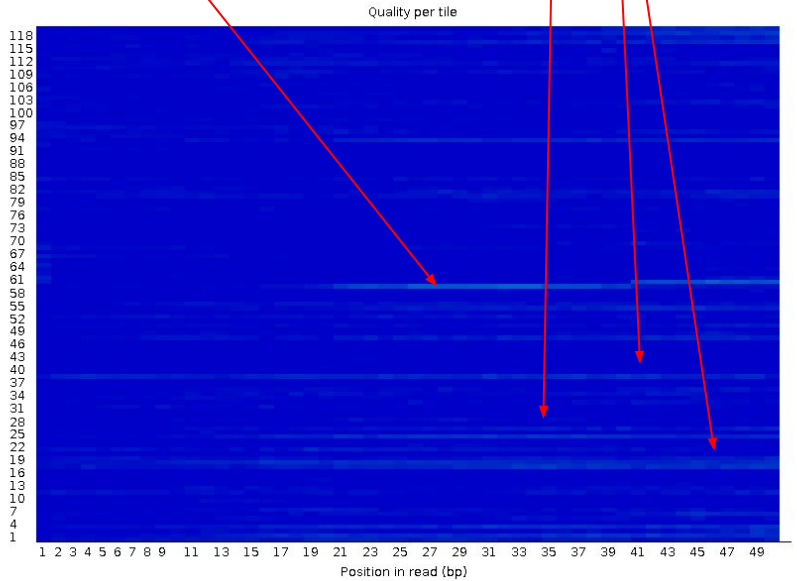


two seq.s are similar, but data#1 has a bit better quality for per tile

Per tile sequence quality

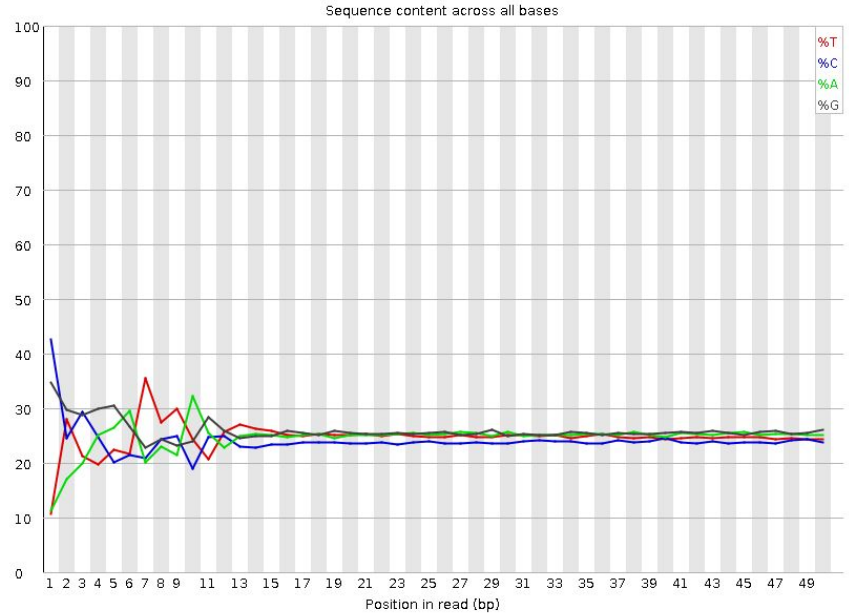


Per tile sequence quality

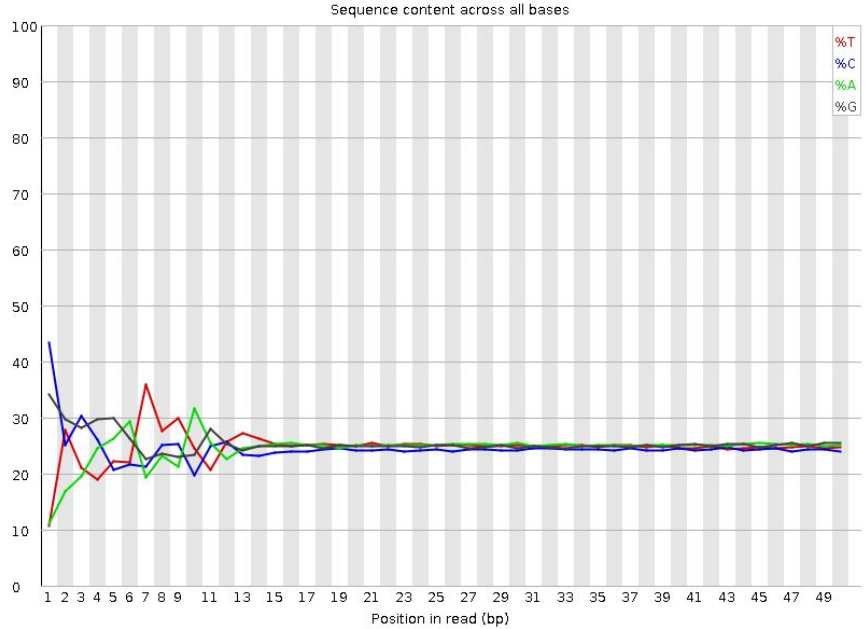


two seq.s are similar

Per base sequence content

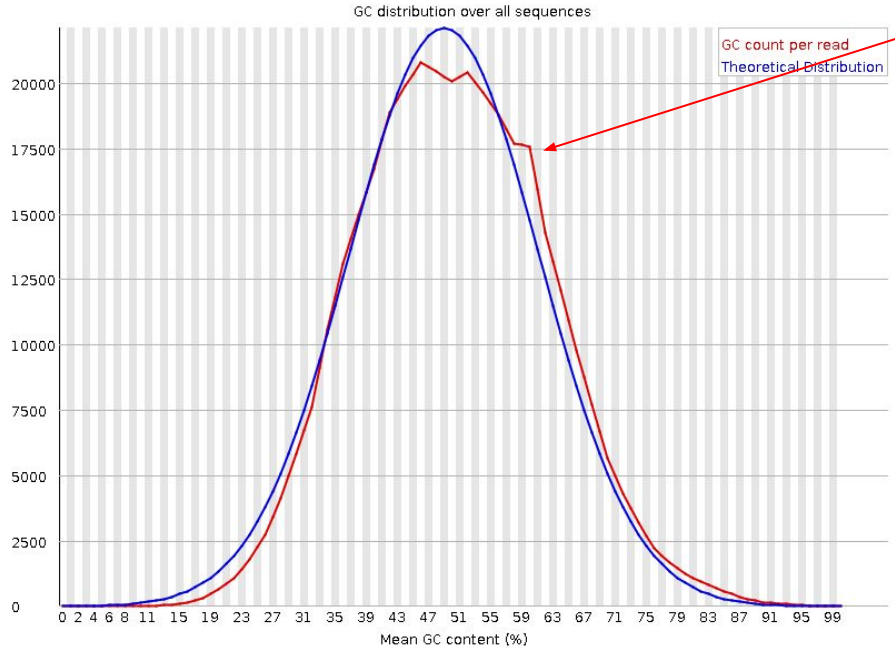


Per base sequence content

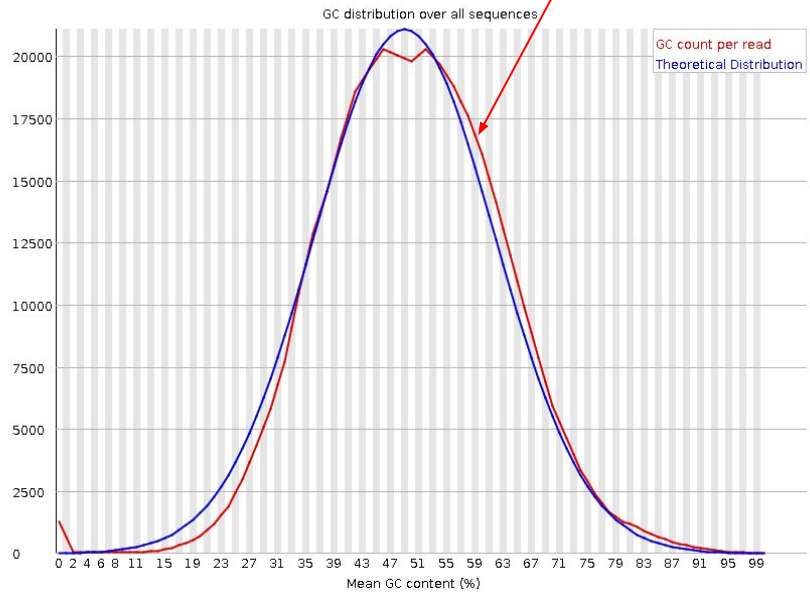


two seq.s are similar with small differences

Per sequence GC content

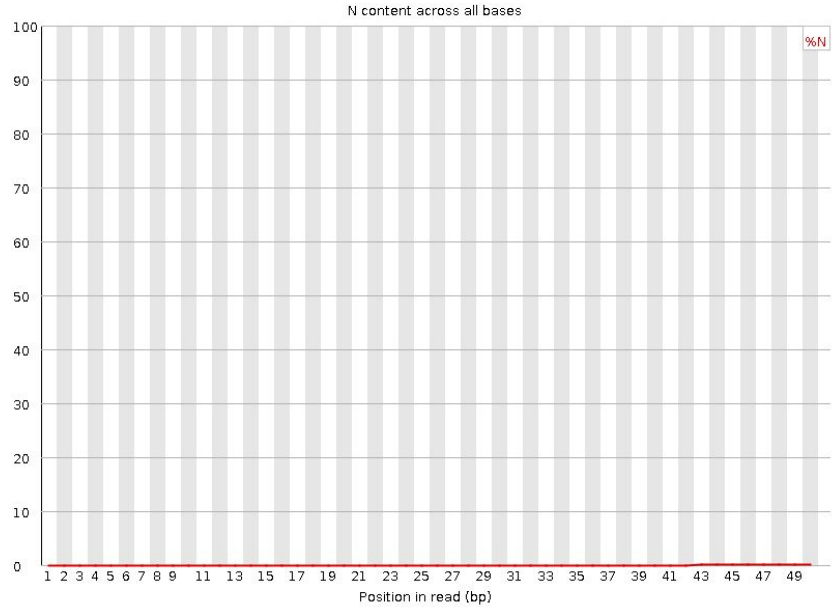


Per sequence GC content

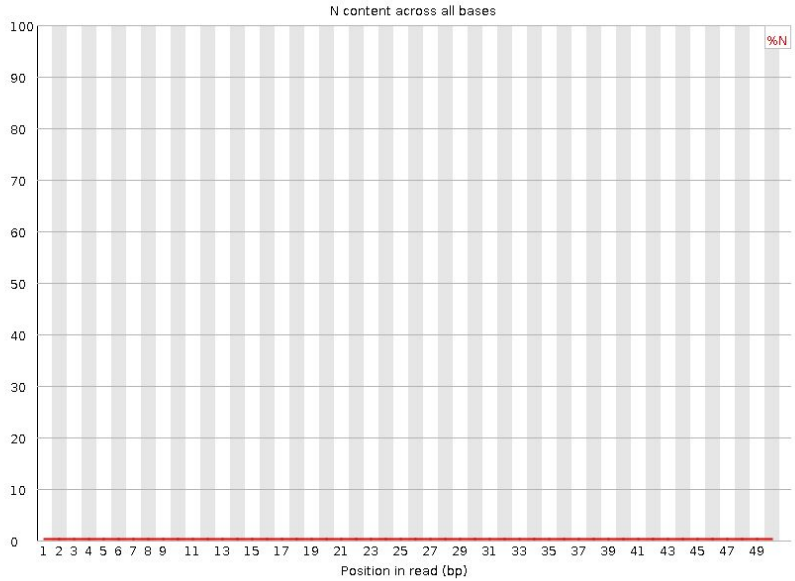


two seq.s are similar

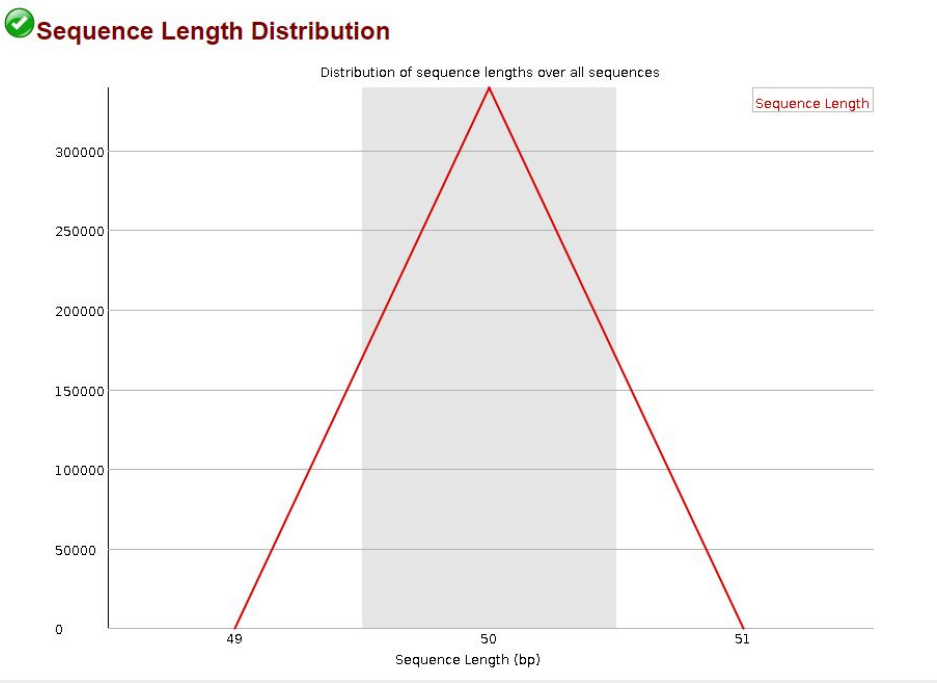
Per base N content

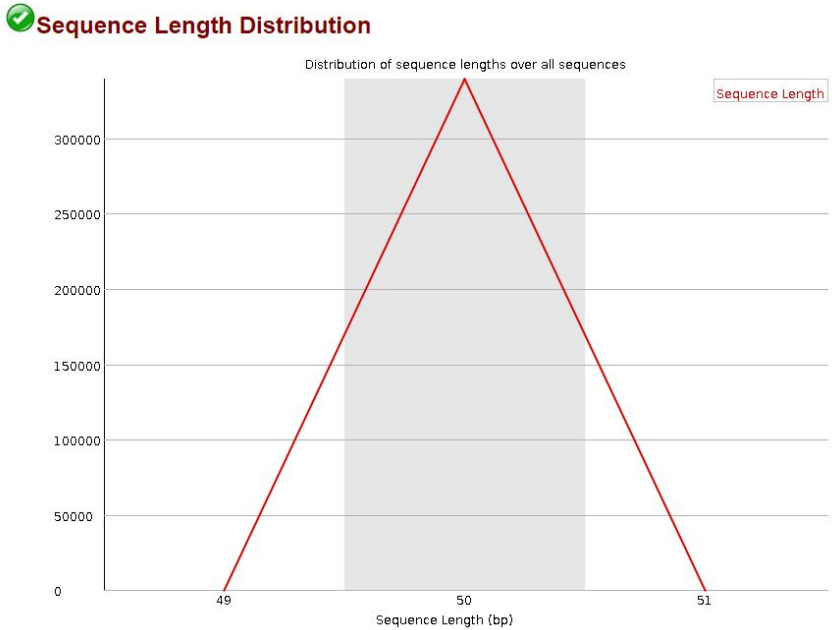


Per base N content



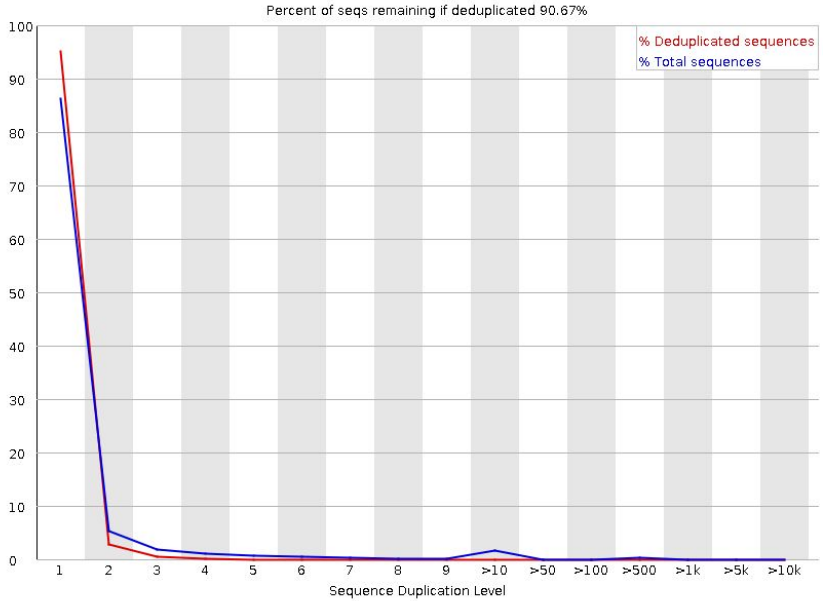
two seq.s are similar



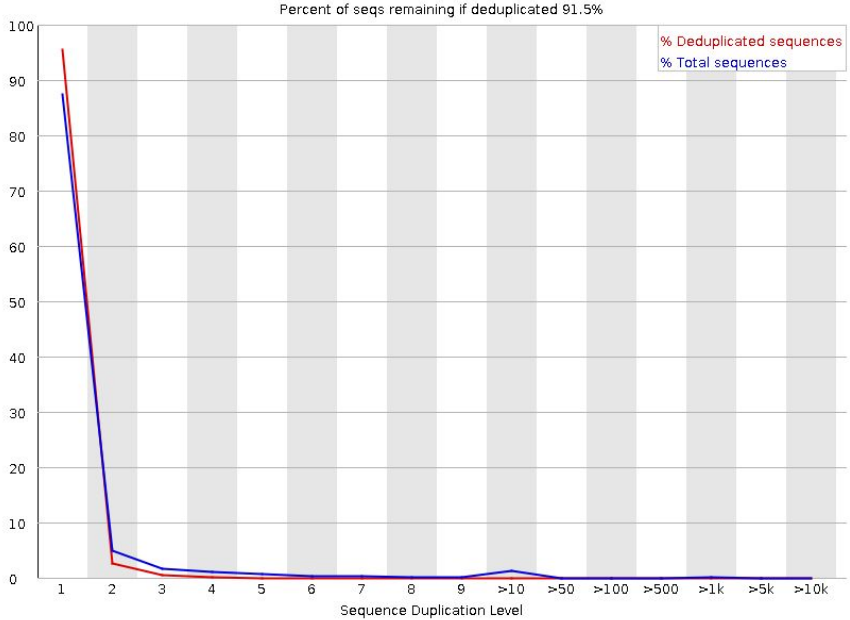


two seq.s are similar

Sequence Duplication Levels



Sequence Duplication Levels



two seq.s are similar with a small difference data #1 has more slightly has overexpressed seq.s than data#2

Overrepresented sequences

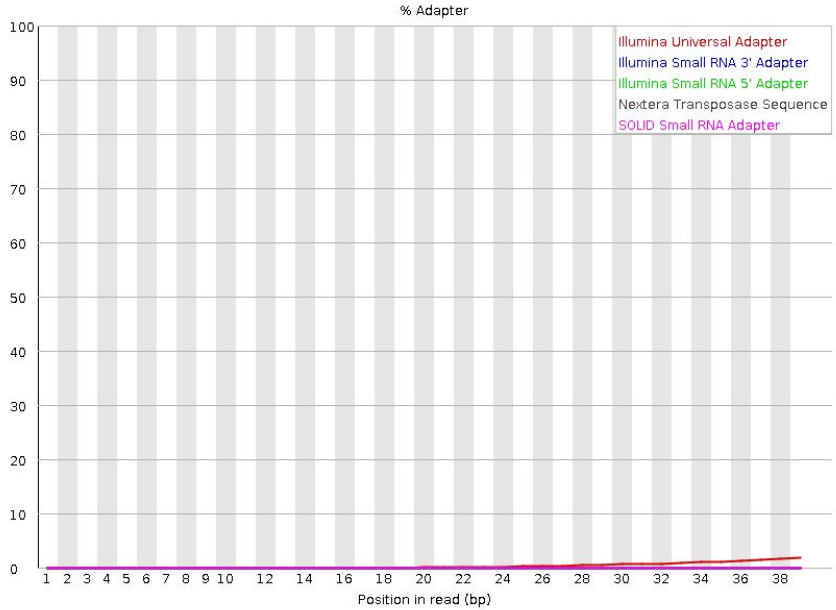
Sequence	Count	Percentage	Possible Source
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCG	957	0.28207123403954304	Illumina Paired End PCR Primer 2 (100% over 31bp)
CGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGCAG	506	0.14914111225079288	Illumina Paired End PCR Primer 2 (96% over 33bp)

Overrepresented sequences

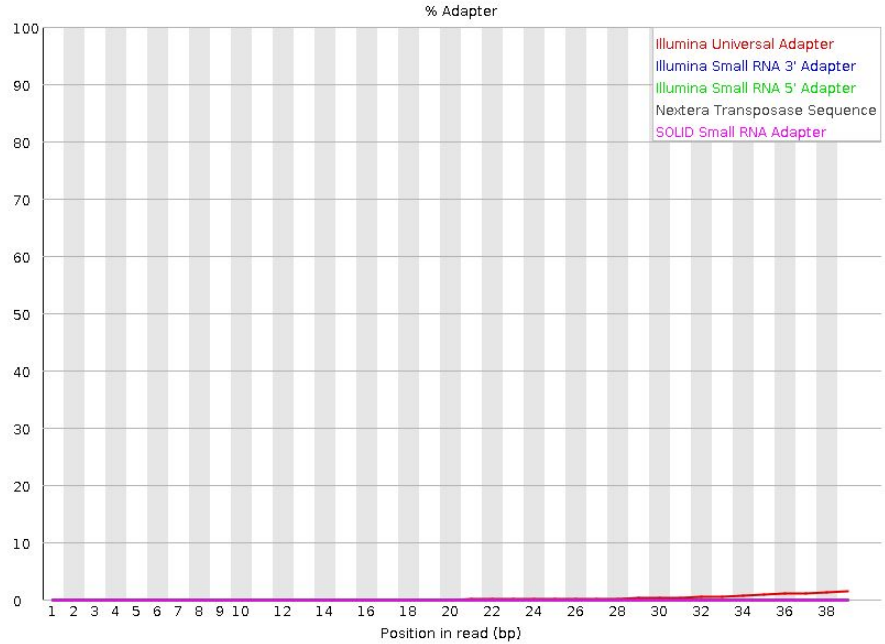
Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	1220	0.35958924297621997	No Hit

two seq.s are similar

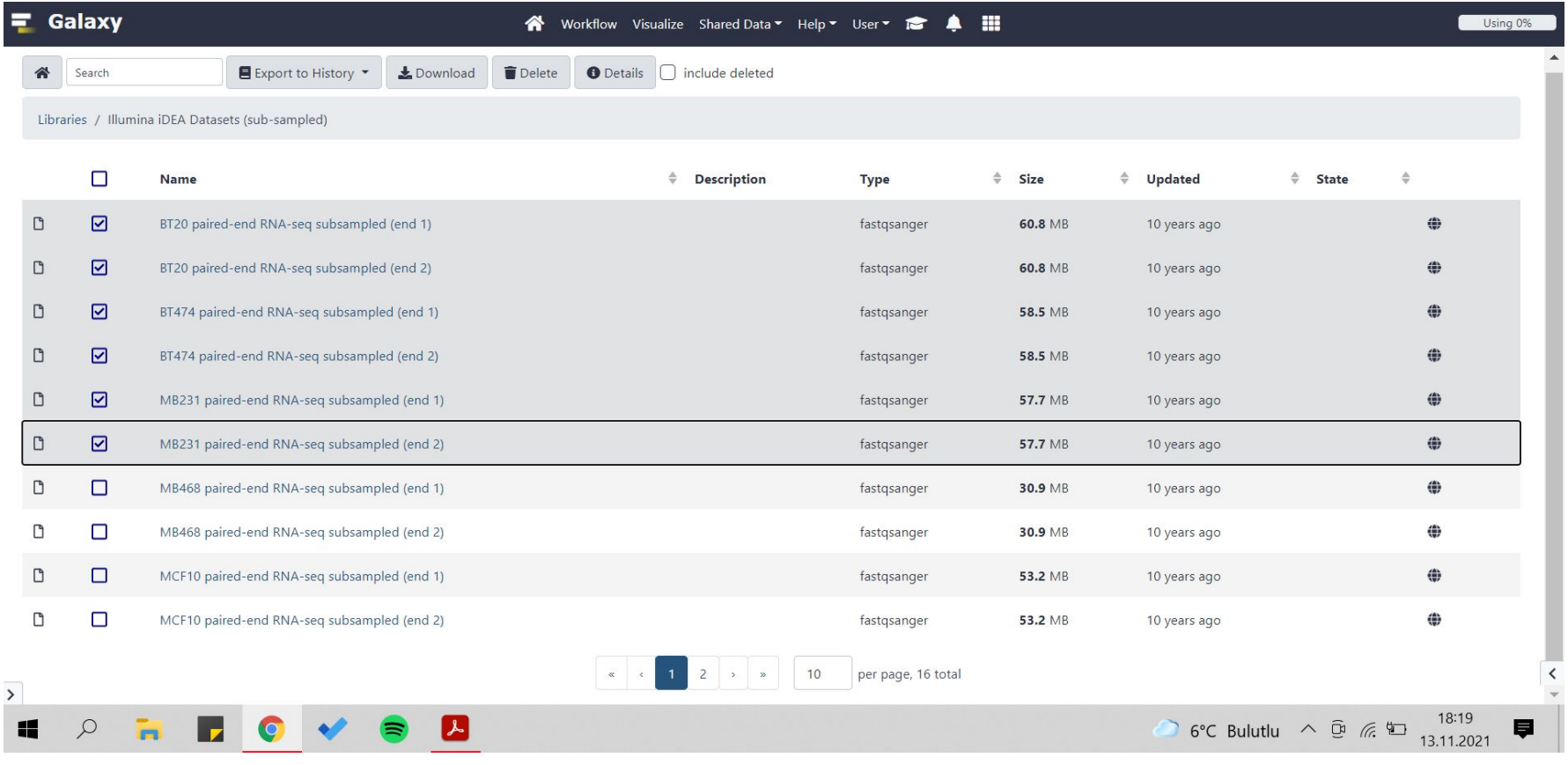
Adapter Content



Adapter Content



part3: getting the data : https://usegalaxy.org/u/leman_nur_nehri/h/assignment3part3



Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **1: BT20 paired-end RNA-seq subsampled (end 1)**

It produces 2 outputs:

- **7: fastp on data 1: Read 1 output**
- **8: fastp on data 1: HTML report**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change

Support, contact, and community

Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **2: BT20 paired-end RNA-seq subsampled (end 2)**

It produces 2 outputs:

- **9: fastp on data 2: Read 1 output**
- **10: fastp on data 2: HTML report**

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

Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **3: BT474 paired-end RNA-seq subsampled (end 1)**

It produces 2 outputs:

- **11: fastp on data 3: Read 1 output**
- **12: fastp on data 3: HTML report**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change

Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **4: BT474 paired-end RNA-seq subsampled (end 2)**

It produces 2 outputs:

- **13: fastp on data 4: Read 1 output**
- **14: fastp on data 4: HTML report**

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

Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **5: MB231 paired-end RNA-seq subsampled (end 1)**

It produces 2 outputs:

- **15: fastp on data 5: Read 1 output**
- **16: fastp on data 5: HTML report**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change

Executed **fastp** and successfully added 1 job to the queue.

The tool uses this input:

- **6: MB231 paired-end RNA-seq subsampled (end 2)**

It produces 2 outputs:

- **17: fastp on data 6: Read 1 output**
- **18: fastp on data 6: HTML report**

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

in general statistics

fastq is a fastq pre-processor : quality control, adapter trimming, quality filtering, per-read quality pruning and many other operations with a single scan of the FASTQ data

****

except **sequencing:**single end (50 cycles), **mean length before filtering:** 50bp and **mean length after filtering:** 50bp, other general statistics are differ from data to data.

duplication rate is nearly same for 6 dataset (not included in this study)

base contents before and after filtering is nearly same (not included)

kmer read countings-ignored

fastp report for BT20 paired-end RNA-seq subsampled _end 1_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	10.414465% (may be overestimated since this is SE data)

Before filtering

total reads:	419.200000 K
total bases:	20.960000 M
Q20 bases:	20.569963 M (98.139136%)
Q30 bases:	19.885924 M (94.875592%)
GC content:	51.039447%

After filtering

total reads:	414.337000 K
total bases:	20.716850 M
Q20 bases:	20.462211 M (98.770860%)
Q30 bases:	19.799622 M (95.572551%)
GC content:	50.958872%

Filtering result

reads passed filters:	414.337000 K (98.839933%)
reads with low quality:	4.293000 K (1.024094%)
reads with too many N:	570 (0.135973%)
reads too short:	0 (0.000000%)

fastp report for BT20 paired-end RNA-seq subsampled _end 2_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	10.112833% (may be overestimated since this is SE data)

Before filtering

total reads:	419.200000 K
total bases:	20.960000 M
Q20 bases:	20.343272 M (97.057595%)
Q30 bases:	19.587029 M (93.449566%)
GC content:	51.058717%

After filtering

total reads:	410.091000 K
total bases:	20.504550 M
Q20 bases:	20.234911 M (98.684980%)
Q30 bases:	19.505922 M (95.129725%)
GC content:	51.123356%

Filtering result

reads passed filters:	410.091000 K (97.827052%)
reads with low quality:	8.939000 K (2.132395%)
reads with too many N:	170 (0.040553%)
reads too short:	0 (0.000000%)

fastp report for BT474 paired-end RNA-seq subsampled _end 1_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	7.947497% (may be overestimated since this is SE data)

Before filtering

total reads:	403.191000 K
total bases:	20.159550 M
Q20 bases:	19.638378 M (97.414764%)
Q30 bases:	18.804090 M (93.276338%)
GC content:	49.871803%

After filtering

total reads:	397.570000 K
total bases:	19.878500 M
Q20 bases:	19.521395 M (98.203562%)
Q30 bases:	18.710670 M (94.125160%)
GC content:	49.799587%

Filtering result

reads passed filters:	397.570000 K (98.605872%)
reads with low quality:	5.058000 K (1.254492%)
reads with too many N:	563 (0.139636%)
reads too short:	0 (0.000000%)

fastp report for BT474 paired-end RNA-seq subsampled _end 2_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	8.158705% (may be overestimated since this is SE data)

Before filtering

total reads:	403.191000 K
total bases:	20.159550 M
Q20 bases:	19.411595 M (96.289823%)
Q30 bases:	18.543744 M (91.984910%)
GC content:	49.817689%

After filtering

total reads:	392.634000 K
total bases:	19.631700 M
Q20 bases:	19.293687 M (98.278229%)
Q30 bases:	18.454989 M (94.006067%)
GC content:	49.886001%

Filtering result

reads passed filters:	392.634000 K (97.381638%)
reads with low quality:	10.437000 K (2.588599%)
reads with too many N:	120 (0.029763%)
reads too short:	0 (0.000000%)

fastp report for MB231 paired-end RNA-seq subsampled _end 1_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	10.081466% (may be overestimated since this is SE data)

Before filtering

total reads:	397.299000 K
total bases:	19.864950 M
Q20 bases:	19.532563 M (98.326766%)
Q30 bases:	18.954842 M (95.418524%)
GC content:	51.589040%

After filtering

total reads:	392.976000 K
total bases:	19.648800 M
Q20 bases:	19.435880 M (98.916371%)
Q30 bases:	18.875352 M (96.063637%)
GC content:	51.523834%

Filtering result

reads passed filters:	392.976000 K (98.911903%)
reads with low quality:	3.747000 K (0.943118%)
reads with too many N:	576 (0.144979%)
reads too short:	0 (0.000000%)

fastp report for MB231 paired-end RNA-seq subsampled _end 2_.fastq

Summary

General

fastp version:	0.20.1 (https://github.com/OpenGene/fastp)
sequencing:	single end (50 cycles)
mean length before filtering:	50bp
mean length after filtering:	50bp
duplication rate:	9.815810% (may be overestimated since this is SE data)

Before filtering

total reads:	397.299000 K
total bases:	19.864950 M
Q20 bases:	19.322485 M (97.269236%)
Q30 bases:	18.667587 M (93.972484%)
GC content:	51.501474%

After filtering

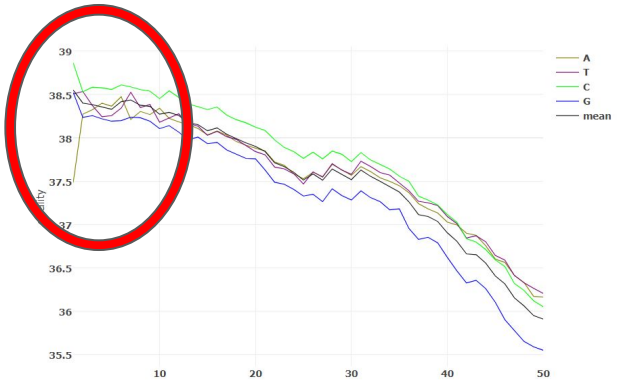
total reads:	389.092000 K
total bases:	19.454600 M
Q20 bases:	19.224503 M (98.817262%)
Q30 bases:	18.593137 M (95.571932%)
GC content:	51.587979%

Filtering result

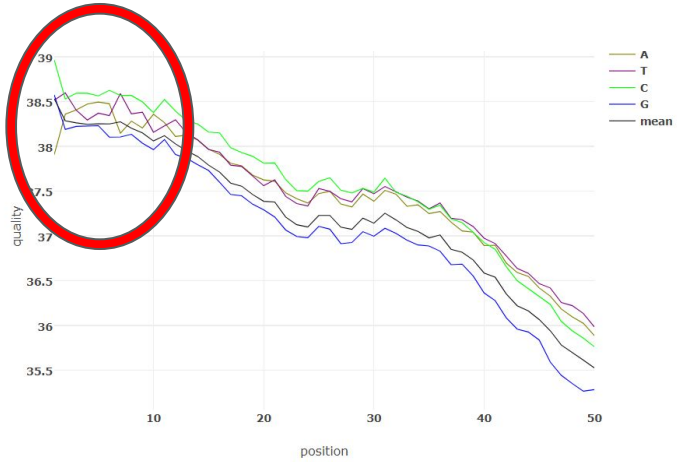
reads passed filters:	389.092000 K (97.934301%)
reads with low quality:	8.066000 K (2.030209%)
reads with too many N:	141 (0.035490%)
reads too short:	0 (0.000000%)

read quality before filtering: small differences // quality control

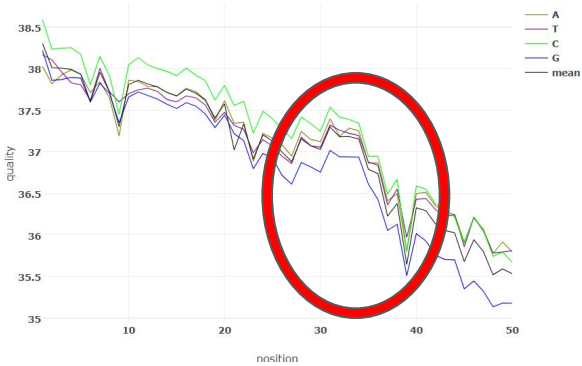
dataset:1



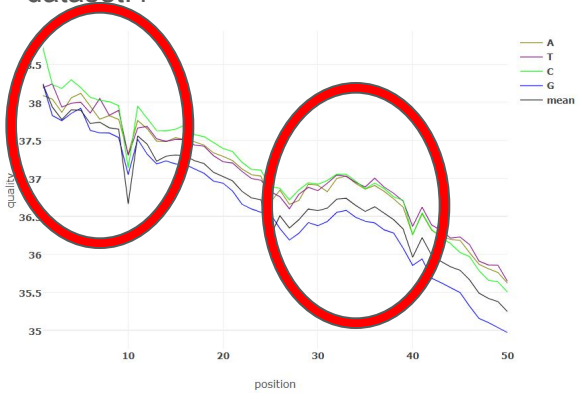
dataset:2



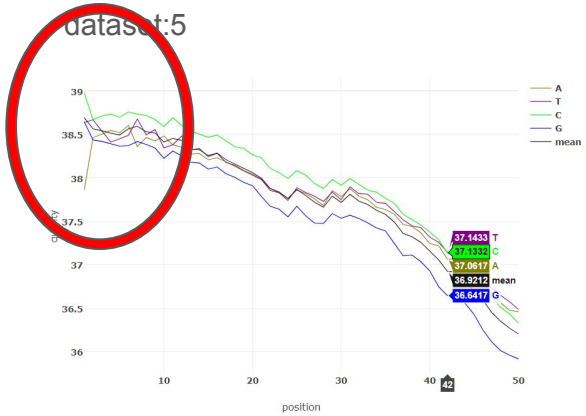
dataset:3



dataset:4



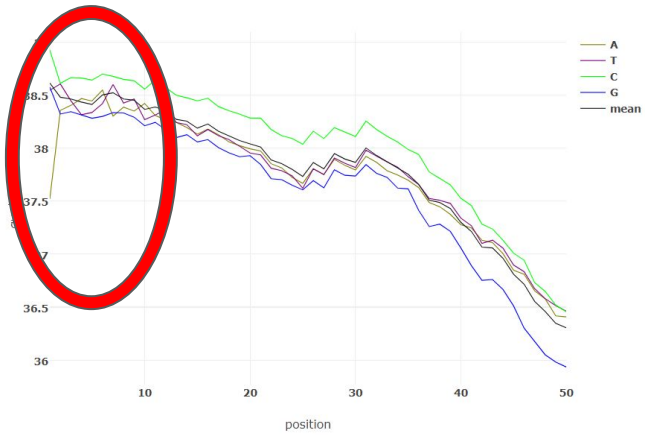
dataset:5



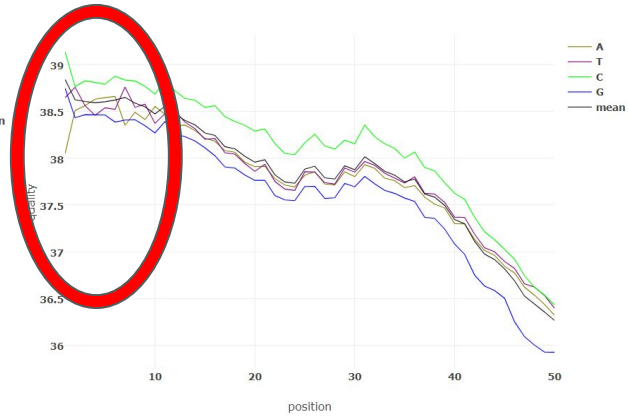
dataset:6

read quality after filtering: small differences
// quality control

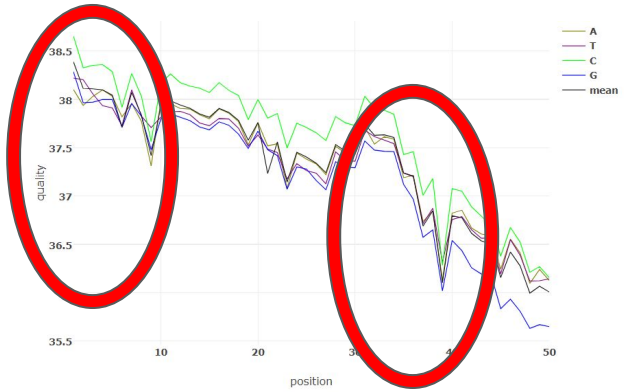
dataset:1



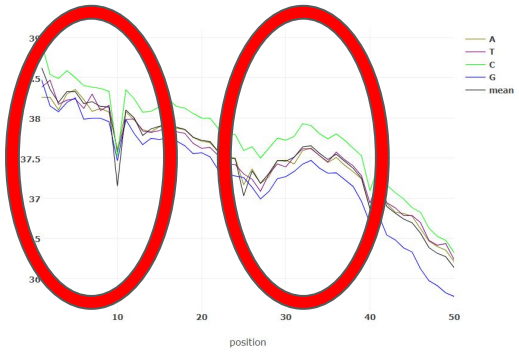
dataset:2



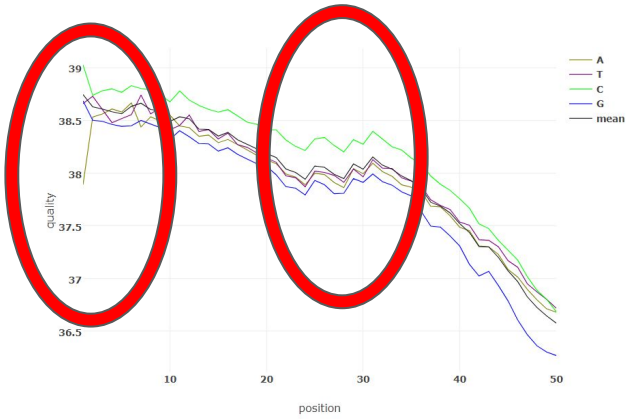
dataset:3



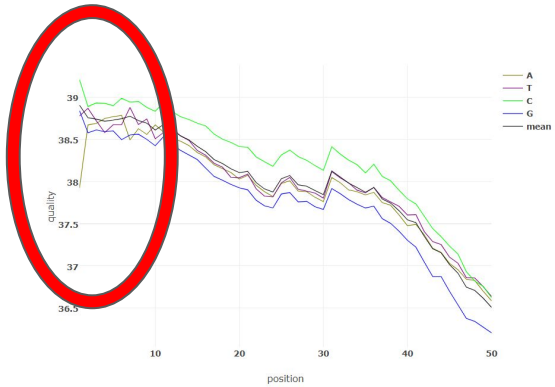
dataset:4



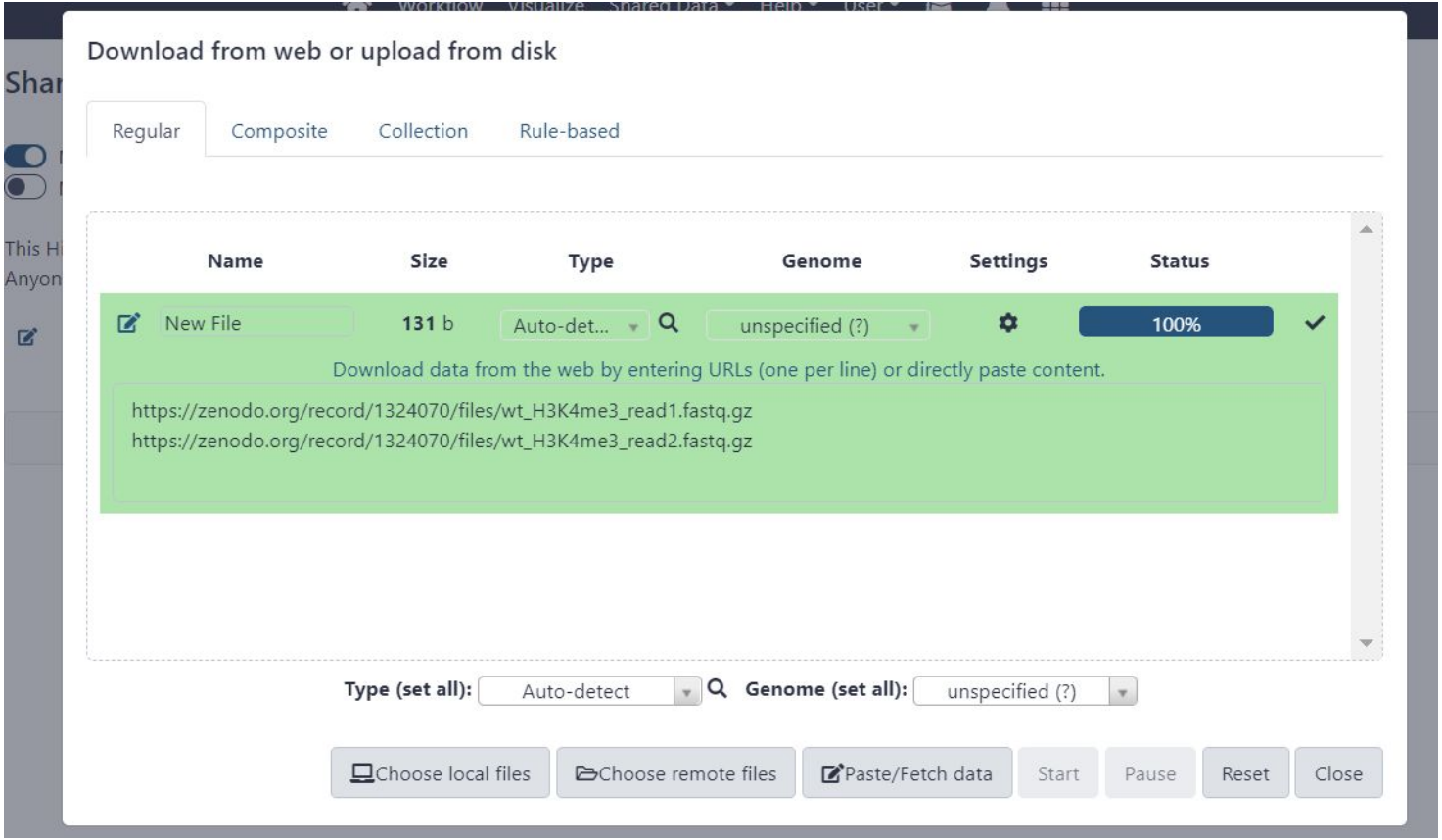
dataset:5



dataset:6



part4 getting the data: https://usegalaxy.org/u/leman_nur_nehri/h/assignment3part4



Is this single or paired library

Paired-end

mapping with bowtie: to align sequences fastly

FASTA/Q file #1

1: wt_H3K4me3_read1.fastq.gz

Must be of datatype "fastqsanger"or "fasta"

FASTA/Q file #2

2: wt_H3K4me3_read2.fastq.gz

Must be of datatype "fastqsanger"or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Do you want to set paired-end options?

No

See "Alignment Options" section of Help below for information

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

Mouse (Mus musculus): mm10

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1: Default setting only

Do you want to use presets?

- (x) No, just use defaults
- ( ) Very fast end-to-end (--very-fast)
- ( ) Fast end-to-end (--fast)
- ( ) Sensitive end-to-end (--sensitive)
- ( ) Very sensitive end-to-end (--very-sensitive)
- ( ) Very fast local (--very-fast-local)
- ( ) Fast local (--fast-local)
- ( ) Sensitive local (--sensitive-local)
- ( ) Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

Do you want to tweak SAM/BAM Options?

- Fast local (--fast-local)
- Sensitive local (--sensitive-local)
- Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

Do you want to tweak SAM/BAM Options?

No

See "Output Options" section of Help below for information

Save the bowtie2 mapping statistics to the history

Yes

Job Resource Parameters

Use default job resource parameters

Email notification

Send an email notification when the job completes.

Execute

Bowtie2 Overview

Executed **Bowtie2** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- **1: wt_H3K4me3_read1.fastq.gz**
- **2: wt_H3K4me3_read2.fastq.gz**

It produces 2 outputs:

- **3: Bowtie2 on data 2 and data 1: alignments**
- **4: Bowtie2 on data 2 and data 1: mapping stats**

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

inspecting the mapping statistics:

```
50000 reads; of these:
  50000 (100.00%) were paired; of these:
    1880 (3.76%) aligned concordantly 0 times
    44731 (89.46%) aligned concordantly exactly 1 time
    3389 (6.78%) aligned concordantly >1 times
    ----
    1880 pairs aligned concordantly 0 times; of these:
      275 (14.63%) aligned discordantly 1 time
    ----
    1605 pairs aligned 0 times concordantly or discordantly; of these:
      3210 mates make up the pairs; of these:
        1882 (58.63%) aligned 0 times
        947 (29.50%) aligned exactly 1 time
        381 (11.87%) aligned >1 times
98.12% overall alignment rate
```

summary of the alignment mapping statistics

inspecting the output

a bam file -which is the version of sam file- includes some standard format attributes like header, alignment, RC, Bc, etc.

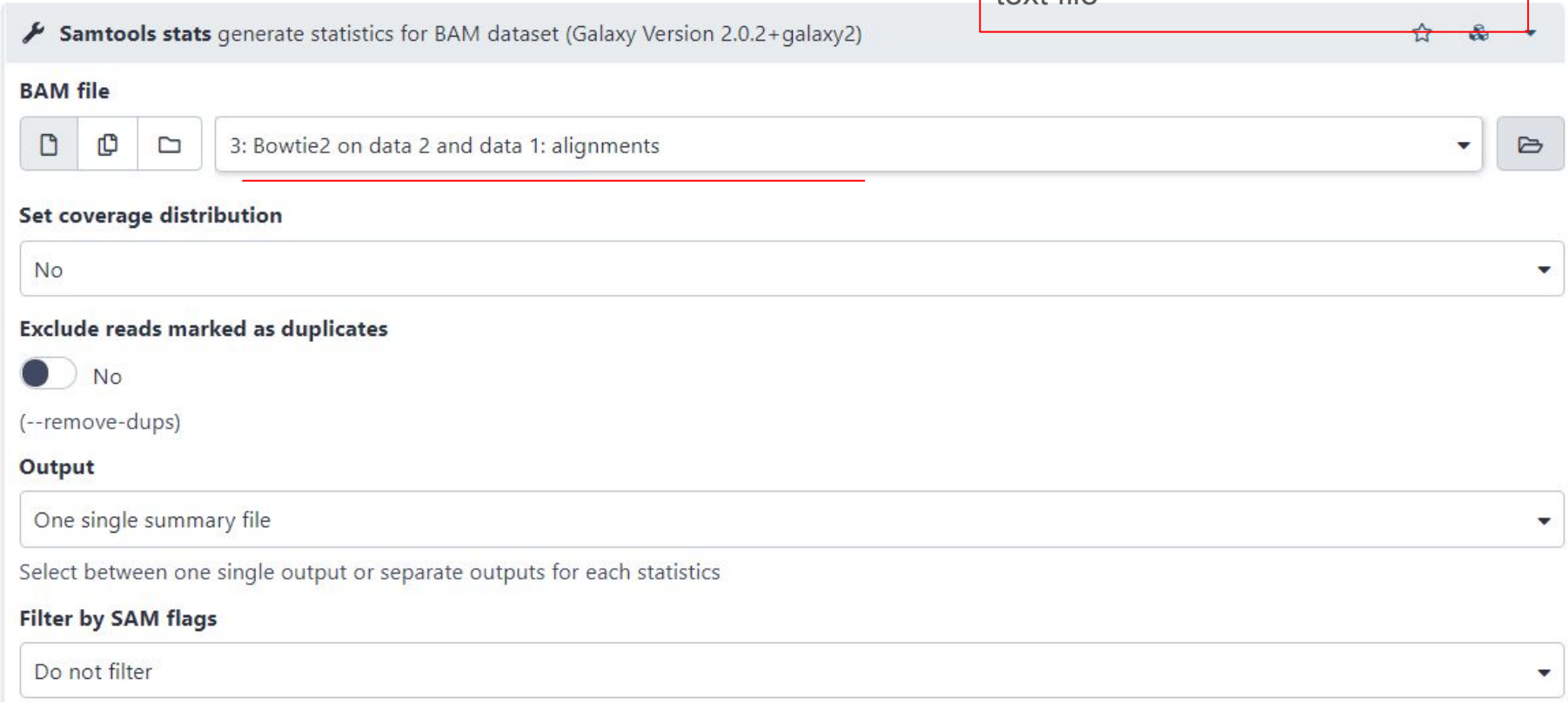
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	C
@HD VN:1.0 SO:coordinate										
@SQ SN:chr1 LN:195471971										
@SQ SN:chr10 LN:130694993										
@SQ SN:chr11 LN:122082543										
@SQ SN:chr12 LN:120129022										
@SQ SN:chr13 LN:120421639										
@SQ SN:chr14 LN:124902244										
@SQ SN:chr15 LN:104043685										
@SQ SN:chr16 LN:98207768										
@SQ SN:chr17 LN:94987271										
@SQ SN:chr18 LN:90702639										
@SQ SN:chr19 LN:61431566										
@SQ SN:chr1_GL456210_random LN:169725										
@SQ SN:chr1_GL456211_random LN:241735										
@SQ SN:chr1_GL456212_random LN:153618										
@SQ SN:chr1_GL456213_random LN:39340										
@SQ SN:chr1_GL456221_random LN:206961										
@SQ SN:chr2 LN:182113224										
@SQ SN:chr3 LN:160039680										
@SQ SN:chr4 LN:156508116										
@SQ SN:chr4_GL456216_random LN:66673										
@SQ SN:chr4_JH584292_random LN:14945										
@SQ SN:chr4_GL456350_random LN:227966										
@SQ SN:chr4_JH584293_random LN:207968										
@SQ SN:chr4_JH584294_random LN:191905										
@SQ SN:chr4_JH584295_random LN:1976										
@SQ SN:chr5 LN:151834684										
@SQ SN:chr5_JH584296_random LN:199368										
@SQ SN:chr5_JH584297_random LN:205776										
@SQ SN:chr5_JH584298_random LN:184189										
@SQ SN:chr5_GL456354_random LN:195993										
@SQ SN:chr5_JH584299_random LN:953012										

@PG ID:samtools PN:samtools PP:bowtie2 VN:1.11 CL:samtools sort -@6 -T /corral4/main/jobs/039/030/39030123/_job_tmp -O bam -o /corral4/main/objects/9/c/f/dataset_9cf92c22-c0ae-4a51-bd7d-75f5f7d9d7af.dat

SRR5680996.8223326	81	chr1	3020404	1	51M	chr10	48032637	0	TTTTGCCATCTGGCAATCTCTAGAGTTAGTTGTTATAGTTGTCTTTGTTTA	HJJJJJJJJ
SRR5680996.2245934	99	chr1	3119659	42	29M1I21M	=	3119797	189	TCAGAAGTCCTGGTTGGGCTGGAGCTCACAAGTGTCCCAGCCTAGCCTTGA	CCCFFFF
SRR5680996.2245934	147	chr1	3119797	42	51M	=	3119659	-189	TGCAAAGAGTAAGCACTGTTCATATAGTCATAGGTACTCTTTATGTCTCTT	GJJJJJJJJ
SRR5680996.20000888	163	chr1	3165868	42	51M	=	3166035	218	TGAGTTTGGCTTTTCTCTAGAGGCTGGAATAACATGGAAATGCTGAAGAAG	@C@FFI
SRR5680996.20000888	83	chr1	3166035	42	51M	=	3165868	-218	TGCTCCTGAAATCATCAGGAACTTCTTGGTATGACTACTGTTGTCTTCACA	JJIGGJIJJ
SRR5680996.5728638	163	chr1	3636076	42	51M	=	3636146	121	CAAAACATACAAACCTCACAGAAGCAGTGTGGTCTTACATCATATTCACGT	CCCFFFF
SRR5680996.5728638	83	chr1	3636146	42	51M	=	3636076	-121	ATGCTGCCATGCATTTCAAGTAATGTTTACATGAGTCAAAAGCATTTTTTC	JJJJJJJJJJ
SRR5680996.10098836	99	chr1	3671374	42	51M	=	3671514	191	CTCCTTTATCTGACACCTGTCCCTGCTGTCTCCTCCTCCCGCCACCTCCTC	?8?DDD
SRR5680996.10098836	147	chr1	3671514	42	51M	=	3671374	-191	CTGGGCCCGGCTAGCCGCACGCTCGGGACCCGCAGGAGCGCCGCCTGGCTG	2B<85@
SRR5680996.15042665	99	chr1	4571483	42	51M	=	4571653	221	AAAGAACGCAAGATGCTTAGGCTGAGGAAGGTATCTAAGTATTTAAAACCC	CCCFFFF
SRR5680996.15042665	147	chr1	4571653	42	51M	=	4571483	-221	AGCGTCTGCAGTTCCCAGGGAAAAGGTGCTACAAGCCGTCTCTGCCCTGCG	GJJJIIGE
SRR5680996.30040942	177	chr1	4779121	2	51M	chr14	47275415	0	TTTTTTTTCCTCCCTTTTTTTATTTTTTTTTTTTTTTTTTTTTTTTTTTA	########
SRR5680996.151987	99	chr1	4785607	42	51M	=	4785764	208	GCCAGGCTCACTCTCGGCAAGGACCTCAGCAGGTCCAGGCTGGTCCCGCAG	?1=B?D=
SRR5680996.151987	147	chr1	4785764	42	51M	=	4785607	-208	GGCCGTTCTGCCCCGCCCCTGTCGTGCTGCAGAGCGACCGCATCCGCGTCG	########
SRR5680996.18610818	99	chr1	4785800	42	51M	=	4785947	198	ACCGCATCCGCGTCGGTAGGCTATGTCTTCCAGCTTCCAGGCTTGATGTCT	CC@FFF
SRR5680996.13469690	163	chr1	4785879	42	51M	=	4786057	229	TGTGAGATCATGCTTTATCCTCTGGCTTTGGTACCATTTTGTGAACGAAAA	CCCFFFF
SRR5680996.18610818	147	chr1	4785947	42	51M	=	4785800	-198	GTAATGACTGAACACGCTGTACGGTAAAAGTGGGGCCTCATGAAGCGAAAA	CIHEIGH
SRR5680996.13469690	83	chr1	4786057	42	51M	=	4785879	-229	AGTGCCTGGCTCTAGAATTAATTGTCTAATTAGGTTCAACCTAATAGCGTG	IJJIJJJJJJ
SRR5680996.25914913	99	chr1	4807631	42	51M	=	4807897	317	GGTCCCTGCTACCTGCCGCTGGTCCAGGCACTGCGGGCGTGGATGGACAGC	CCCFFFF
SRR5680996.25914913	147	chr1	4807897	42	51M	=	4807631	-317	CGCCCGCCAGCCGGTGGATGTGCGGCAACAACATGTCCGCTCCGATGCCCG	###@A:
SRR5680996.3661534	163	chr1	4807909	42	51M	=	4808104	246	GGTGGATGTGCGGCAACAACATGTCCGCTCCGATGCCCGCCGTTGTGCCGG	?@1=D:
SRR5680996.27461678	99	chr1	4807925	42	51M	=	4808010	136	CAACATGTCCGCTCCGATGCCCGCCGTTGTGCCGGCCGCCCGGAAGGCCAC	CCCFFFF
SRR5680996.27461678	147	chr1	4808010	42	51M	=	4807925	-136	ACCAGCGAGACCCAGGCCGCCGCGCGTGGATGCGTCCGCGCCGCCCCGCGG	<DDDD
SRR5680996.3661534	83	chr1	4808104	42	51M	=	4807909	-246	CCTGCCTGCCTAGCACTCTCACCCCCACCCCTGAAGCAAATTCCGAGGTCC	35:=8?A
SRR5680996.12925132	163	chr1	4808222	42	51M	=	4808374	203	TCCCGGAAACTTTTGTGATGAAAAAAATTCGACTTTCGTGACTTGACTTTA	CCCFFFF
SRR5680996.12925132	83	chr1	4808374	42	51M	=	4808222	-203	TTAGGTCCTTCATTGGGTTTTATAACTTAACAAAATCTGTAACAGTCCAGT	JJIJIGIJIJ
SRR5680996.27018995	99	chr1	4857847	42	51M	=	4857995	199	CAACGTCCGCGCAGCTTCTCTCGCTTCCCGTCCGCTGTGGCCGGGCGAGGG	CCCFFFF
SRR5680996.27018995	147	chr1	4857995	42	51M	=	4857847	-199	GGCAGGCCGCGGCCGAGTCCGCGACACCGGCCCACGCGGAGTCAGTCAGGG	ADDDDD

samtools stats

samtools stats collects data from bam files and gives an output as text file



No cutoff if left empty (--read-length)

Report only the main part of inserts

default=0.99 (--most-inserts)

BWA trim parameter

default=0 (--trim-quality)

Use a reference sequence

Locally cached

Required for GC-depth and mismatches-per-cycle calculation

Using genome

Mouse (Mus Musculus): mm10

Filter by regions

No

restricts output to only those alignments which overlap the specified region(s)

Suppress absence of insertions

inspecting the stats

the output can be visualized as graphs also

```
# This file was produced by samtools stats (1.9+htslib-1.9) and can be plotted using plot-bamstats
# This file contains statistics for all reads.
# The command line was: stats --ref-seq /cvmfs/data.galaxyproject.org/byhand/mm10/sam_index/mm10.fa -@ 0 infile
# CHK, Checksum	[2]Read Names	[3]Sequences	[4]Qualities
# CHK, CRC32 of reads which passed filtering followed by addition (32bit overflow)
CHK	195ef5c8	83b341d3
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN	raw total sequences:	100000
SN	filtered sequences:	0
SN	sequences:	100000
SN	is sorted:	1
SN	1st fragments:	50000
SN	last fragments:	50000
SN	reads mapped:	98118
SN	reads mapped and paired:	97152
SN	reads unmapped:	1882
SN	reads properly paired:	96240
SN	reads paired:	100000
SN	reads duplicated:	0
SN	reads MQ0:	574
SN	reads QC failed:	0
SN	non-primary alignments:	0
SN	total length:	5100000
SN	total first fragment length:	2550000
SN	total last fragment length:	2550000
SN	bases mapped:	5004018
SN	bases mapped (cigar):	5004018
SN	pairs with other orientation:	16
SN	pairs on different chromosomes:	381
SN	percentage of properly paired reads (%):	96.2
# First Fragment Qualities. Use `grep ^FFQ | cut -f 2-` to extract this part.
# Columns correspond to qualities and rows to cycles. First column is the cycle number.
FFQ	1	0	0	30	0	0
FFQ	2	0	0	59	0	0
FFQ	3	0	0	3	0	0
FFQ	4	0	0	3	0	0
FFQ	5	0	0	5	0	0
FFQ	6	0	0	6	0	0
FFQ	7	0	0	7	0	0
FFQ	8	0	0	8	0	0
FFQ	9	0	0	8	0	0
FFQ	10	0	0	9	0	0
FFQ	11	0	0	12	0	0
FFQ	12	0	0	14	0	0
FFQ	13	0	0	18	0	0
FFQ	14	0	0	20	0	0
FFQ	15	0	0	25	0	0
FFQ	16	0	0	30	0	0
FFQ	17	0	0	36	0	0
FFQ	18	0	0	44	0	0
FFQ	19	0	0	62	0	0
FFQ	20	0	0	77	0	0
FFQ	21	0	0	86	0	0
```

JBrowse genome browser (Galaxy Version 1.16.11+galaxy1)

it is a genome browser, it visualise the map reads

Reference genome to display

Use a built-in genome

Built-in references

Select a reference genome

Mouse (Mus Musculus): mm10

If your genome of interest is not listed, contact the Galaxy team

Output JBrowse

Minimal for viewing (Documentation removed)

Genetic Code

1. The Standard Code

JBrowse-in-Galaxy Action

New JBrowse Instance

Annotation Track

1: Annotation Track

Track Type

BAM Pileups

BAM Track Data

3: Bowtie2 on data 2 and data 1: alignments

Autogenerate SNP Track

Yes

Not recommended for deep coverage BAM files

Maximum size of BAM chunks

5000000

Maximum size in bytes of BAM chunks that the browser will try to deal with. When this is exceeded, most tracks will display 'Too much data' message.

JBrowse Custom Track Config [Advanced]

Track Visibility

On for new users

Override Apollo Plugins

Executed **JBrowse** and successfully added 1 job to the queue.

The tool uses this input:

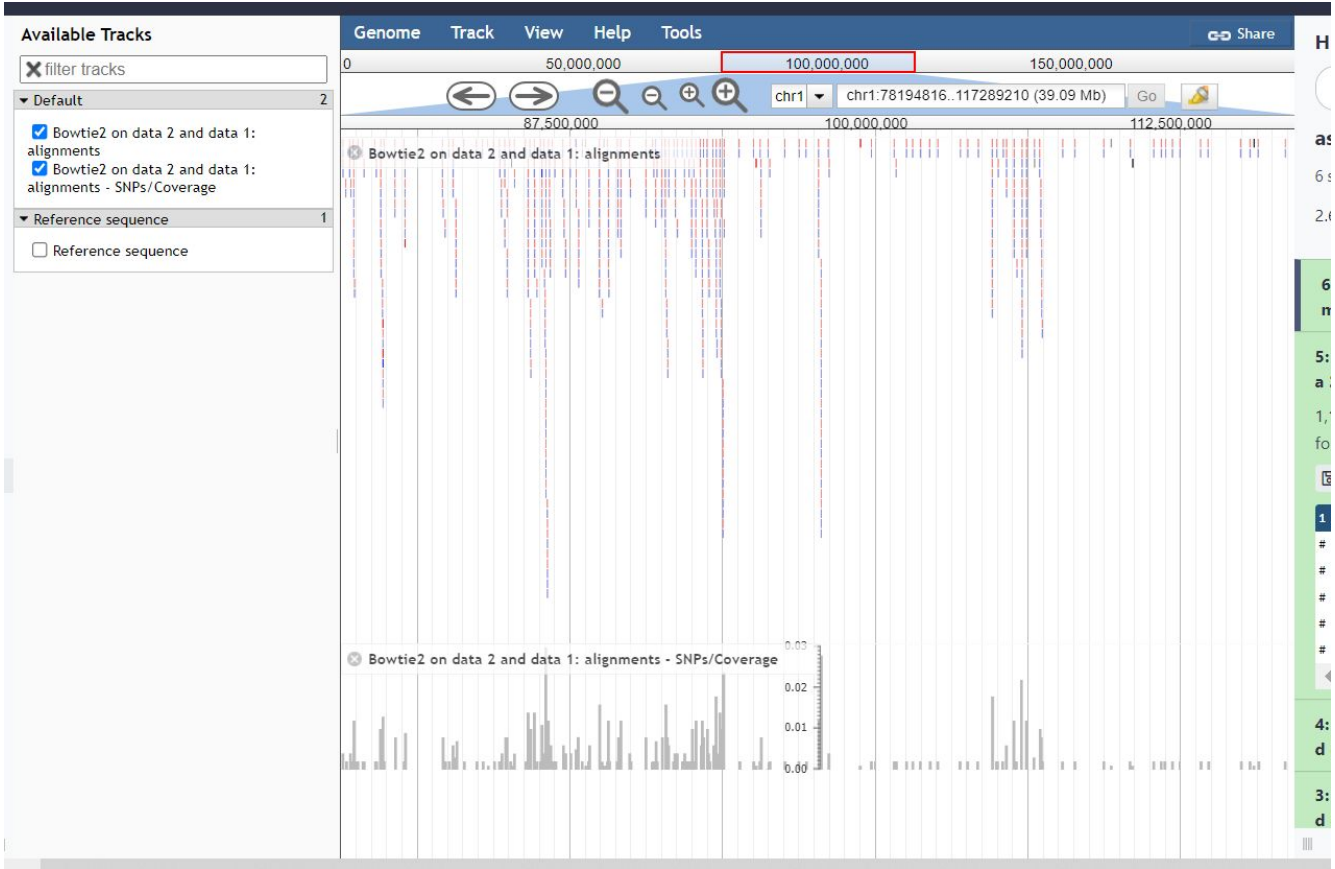
- **3: Bowtie2 on data 2 and data 1: alignments**

It produces this output:

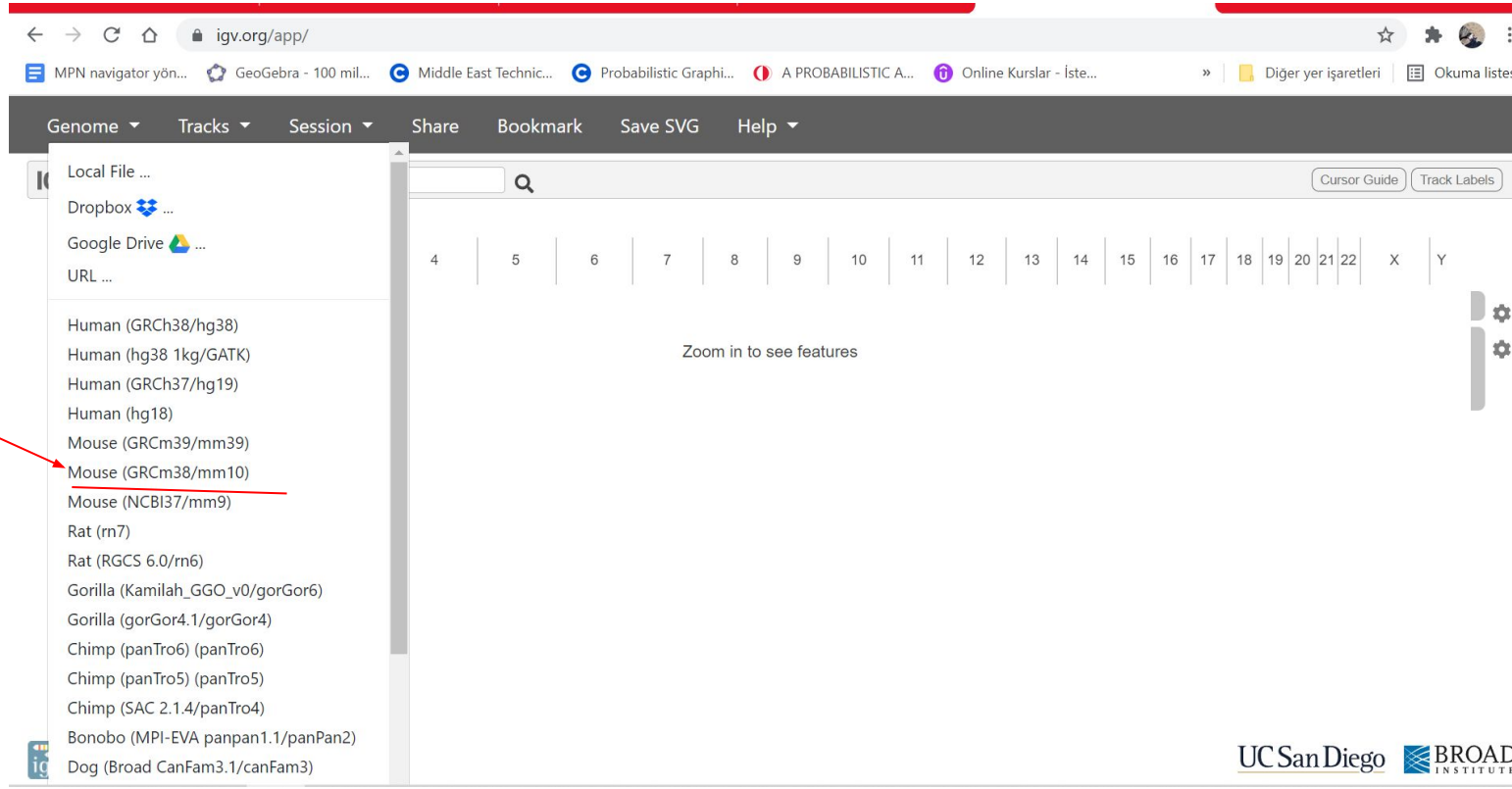
- **6: JBrowse on data 3 - minimal**

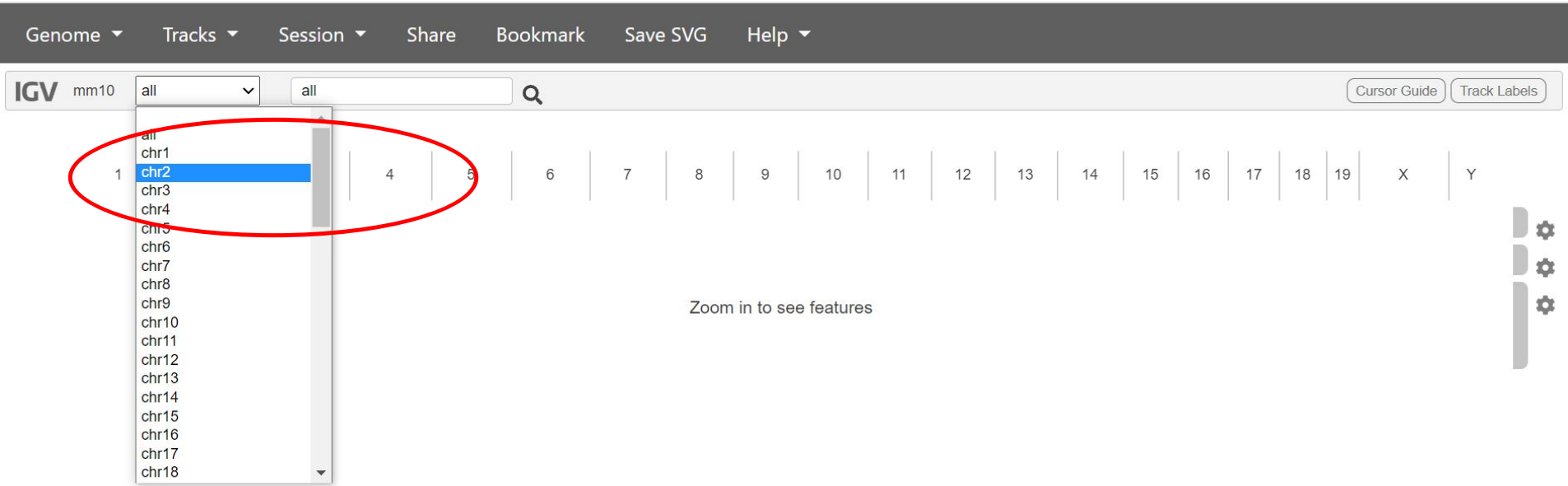
You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

visualizing dataset

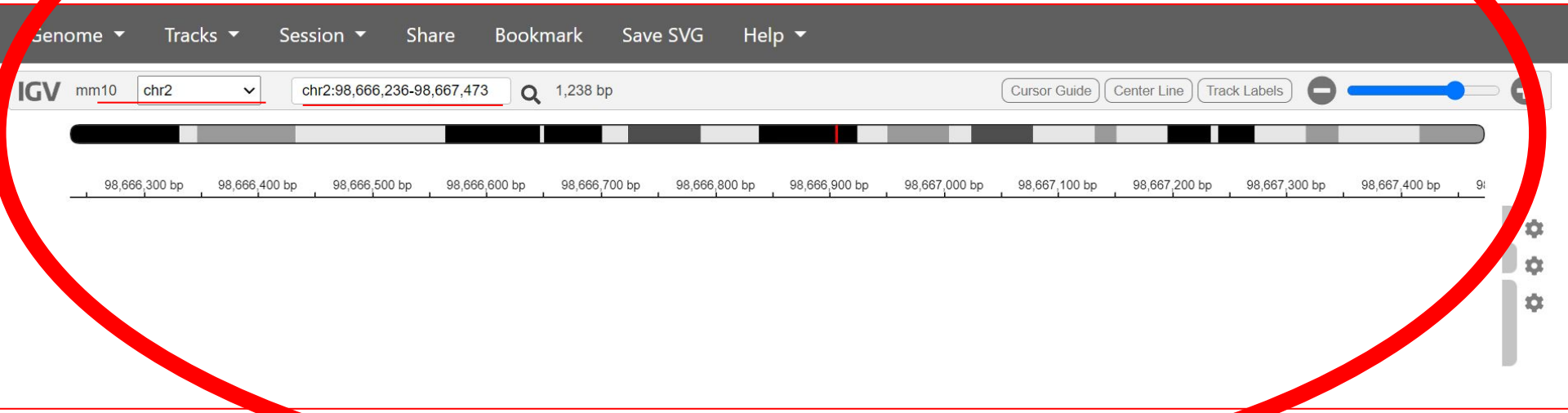


IGV-without downloading




Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
mm10
all
all
Cursor Guide
Track Labels
chr1
chr2
chr3
chr4
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
1
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Y
Zoom in to see features


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
mm10
chr2
chr2:98,666,236-98,667,473
182 mb
Cursor Guide
Center Line
Track Labels
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
100 mb
110 mb
120 mb
130 mb
140 mb
150 mb
160 mb
170 mb
180 mb


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
mm10
chr2
chr2:98,666,236-98,667,473
1,238 bp
Cursor Guide
Center Line
Track Labels
98,666,300 bp
98,666,400 bp
98,666,500 bp
98,666,600 bp
98,666,700 bp
98,666,800 bp
98,666,900 bp
98,667,000 bp
98,667,100 bp
98,667,200 bp
98,667,300 bp
98,667,400 bp

References

https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

https://www.engage-europe.eu/-/media/Sites/engage-europe/Ressources/Protocols-and-Training/D7-1-ENGAGE_Sequencing_quality_report.ashx?la=da&hash=7DB0B76928D1C00B67E41C7A9E76F17A250D263C#:~:text=Per%20base%20sequence%20quality.,FastQ%20file%20(raw%20reads).&text=It%20measures%20the%20GC%20content,normal%20distribution%20of%20GC%20content.

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html

https://academic.oup.com/bioinformatics/article/34/17/i884/5093234