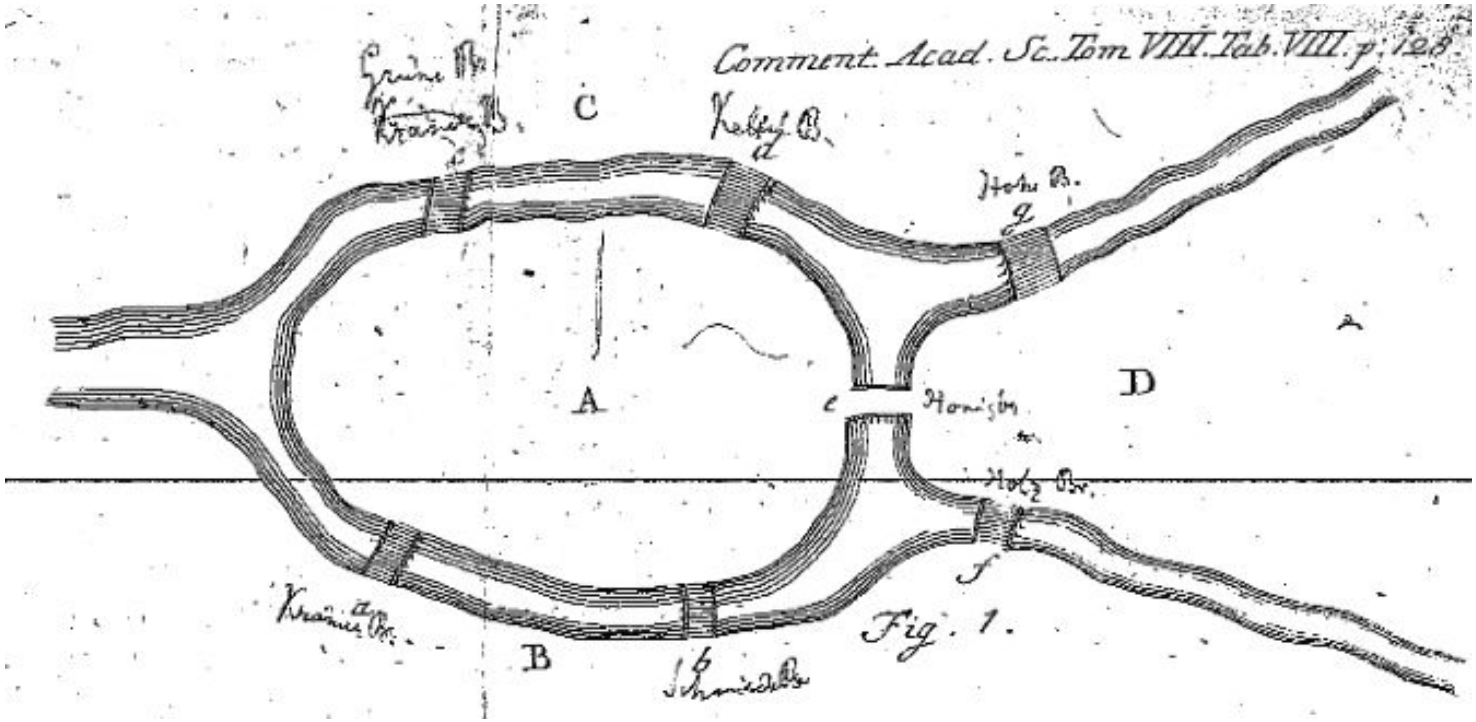
Quality control and cleaning data



By Dr. Richard Allen White III

Lecture 3 - Sep 9th, 2019

Zoom! 404-899-586

Cleaning and quality control

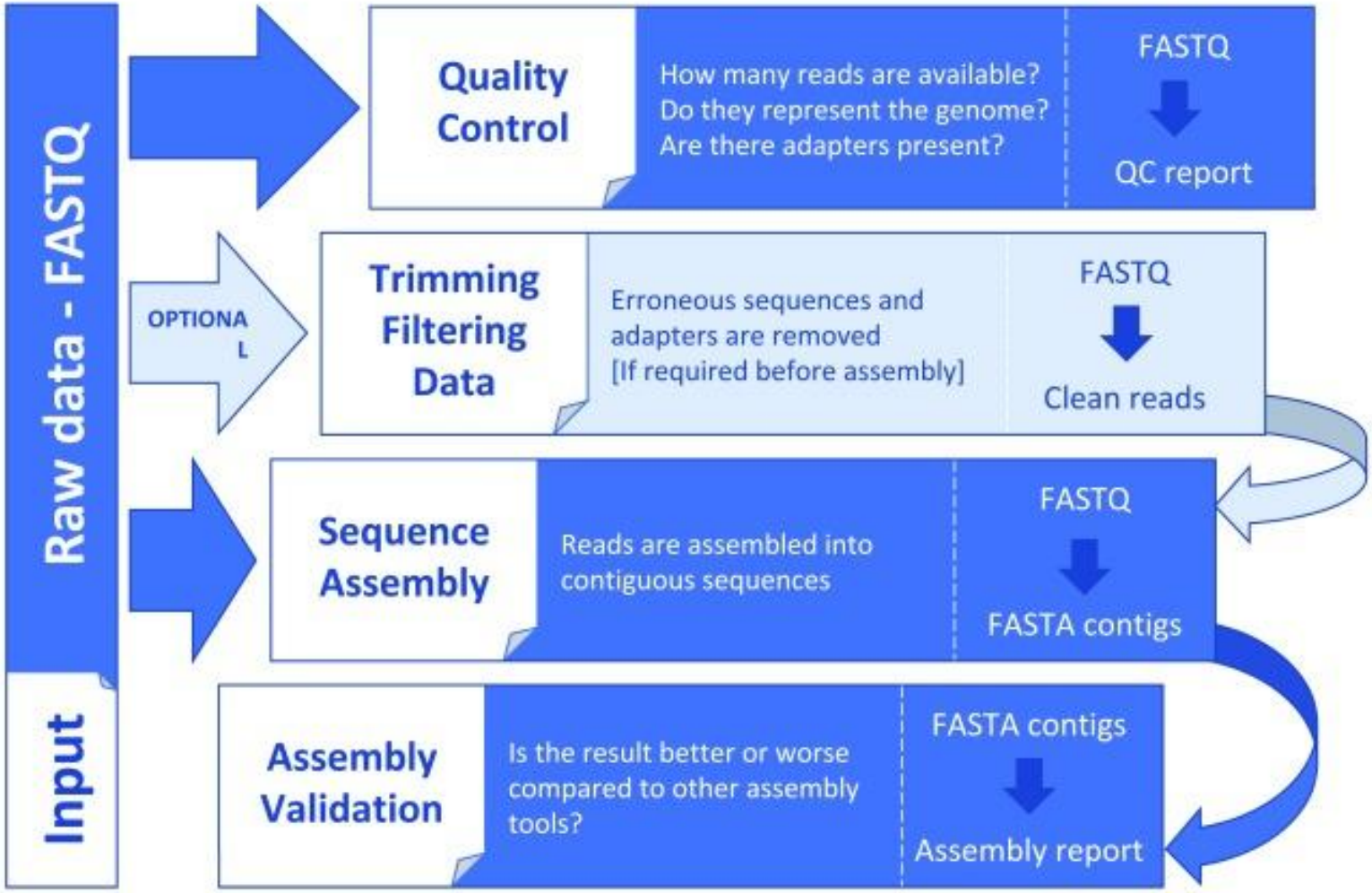
Concepts:

- Sequencing Sanger vs. Illumina
- Error profiles in Illumina
- Review Genomics files
- Good, Bad and ugly Illumina data

Learning Objectives:

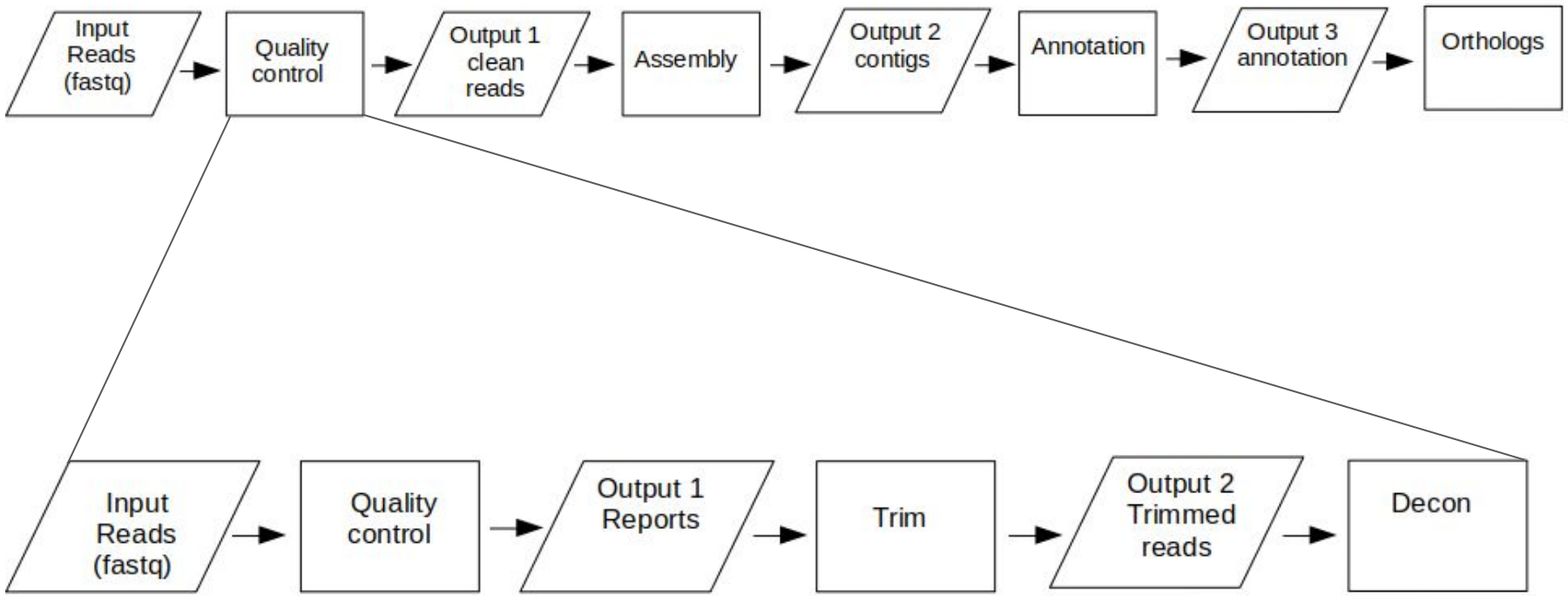
- UNIX command line review/exercise
- Quality control of fastq files
- How to clean and quality control Illumina files

Genome sequencing - flowgraph

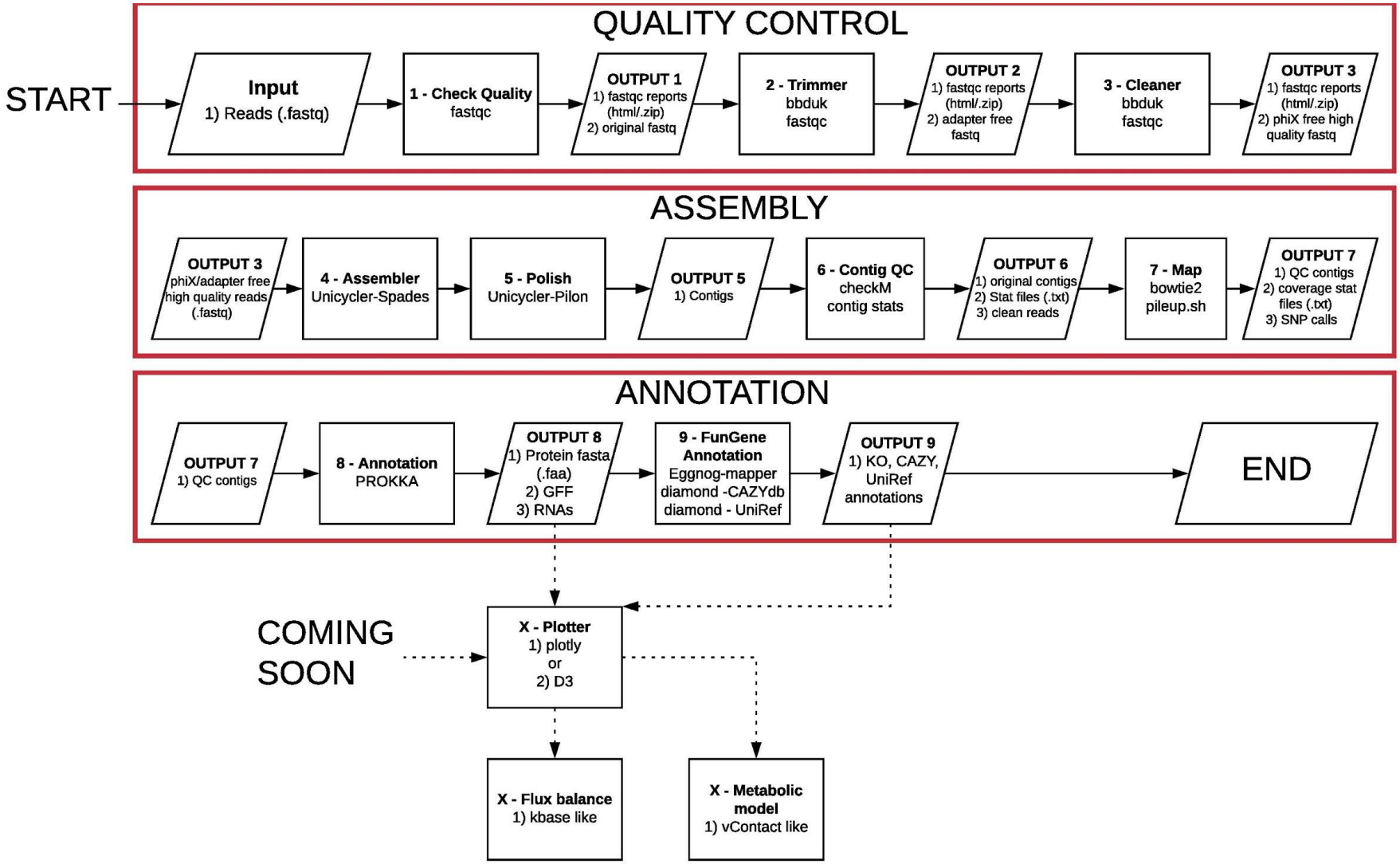


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850084/

Quality control - flowgraph



Example pipeline/workflow



UNIX Review

cd - ?

pwd - ?

mkdir - ?

rm - ?

head - ?

tail - ?

more - ?

| - ?

mv - ?

touch - ?

grep - Global regular expression print.

sed - ?

UNIX Review

cd - change directory
pwd - print working directory
mkdir - make directory
rm - remove
head - prints first 10 lines
tail - prints last 10 lines
more - scroll file top to bottom
| - pipe
mv - move (can rename file/directory)
touch - creates a file
grep - global regular expression print
sed - stream editor

UNIX Review

cat file1.txt file2.txt >>file3.txt ?

What would the command tail file.txt do?

head -100 file.txt | more

touch file.txt file2.txt ?

Head vs. Tail ?

UNIX Tools “grep and sed”

grep - Global regular expression print. The “**hand of the gods,**” of computation.

```
>seq1
ATCCATA
>seq2
GGGTACC
```

COPY and Paste into .txt file.
Count the number of sequences? In unix.

UNIX Tools “grep and sed”

Count the number of sequences? In unix.

= 2

cat file.txt | grep “seq” | wc -l

Is there a shorter way to write this?

Another way to write this?

Count the number of “CC” using grep = answer?

UNIX Tools “grep and sed”

sed - stream editor. The “**dog,**” of computation. Your “best friend.”

```
>seq1
ATCCATA
>seq2
GGGTACC
```

Replace all the T’s with U’s. Using sed in your terminal.

UNIX Tools “grep and sed”

Replace all the T’s with U’s. Using sed in your terminal.

```
sed -i 's/T/U/g' grep.txt
```

95% of your time will be formatting and cleaning data!
Use these tools to help!!

UNIX Challenge

Download these files:
https://www.dropbox.com/s/77csy6stf36r8b0/mystery.fastq?dl=0
https://www.dropbox.com/s/o5duvdpk2iyqo8m/mystery.fasta?dl=0

How can I use UNIX to grab them?

UNIX Challenge

In both files using UNIX ONLY:

- Number of lines total per file?
- Number of sequences per file?

For the mystery.fasta UNIX ONLY:

- How many “sp” or species in the file?
- How many “TAATACA” in the file?

For the mystery.fastq UNIX ONLY:

- How many “AGGCCATT” or species in the file?
- What was the barcode used?
- Is it an R1 or R2 file? (Bonus)

UNIX Challenge

In both files using UNIX ONLY:

- Code used:
- Answers:

For the mystery.fasta UNIX ONLY:

- Code used:
- Answers:

For the mystery.fastq UNIX ONLY:

- Code used:
- Answers:

UNIX Review - resources

https://github.com/raw937/bacterialpopulationgenomics/blob/master/doc/unix_introduction.md

https://github.com/raw937/bacterialpopulationgenomics/blob/master/doc/unix_introduction.md

Branch: master | bacterialpopulationgenomics / doc / unix_introduction.md | Find file | Copy path

raw937 Create unix_introduction.md ef0db79 13 days ago

1 contributor

333 lines (233 sloc) | 8.2 KB | Raw | Blame | History

Introduction to command line basics

A. Bash Basics.

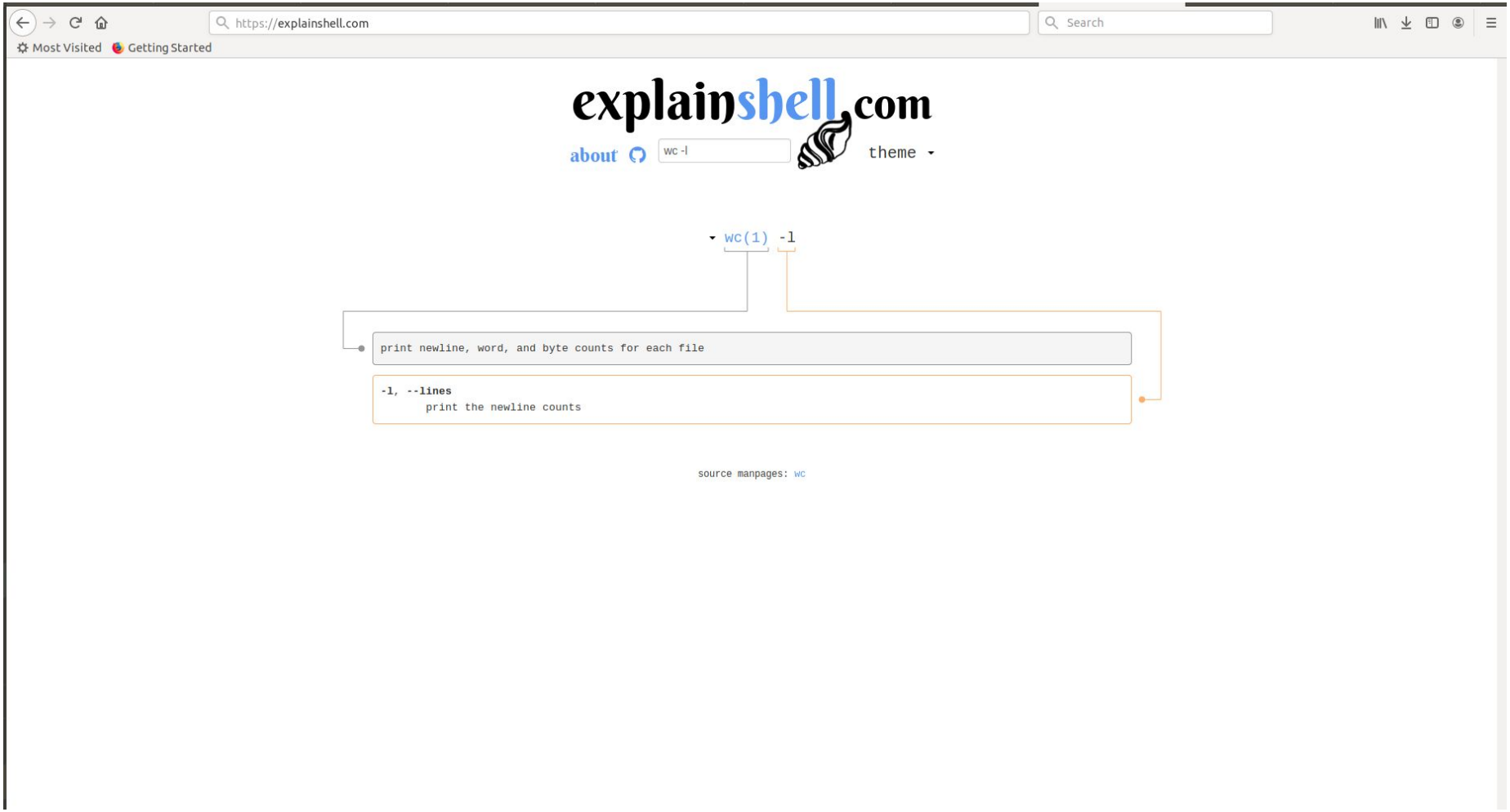
Displays all environment variables

```
export
```

Displays the shell you're using

UNIX Review - resources

https://explainshell.com/



File formats in genomics

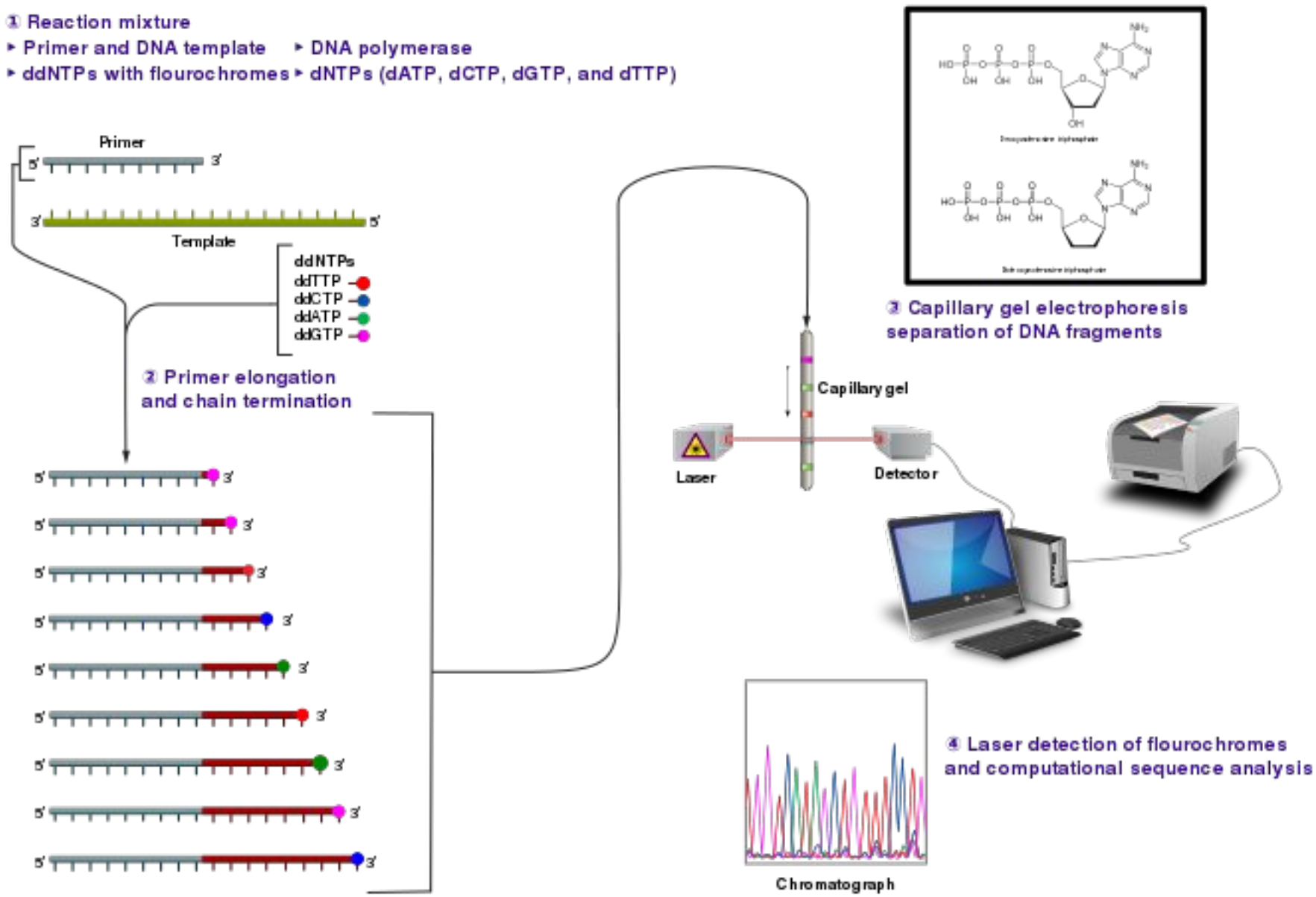
*.fna, .fasta, .fa	?
*.faa	?
*.gff	?
*.gbk	?
*.fq or fastq	?

Read?

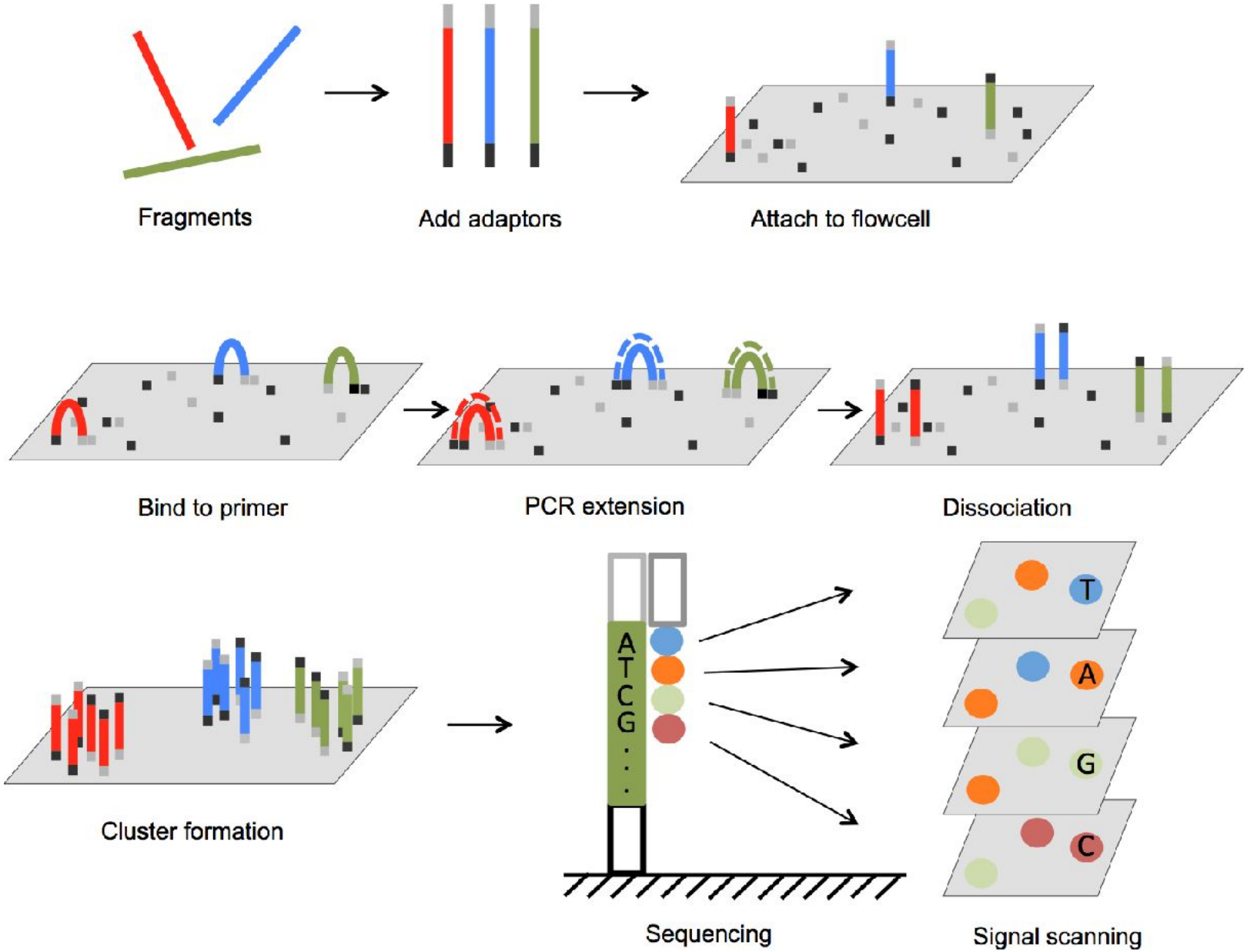
Contig?

kmer?

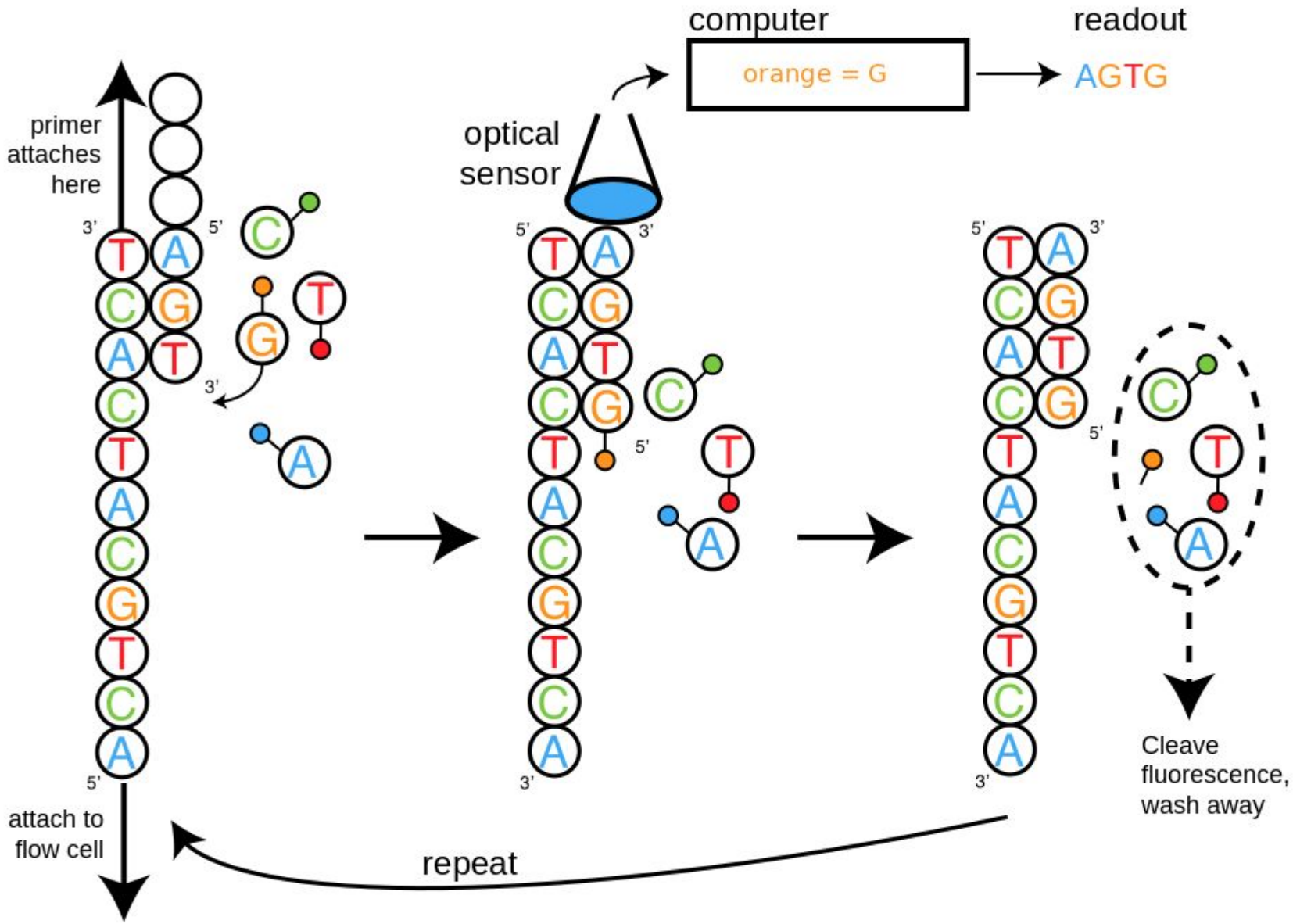
Sanger sequencing: dideoxy-chain termination



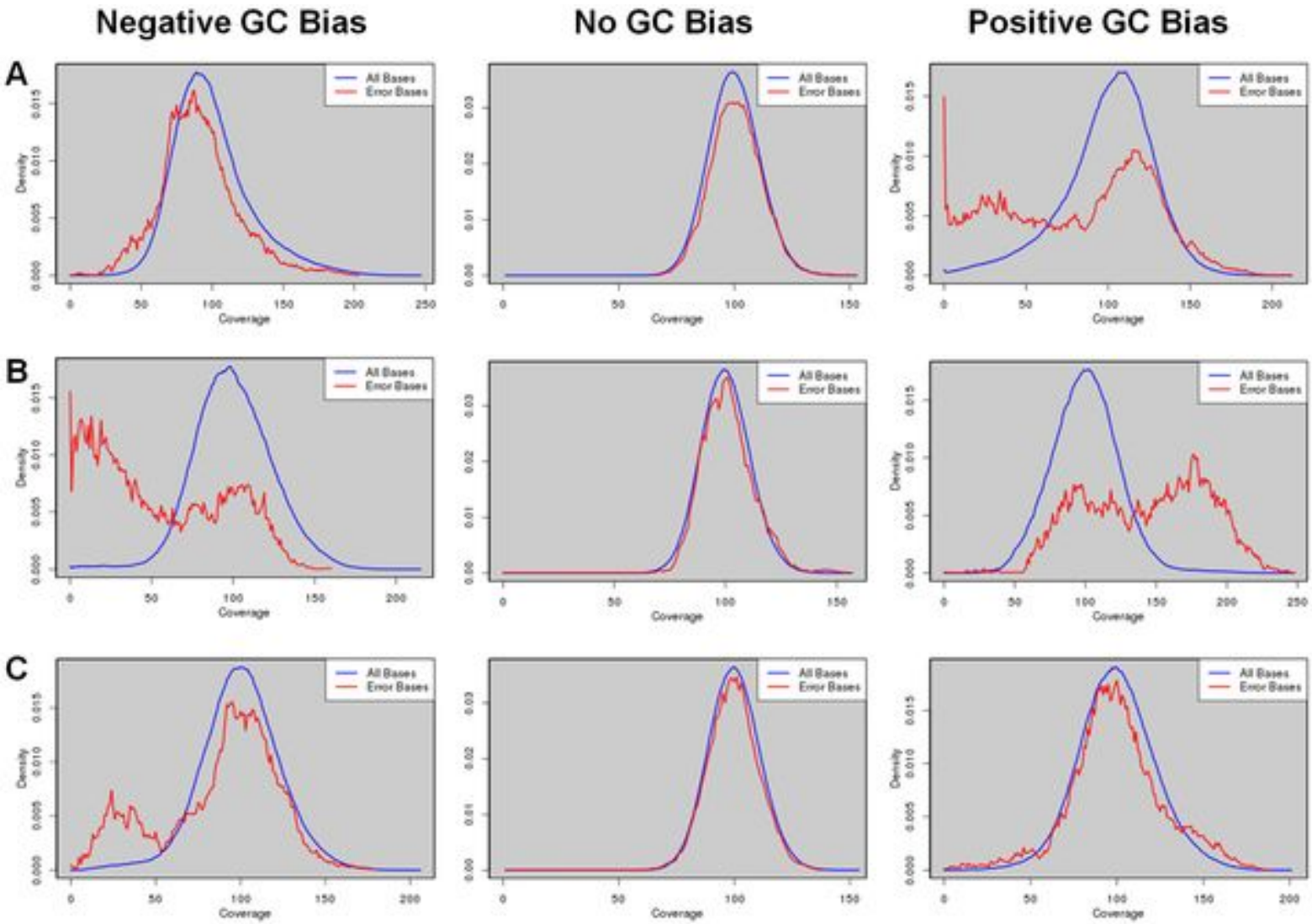
Illumina sequencing - reversible chain termination



Illumina sequencing - substitution error

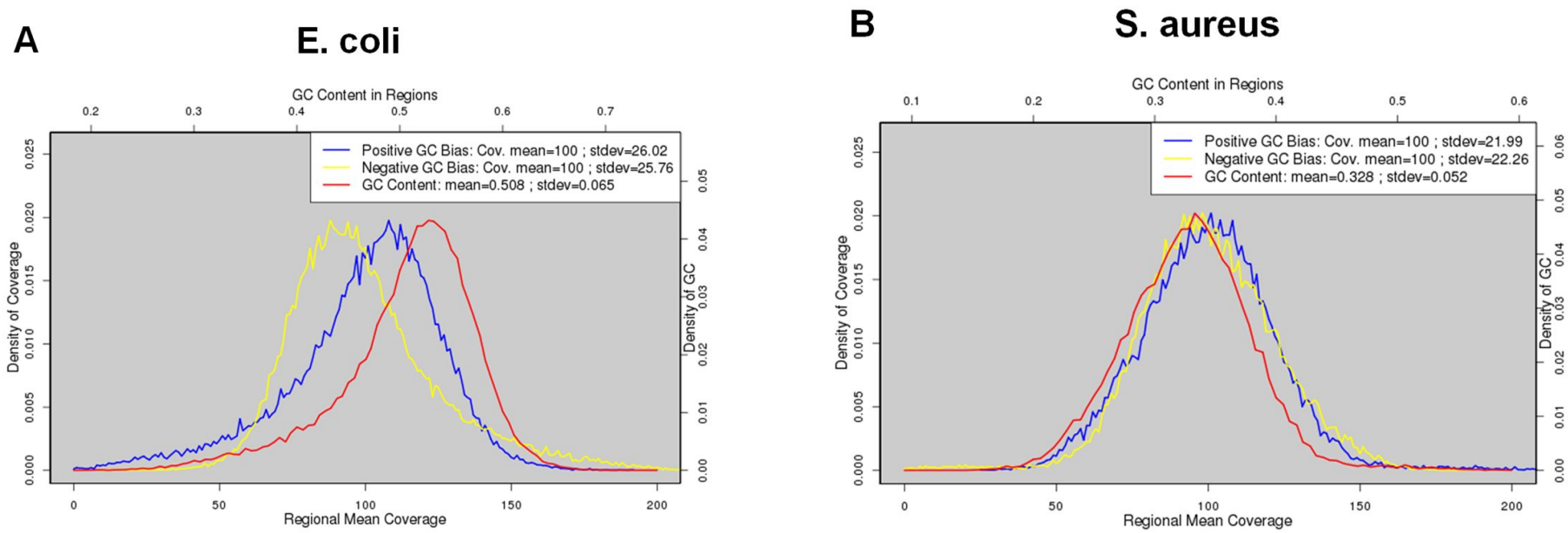


Illumina sequencing - GC bias



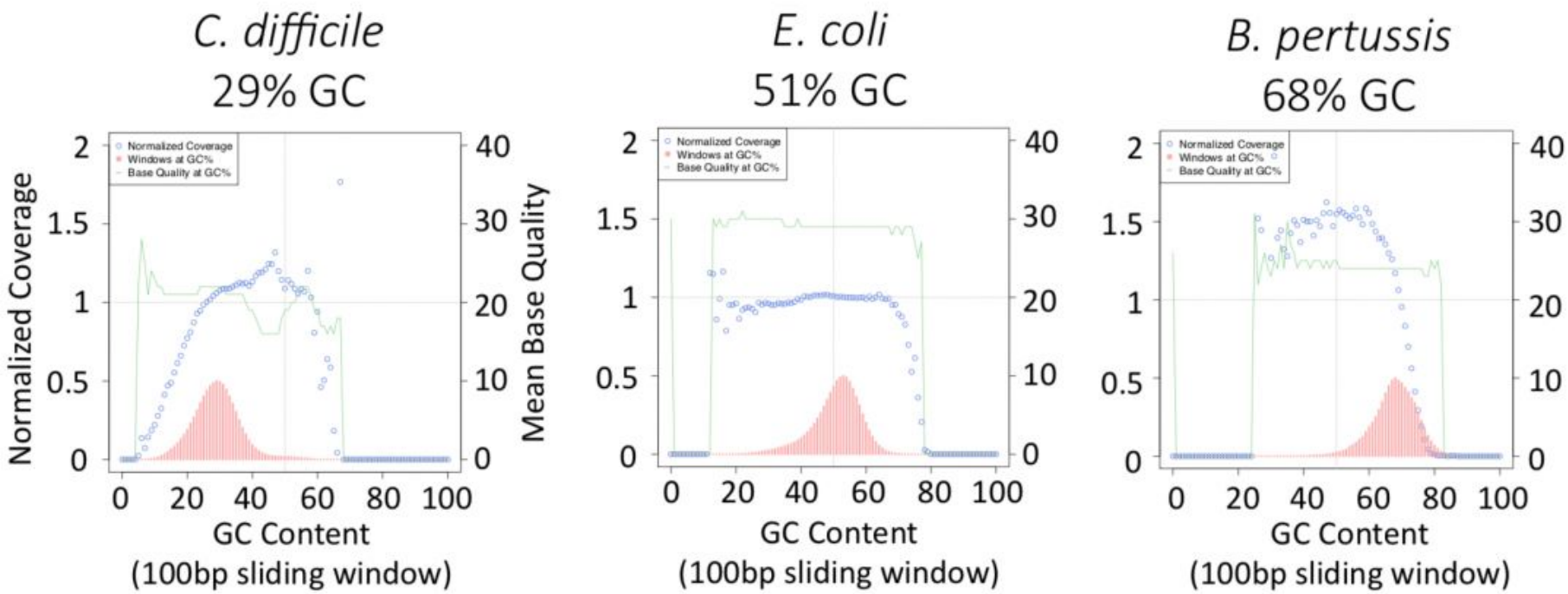
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0062856

Illumina sequencing - GC bias



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0062856

Minimal GC Content Bias



Quality control - fastqc

Download these files:
https://www.dropbox.com/s/5sebhvioluocea2/Test_R1.fastq?dl=0
https://www.dropbox.com/s/qh7tboc4je6hhtm/Test_R2.fastq?dl=0

Login into kamiak

Use UNIX to put them in your home directory!

Quality control - fastqc (GOOD)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Most Visited Getting Started

FastQC Report

Thu 21 Dec 2017
good_sequence_short.txt

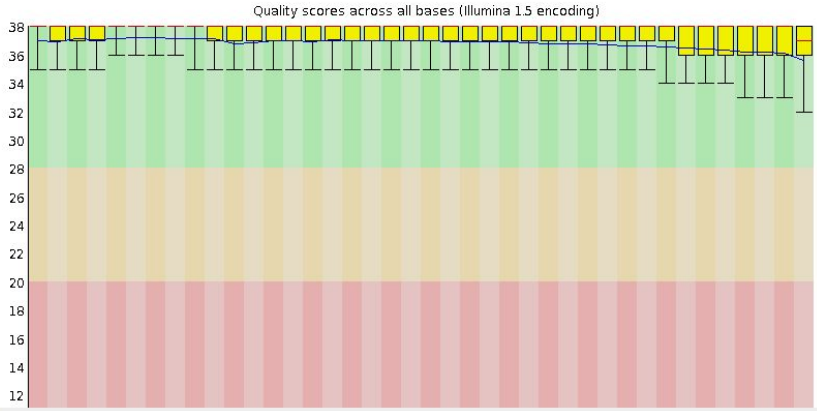
Summary

- Basic Statistics
- Per base sequence quality
- Per tile 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

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

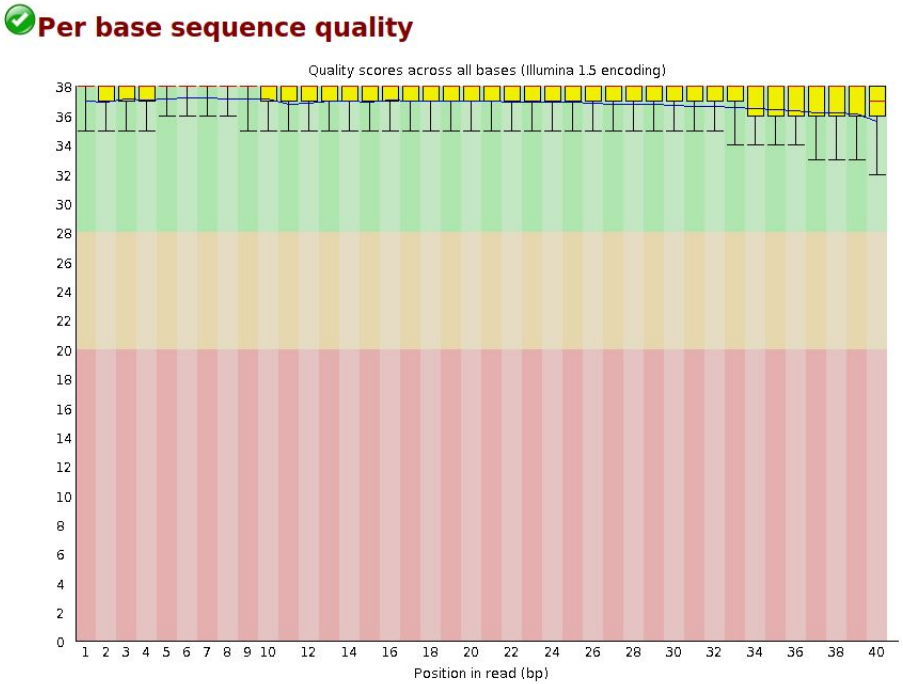
Per base sequence quality

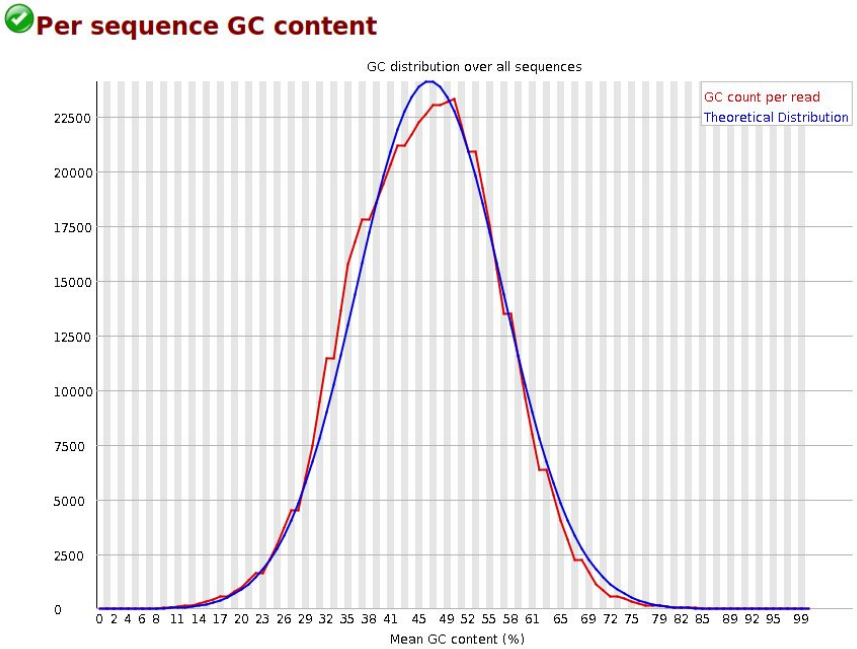


Produced by FastQC (version 0.11.6)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
How from UNIX???

Quality control - fastqc (GOOD)





Quality control - fastqc (BAD)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Most Visited Getting Started

FastQC Report

Thu 21 Dec 2017
bad_sequence.txt

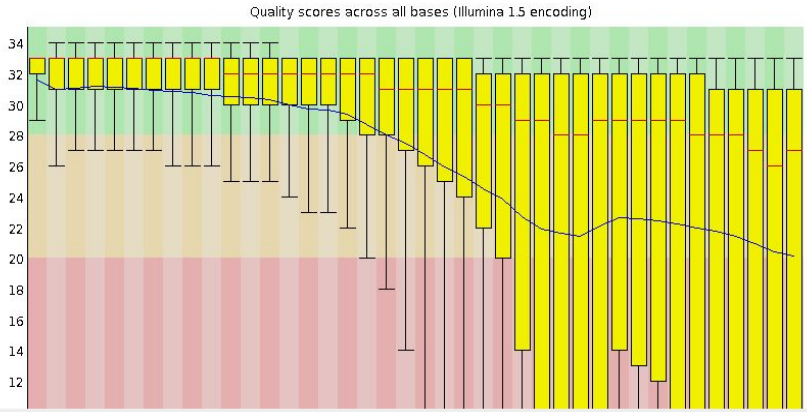
Summary

- Basic Statistics
- Per base sequence quality
- Per tile 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

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

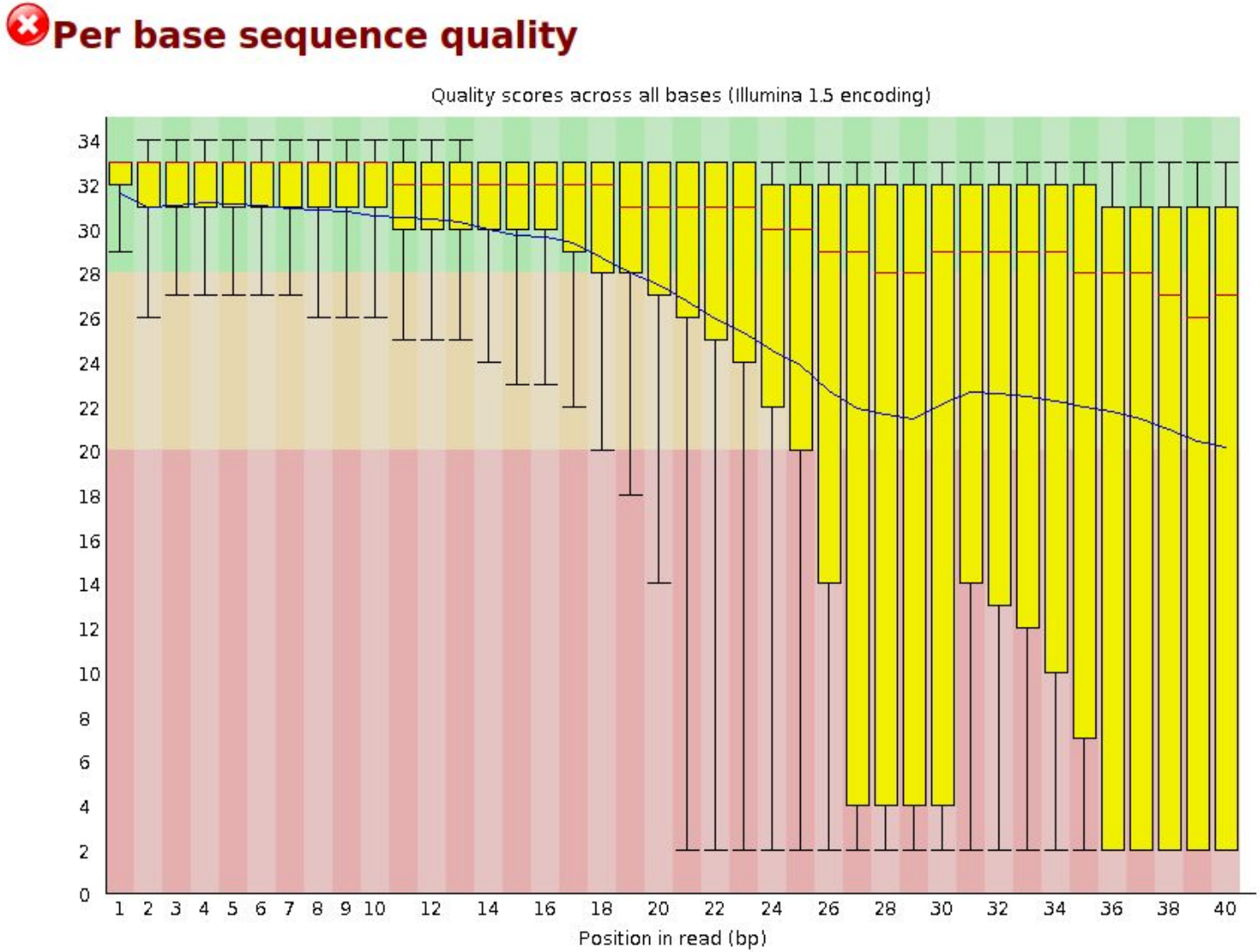
Per base sequence quality



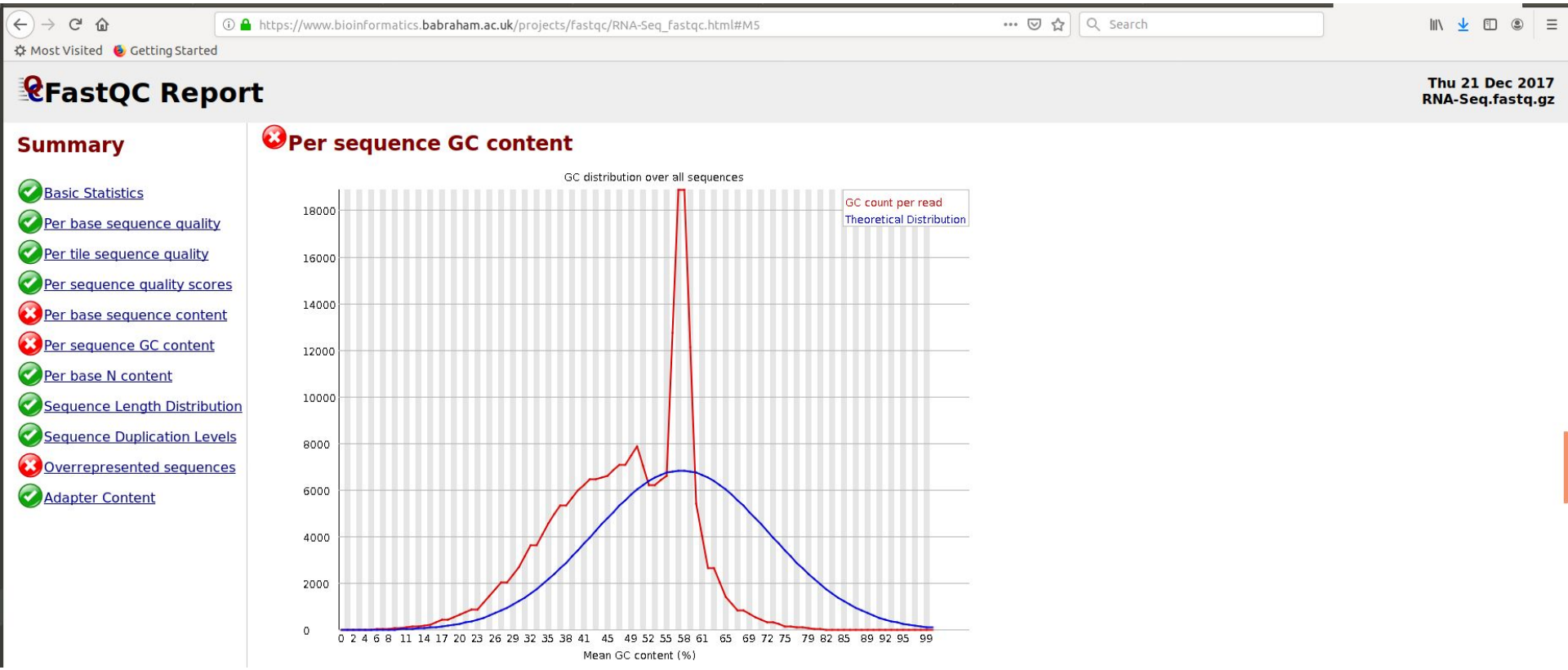
Produced by FastQC (version 0.11.6)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Quality control - fastqc (BAD)

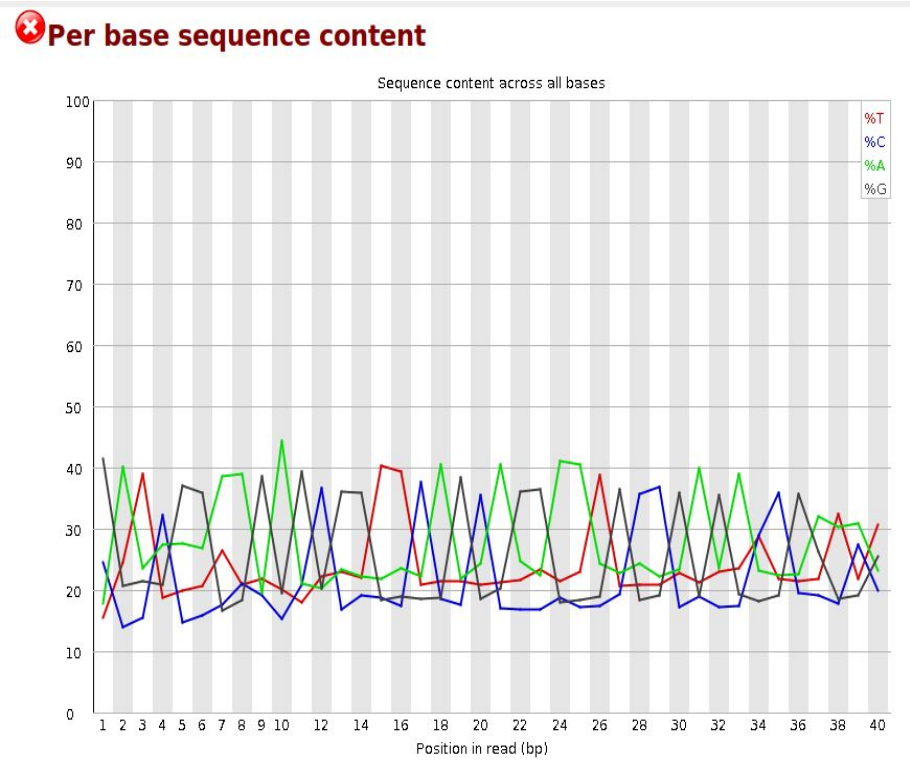


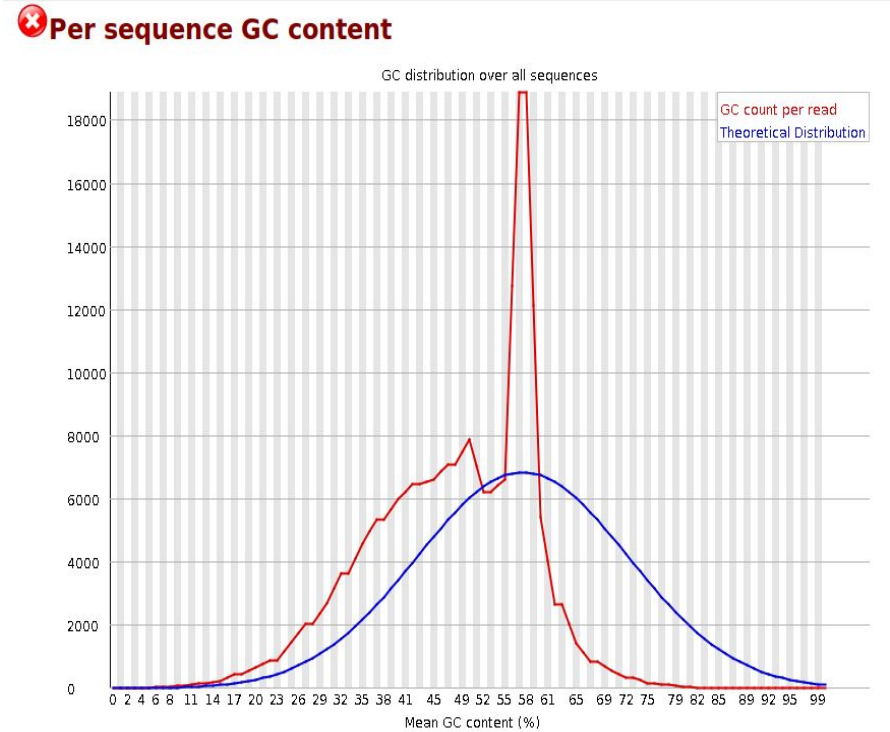
Quality control - fastqc (UGLY)



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/RNA-Seq_fastqc.html#M5

Quality control - fastqc (UGLY)





Quality control - fastqc (UGLY)

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.2350000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.2050000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Quality control - fastqc

Run these commands:

1) /home/richard.white3/FastQC/fastqc Test_R1.fastq

2) /home/richard.white3/FastQC/fastqc Test_R2.fastq

You should get two .zip or .html.

You can use firezilla to grab the htmls.

Open it in terminal with UNIX

https://www.dropbox.com/sh/9v7a9msstj7ffkw/AAD23VVhm6rV4PvRTTiqFDsEa?dl=0

Quality control - fastqc (raw)

file:///home/raw937/Desktop/Test_R1_fastqc.html

Most Visited Getting Started

FastQC Report

Mon 9 Sep 2019
Test_R1.fastq

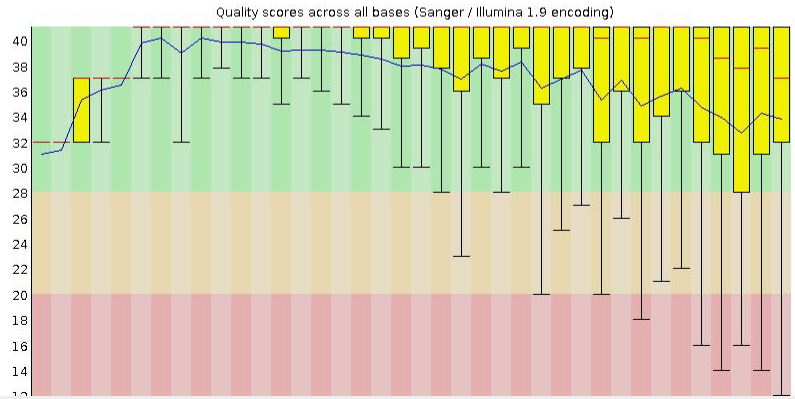
Summary

- Basic Statistics
- Per base sequence quality
- Per tile 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

Basic Statistics

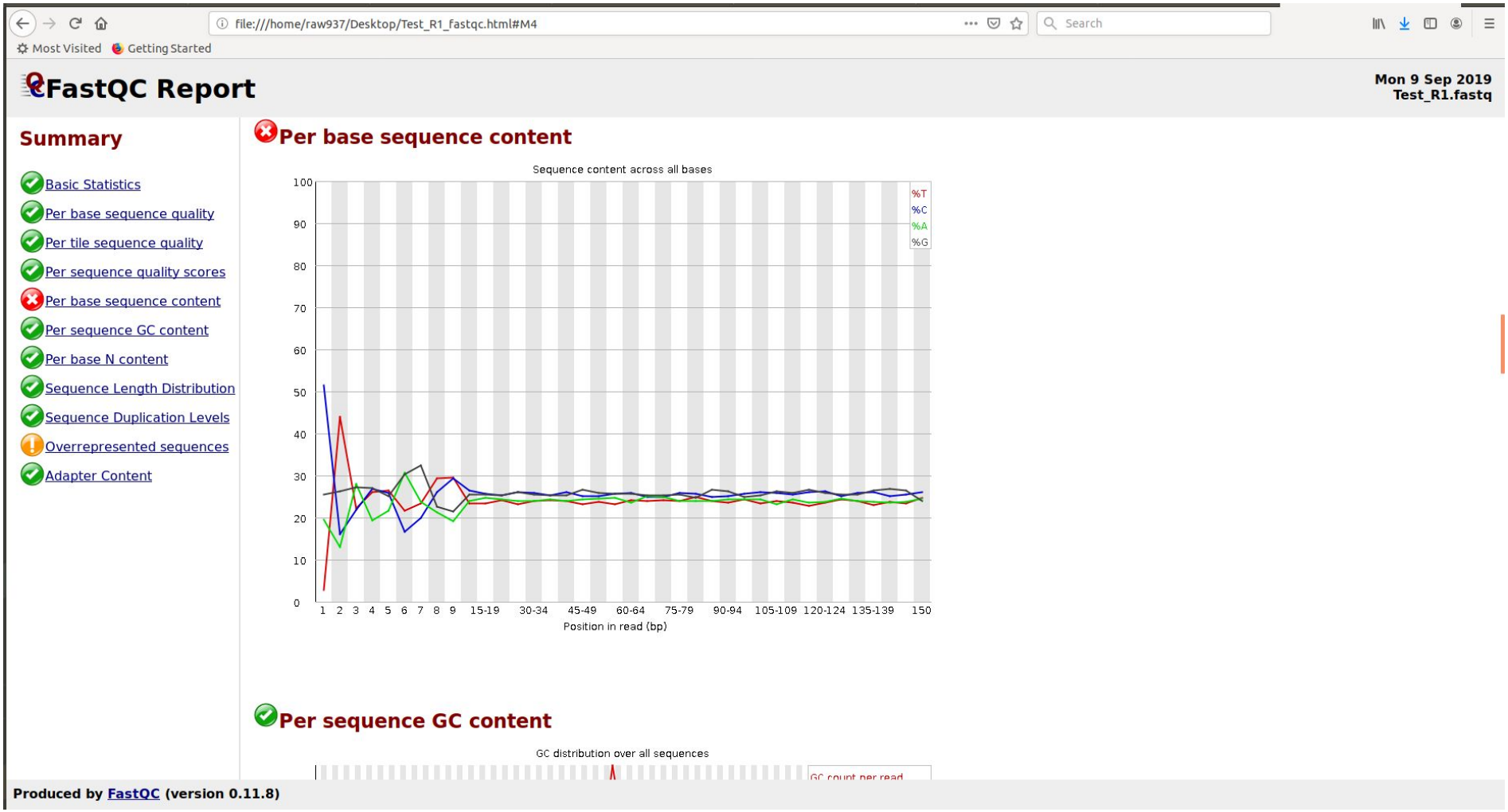
Measure	Value
Filename	Test_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2500
Sequences flagged as poor quality	0
Sequence length	150
%GC	51

Per base sequence quality



Produced by FastQC (version 0.11.8)

Quality control - fastqc (raw)



Quality control - fastqc (trim)

1) /home/richard.white3/bbmap/bbduk.sh -Xmx1g in1=Test_R1.fastq in2=Test_R2.fastq out1=Test_trim_R1.fastq out2=Test_trim_R2.fastq ref=~/bbmap/resources/adapters.fa ktrim=r k=21 mink=11 hdist=2 tpe tbo

2) /home/richard.white3/bbmap/fastqc Test_trim_R1.fastq

3) firefox Test_trim_R1.fastq.html (not from KAMAIK)

https://www.dropbox.com/sh/9v7a9msstj7ffkw/AAD23VVhm6rV4PvRTTiqFDsEa?dl=0

Quality control - fastqc (trim)

file:///home/raw937/Desktop/Test_trim_R1_fastqc.html

Most Visited Getting Started

FastQC Report

Mon 9 Sep 2019
Test_trim_R1.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per tile 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

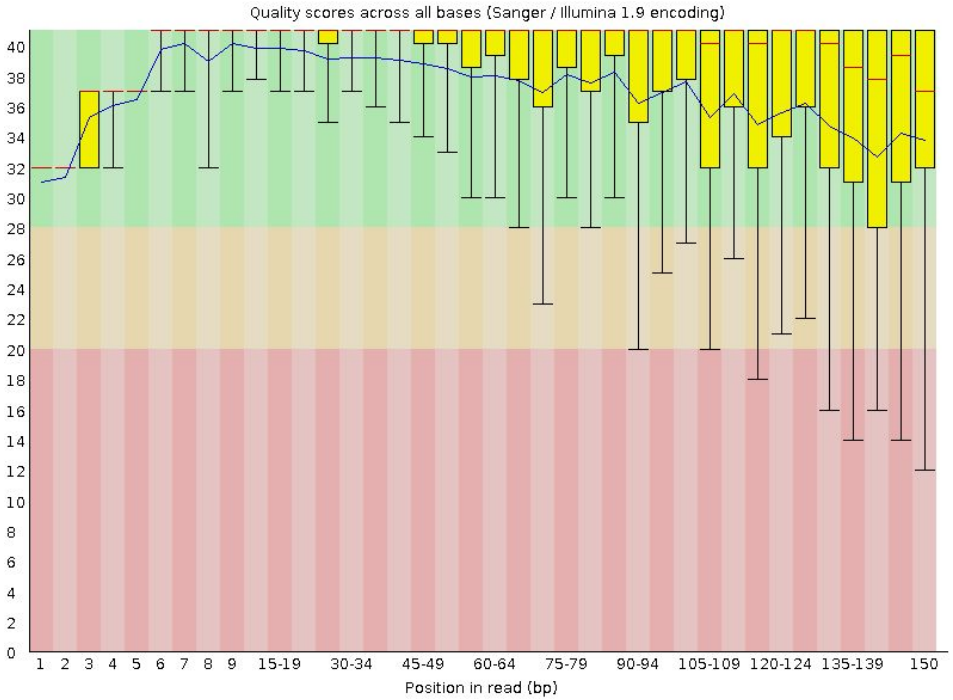
Basic Statistics

Measure	Value
Filename	Test_trim_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2499
Sequences flagged as poor quality	0
Sequence length	38-150
%GC	51

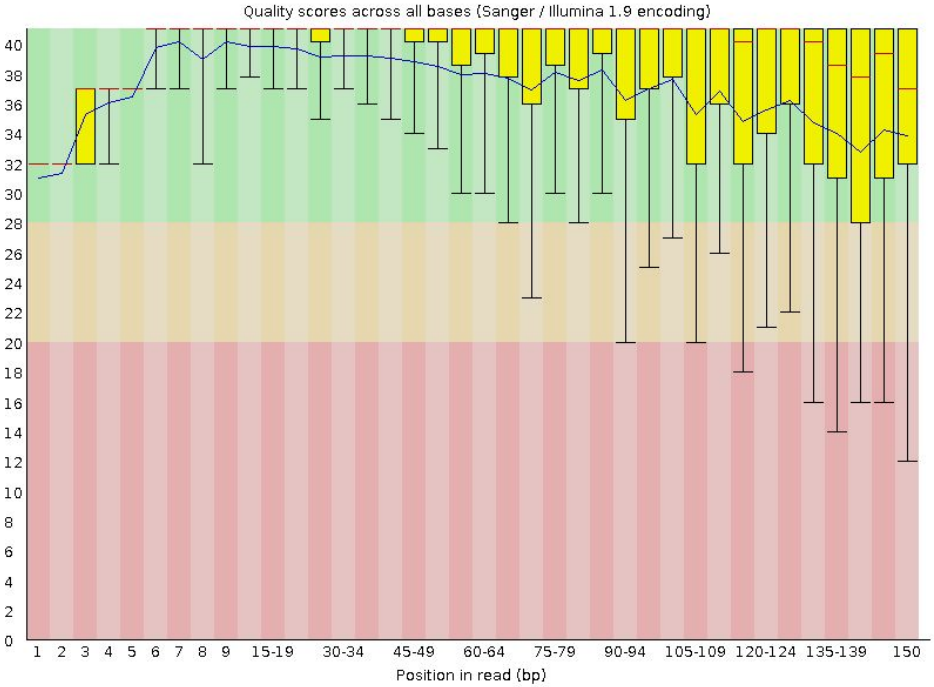
Per base sequence quality

Produced by FastQC (version 0.11.8)

Quality control - fastqc (trim)



raw



trim

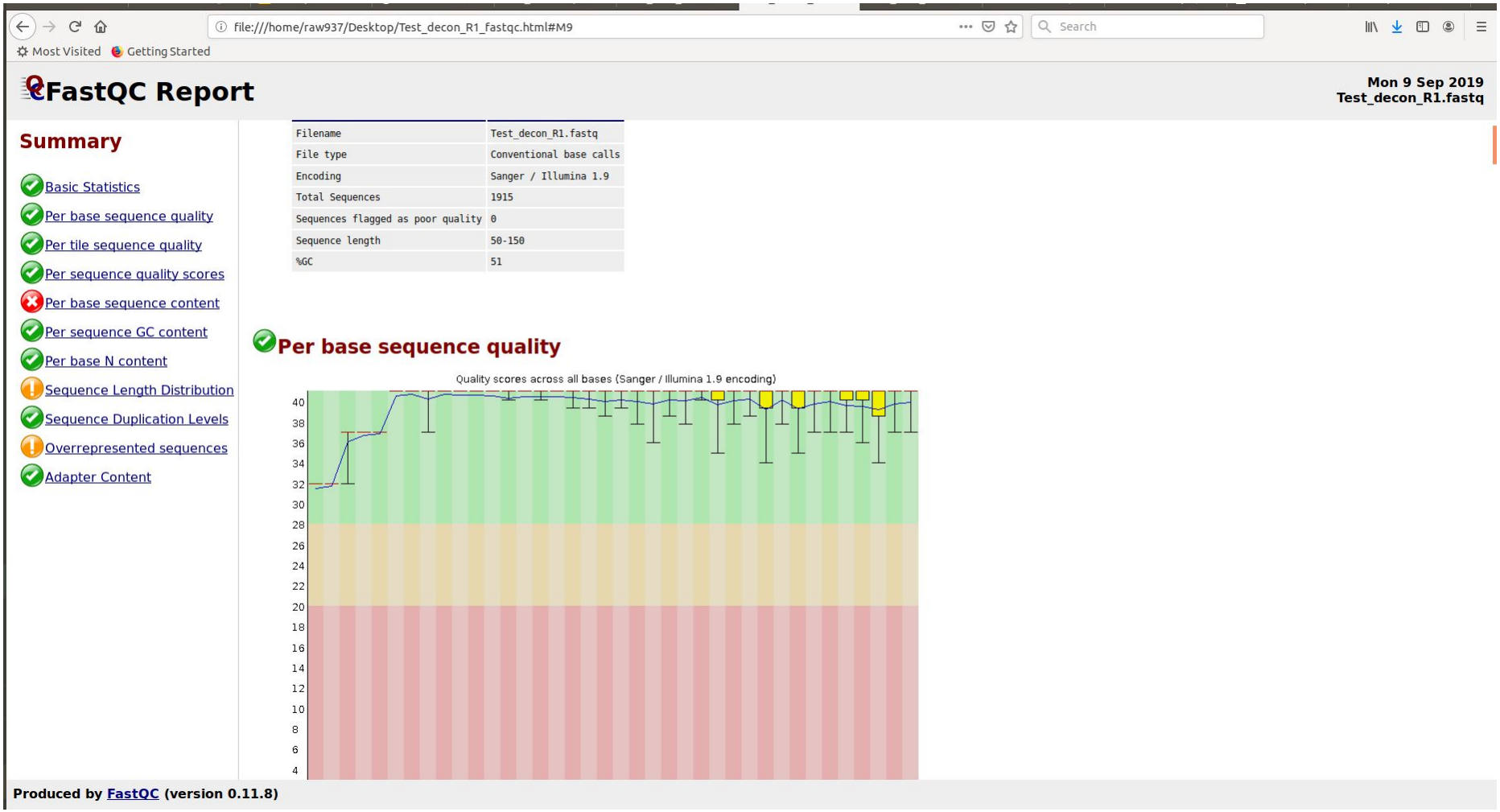
Quality control - fastqc (decon)

1) /home/richard.white3/bbmap/bbduk.sh in1=Test_trim_R1.fastq in2=Test_trim_R2.fastq out1=Test_decon_R1.fastq out2=Test_decon_R2.fastq qtrim=r trimq=25 maq=25 minlen=50 ref=~/bbmap/resources/phix174_ill.ref.fa.gz k=31 hdist=1 stats=decon_out.txt

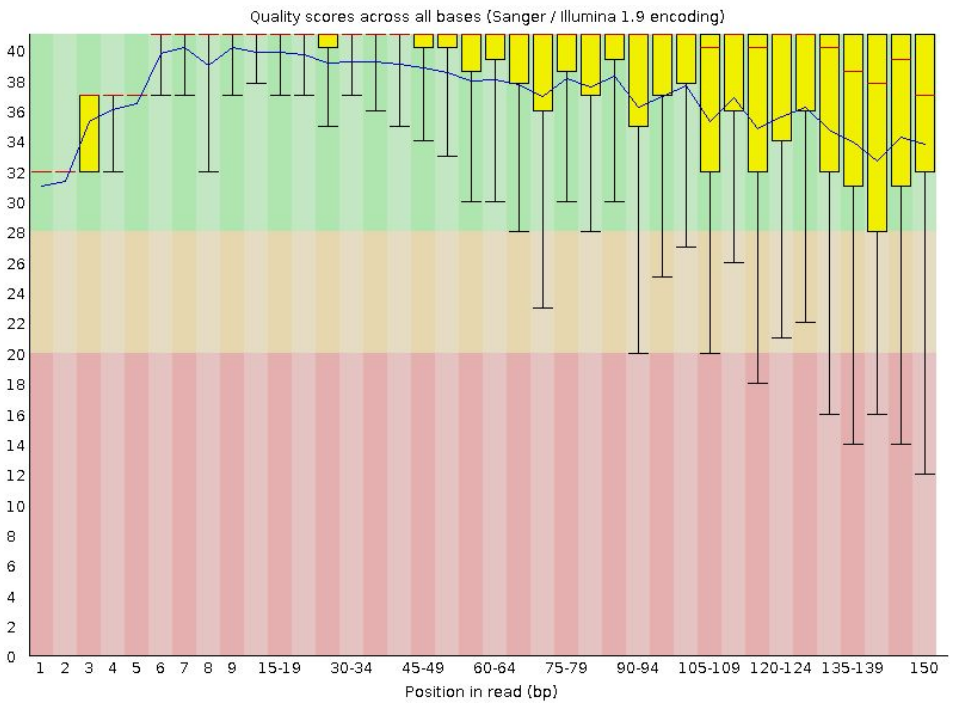
2) /home/richard.white3/bbmap/fastqc Test_decon_R1.fastq

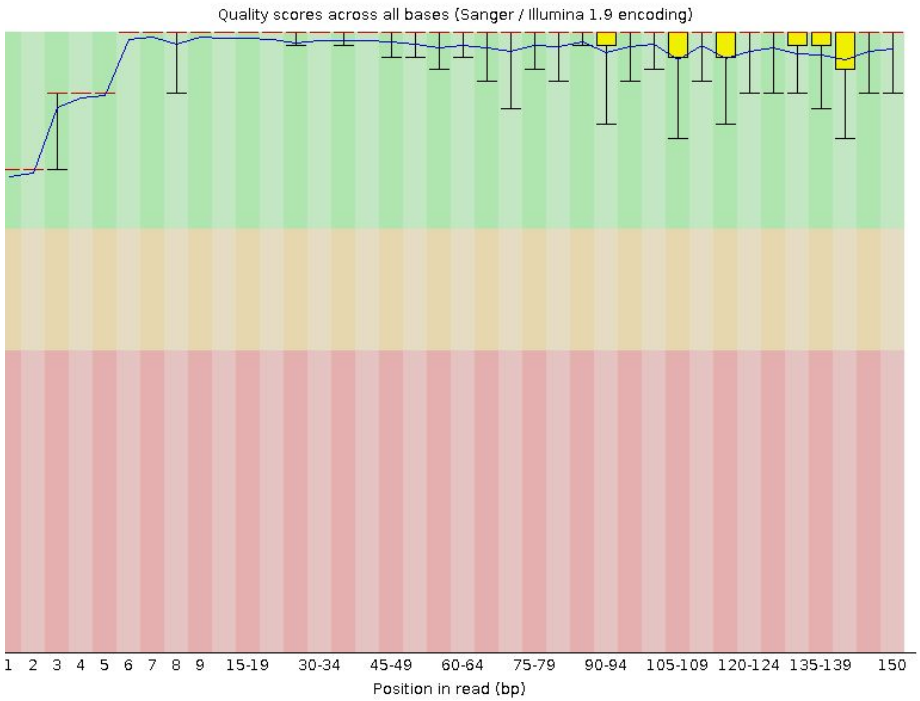
3) firefox Test_decon_R1.fastq.html (not from KAMAIK)

Quality control - fastqc (decon)



Quality control - fastqc (decon)





raw

decon

Quality control - check via mapping

1) module load bowtie2/2.3.4

2) /home/richard.white3/bowtie2 -x ~/bbmap/resources/phix -1 Test_decon_R1.fastq -2 Test_decon_R2.fastq --very-sensitive -S out.sam

Quality control - check via mapping

Results:

```
1915 reads; of these:
  1915 (100.00%) were paired; of these:
      1915 (100.00%) aligned concordantly 0 times
      0 (0.00%) aligned concordantly exactly 1 time
      0 (0.00%) aligned concordantly >1 times
      ----
      1915 pairs aligned concordantly 0 times; of these:
      0 (0.00%) aligned discordantly 1 time
      ----
      1915 pairs aligned 0 times concordantly or discordantly; of these:
      1915 mates make up the pairs; of these:
      1915 (100.00%) aligned 0 times
      0 (0.00%) aligned exactly 1 time
      0 (0.00%) aligned >1 times
0.00% overall alignment rate
```