Informatics on High-throughput Sequencing Data

(Summer Course 2020)

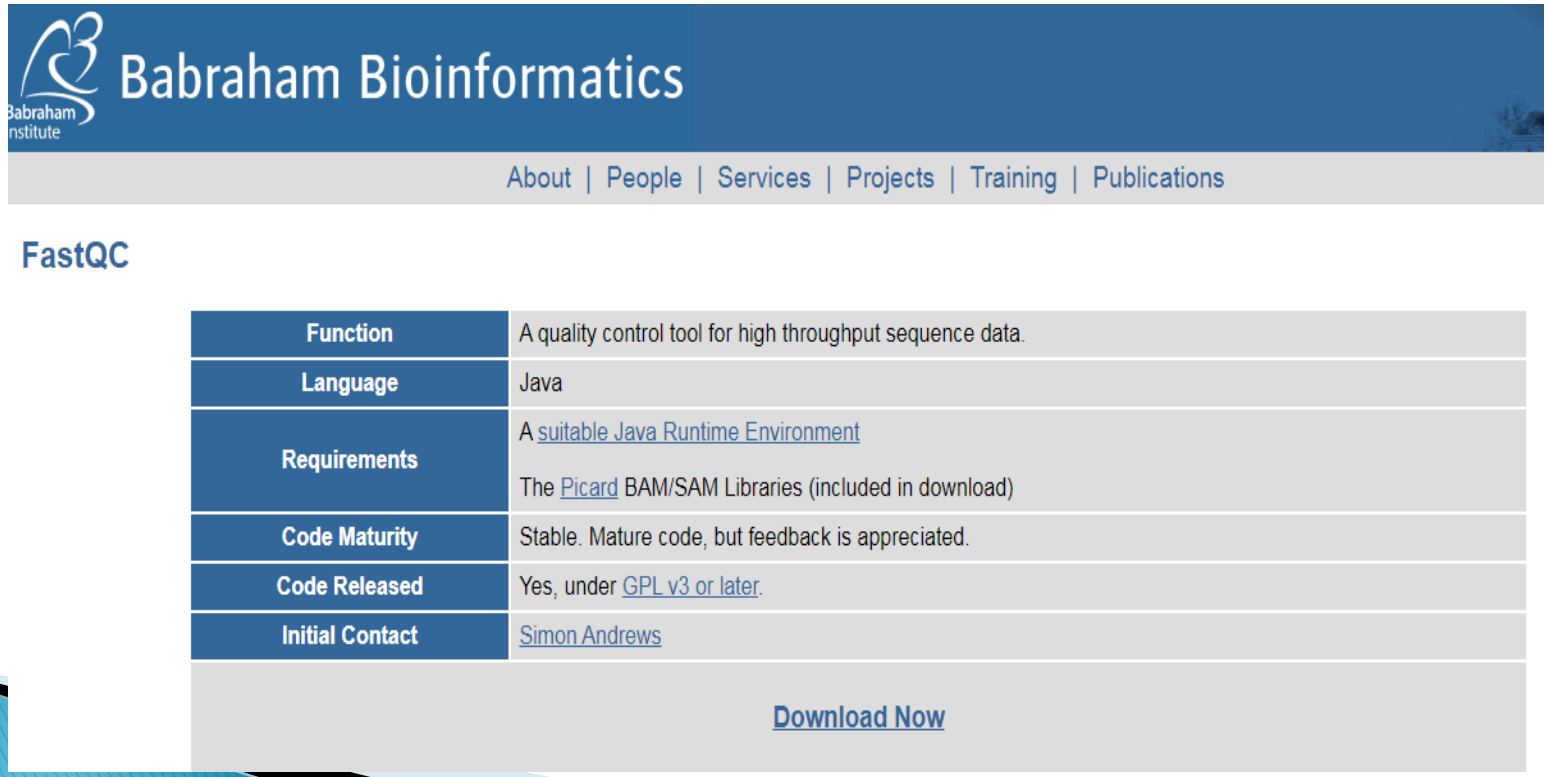
Day 11

Note (SRA toolkit)

```
mkdir ~/glibc_install
cd ~/glibc_install

wget http://ftp.gnu.org/gnu/glibc/glibc-2.14.tar.gz

tar zxvf glibc-2.14.tar.gz

cd glibc-2.14

mkdir build

cd build

../configure --prefix=/opt/glibc-2.14

make -j4

sudo make install

export LD_LIBRARY_PATH="/opt/glibc-2.14/lib${LD_LIBRARY_PATH:+:$LD_LIBRARY_PATH}"
sudo chmod 777 /etc/environment
vi /etc/environment
export LD_LIBRARY_PATH="/opt/glibc-2.14/lib${LD_LIBRARY_PATH:+:$LD_LIBRARY_PATH}"
source /etc/environment
```

FastQC

- **FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.**



Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

FastQC

github.com/s-andrews/FastQC

Search or jump to... / Pull requests Issues Marketplace Explore

s-andrews / FastQC

Code Issues 18 Pull request

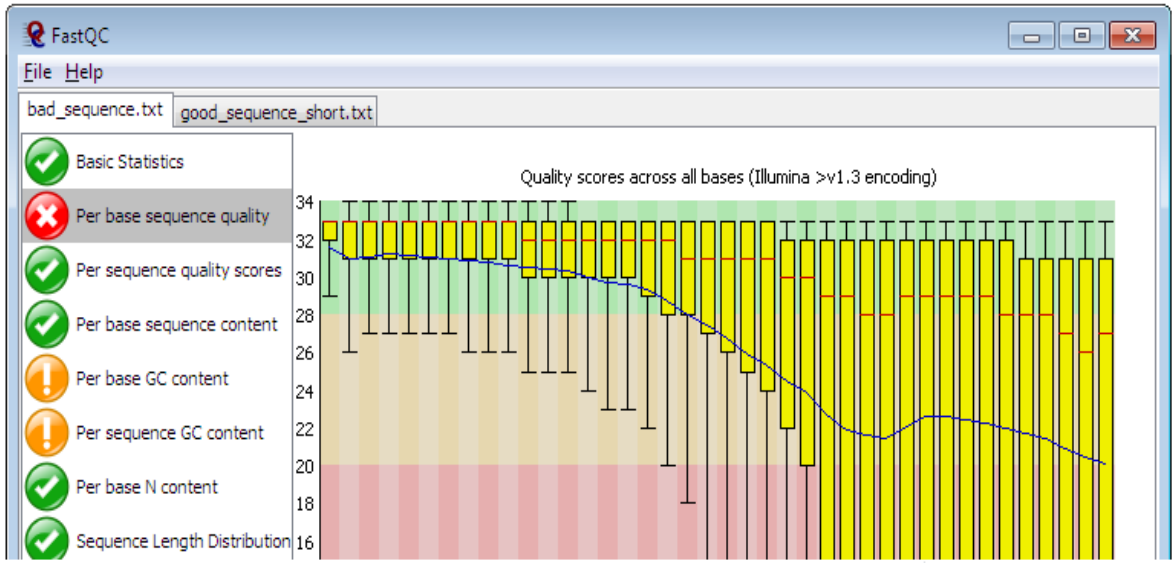
master 3 branches 4 tag

s-andrews Merge pull request #28 fror

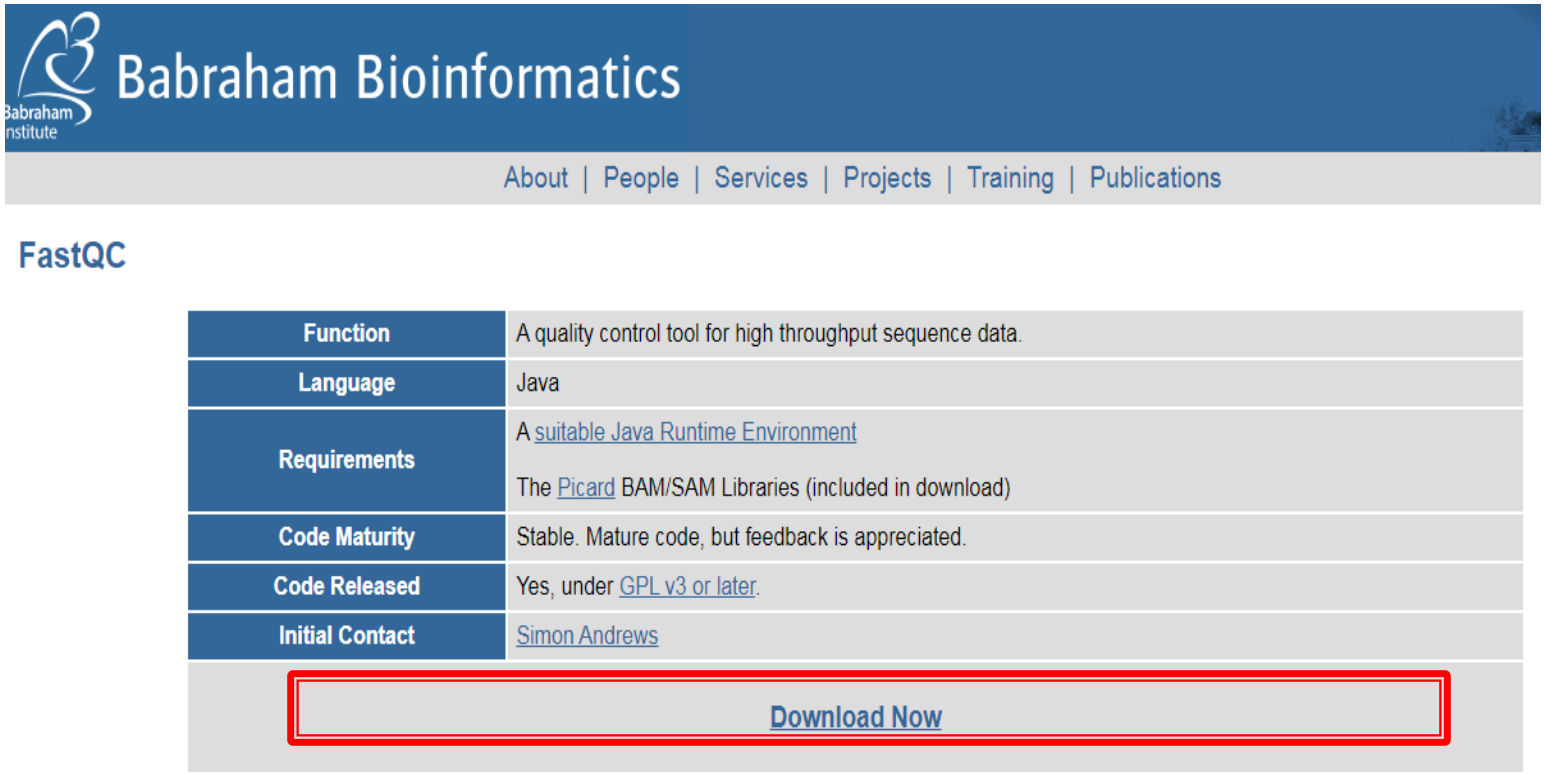
- .settings
- Configuration
- Help
- Templates
- net/sourceforge/iharder/base64
- org/apache/commons/math3
- uk/ac/babraham/FastQC

FastQC

FastQC is a program designed to spot potential problems in high througput sequencing datasets. It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.



FastQC


Babraham Institute
Babraham Bioinformatics
About | People | Services | Projects | Training | Publications
FastQC
Function
A quality control tool for high throughput sequence data.
Language
Java
Requirements
A suitable Java Runtime Environment
The Picard BAM/SAM Libraries (included in download)
Code Maturity
Stable. Mature code, but feedback is appreciated.
Code Released
Yes, under GPL v3 or later.
Initial Contact
Simon Andrews
Download Now

FastQC

FastQC A quality control application for high throughput sequence data

- README
- Installation and setup instructions
- Release Notes Please read these before using the program.
- FastQC v0.11.9 (Win/Linux zip file)
- FastQC v0.11.9 (Mac DMG image)
- Source Code for the latest FastQC release

```
Ubuntu / Mint: sudo apt install default-jre

CentOS / Redhat: sudo yum install java-1.8.0-openjdk

You can check whether java is installed by opening the 'cmd' program on windows, or
any shell on linux and typing:

java -version

You should see something like:

>java -version
openjdk version "11.0.2" 2019-01-15
OpenJDK Runtime Environment AdoptOpenJDK (build 11.0.2+9)
OpenJDK 64-Bit Server VM AdoptOpenJDK (build 11.0.2+9, mixed mode)
```

FastQC

FastQC A quality control application for high throughput sequence data

- README
- Installation and setup instructions
- Release Notes Please read these before using the program.
- FastQC v0.11.9 (Win/Linux zip file)
- FastQC v0.11.9 (Mac DMG image)
- Source Code for the latest FastQC release

```
Running FastQC Interactively
----------------------------
Windows: Simply double click on the run_fastqc bat file.  If you want to make a pretty
shortcut then we've included an icon file in the top level directory so you don't have
to use the generic bat file icon.

MacOSX: Double click on the FastQC application icon.

Linux:  We have included a wrapper script, called 'fastqc' which is the easiest way to
start the program.  The wrapper is in the top level of the FastQC installation.  You
may need to make this file executable:

chmod 755 fastqc

..but once you have done that you can run it directly

./fastqc
```

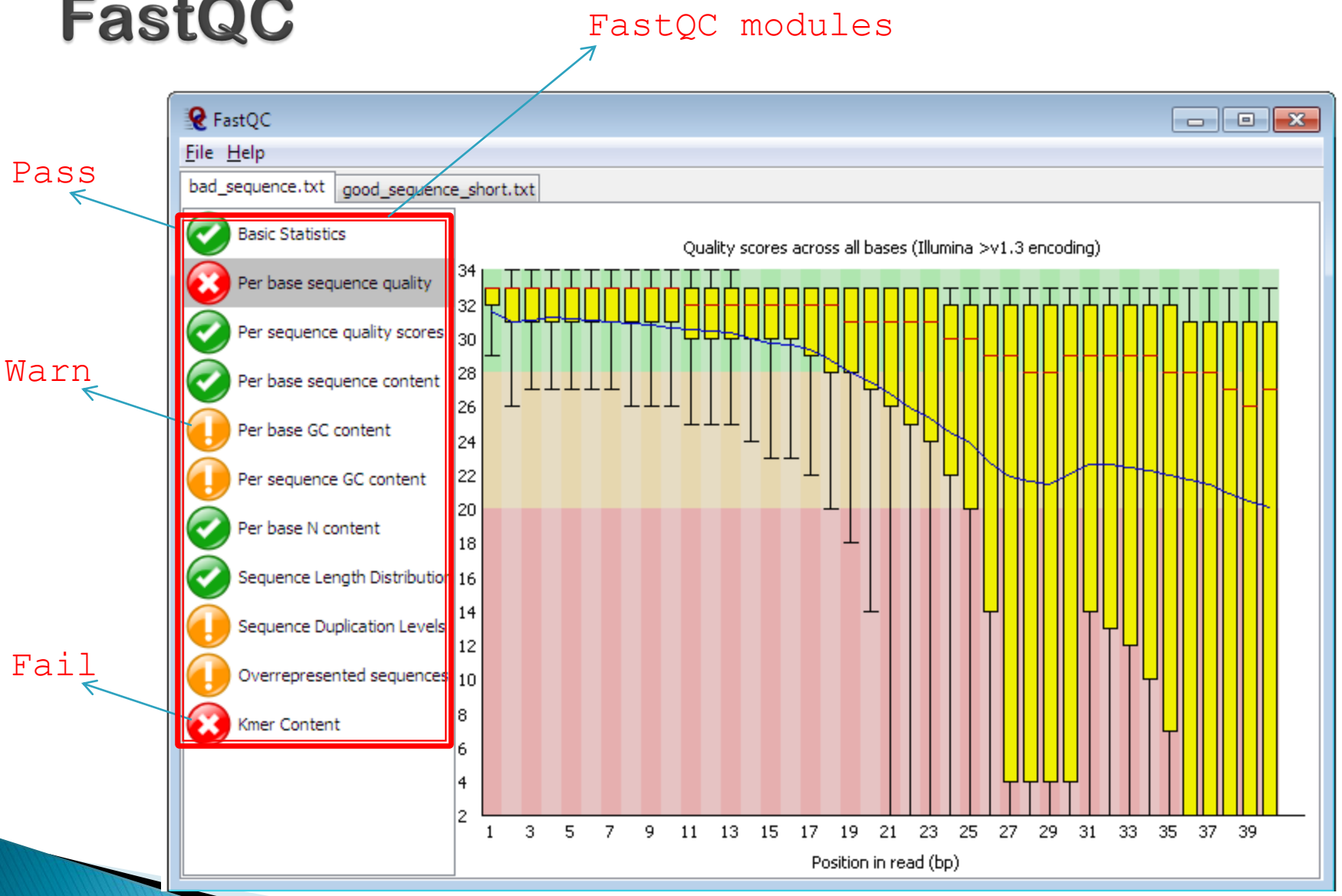
FastQC

FastQC A quality control application for high throughput sequence data

- README
- Installation and setup instructions
- Release Notes Please read these before using the program
- FastQC v0.11.9 (Win/Linux zip file)
- FastQC v0.11.9 (Mac DMG image)
- Source Code for the latest FastQC release

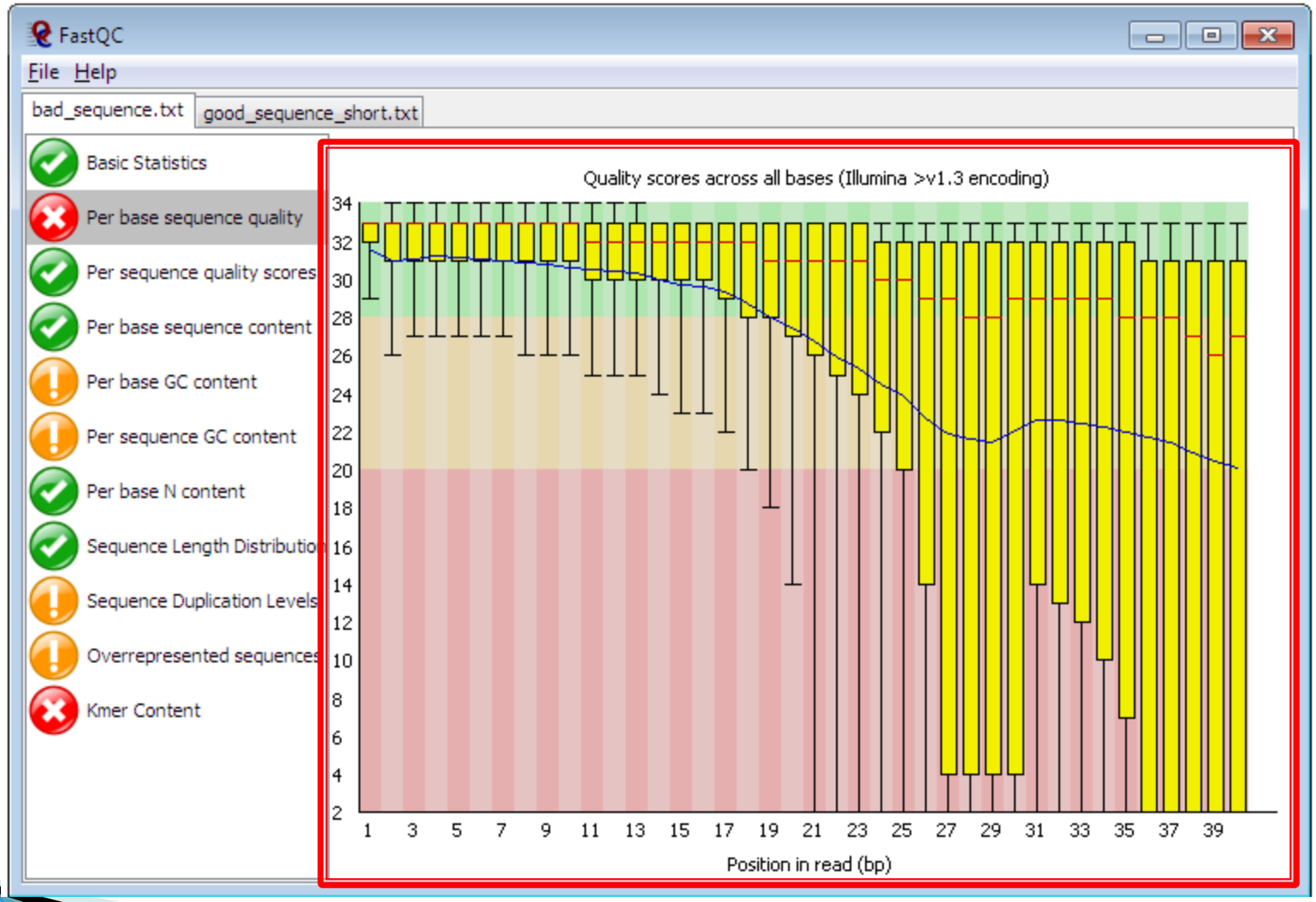
- `wget download link fastqc`
- `chmod 755 fastqc`
- `./fastqc`

FastQC


FastQC modules
Pass
Warn
Fail
FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
Position in read (bp)

FastQC

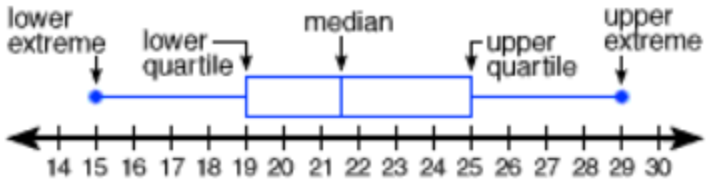
Box and whisker plot



FastQC

Example 1 – Box and whisker plots

51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.



Box and whisker plot

Answers

1. First, put the data in ascending order. Then find the median.

   6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.
   Median = (12th + 1st) ÷ 2 = 6.5th value
   = (sixth + seventh observations) ÷ 2
   = (25 + 39) ÷ 2
   = **32**

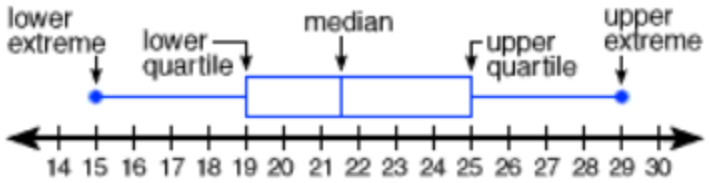
   There are six numbers below the median, namely: 6, 7, 13, 17, 20, 25.
   Q1 = the median of these six items
   = (6 + 1) ÷ 2= 3.5^{th} value
   = (third + fourth observations) ÷ 2
   = (13 + 17) ÷ 2
   = **15**

   Here are six numbers above the median, namely: 39, 41, 43, 49, 51, 62.
   Q_3 = the median of these six items
   = (6 + 1) ÷ 2= 3.5^{th} value
   = (third + fourth observations) ÷ 2
   = **46**

   The five-number summary 6, 15, 32, 46, 62.

https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214889-eng.htm#:~:text=A%20box%20and%20whisker%20plot%20is%20a%20way%20of%20summarizing,used%20in%20explanatory%20data%20analysis.&text=In%20a%20box%20and%20whisker,vertical%20line%20inside%20the%20box

FastQC



Box and whisker plot

11, 10, 12, 23, 17, 16, 17, 14, 24, 22, 14

sorted:

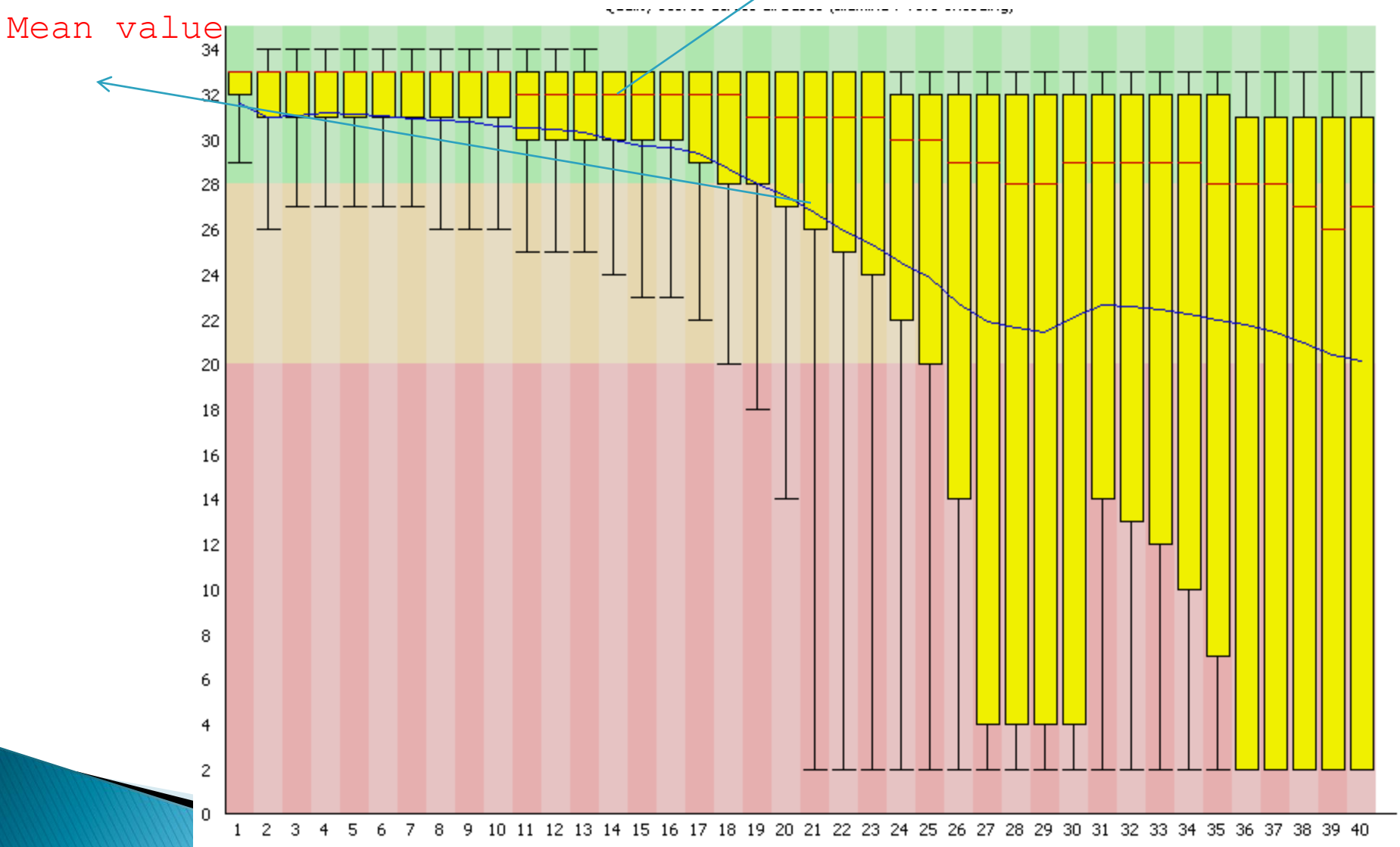
10, 11, 12, 14, 14, 16, 17, 17, 22, 23, 24

resulting:

- 10th percentile: **11**
- 50th percentile: **16**
- 90th percentile: **23**

0.9*11=9.9=[10]=23.
0.1*11=1.1=[2]=11.

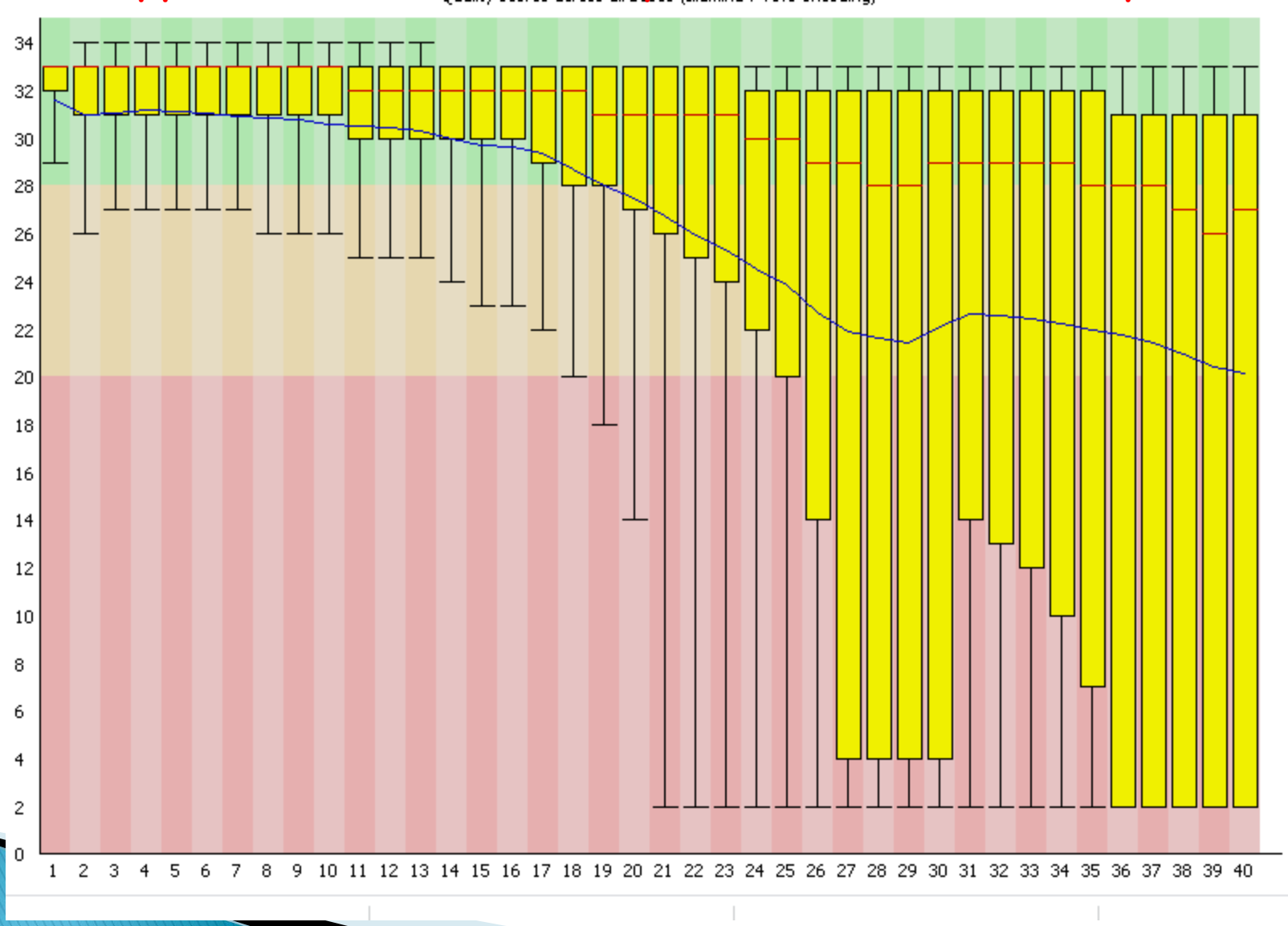
FastQC



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

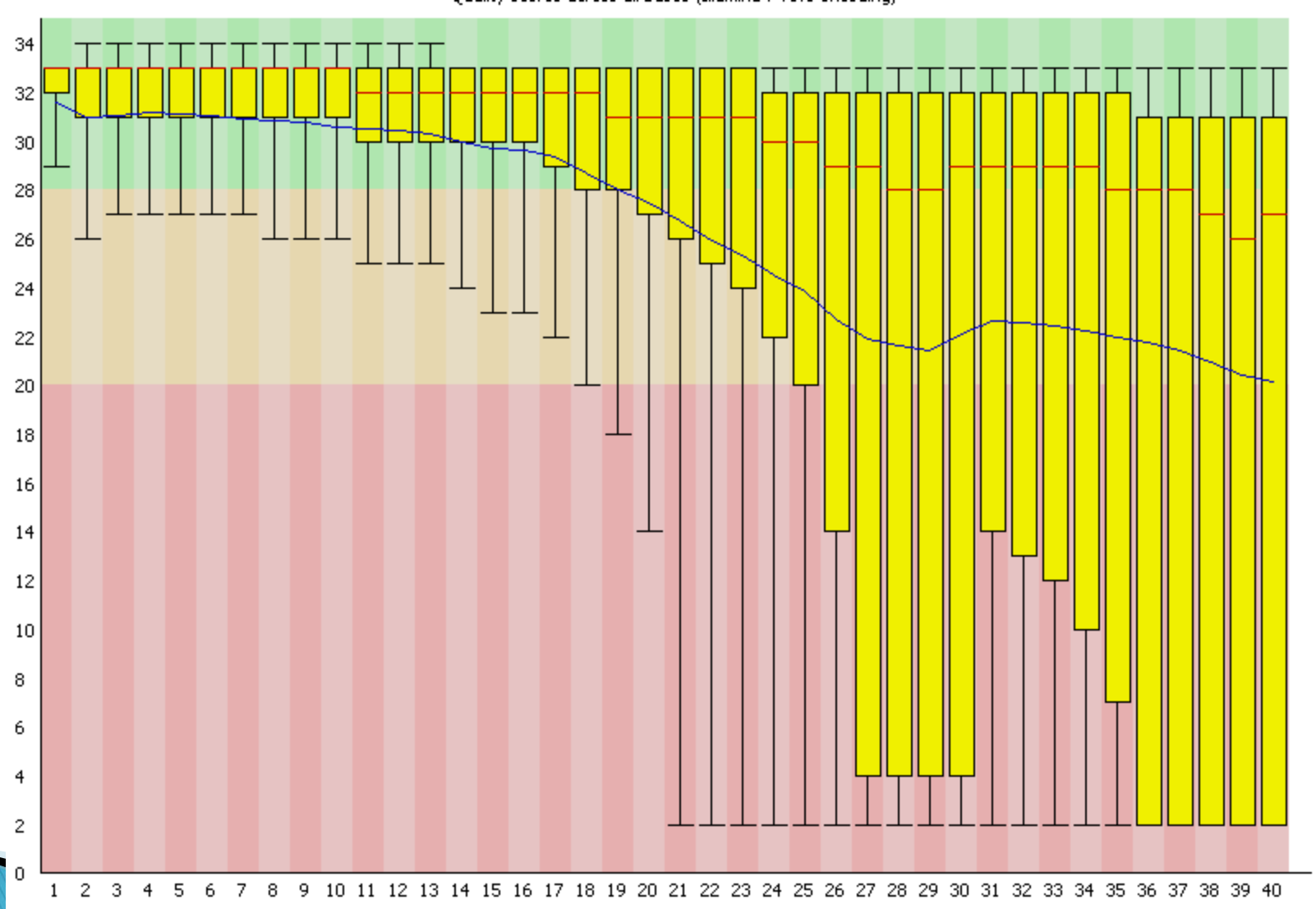
FastQC

The upper and lower whiskers represent the 10% and 90% points



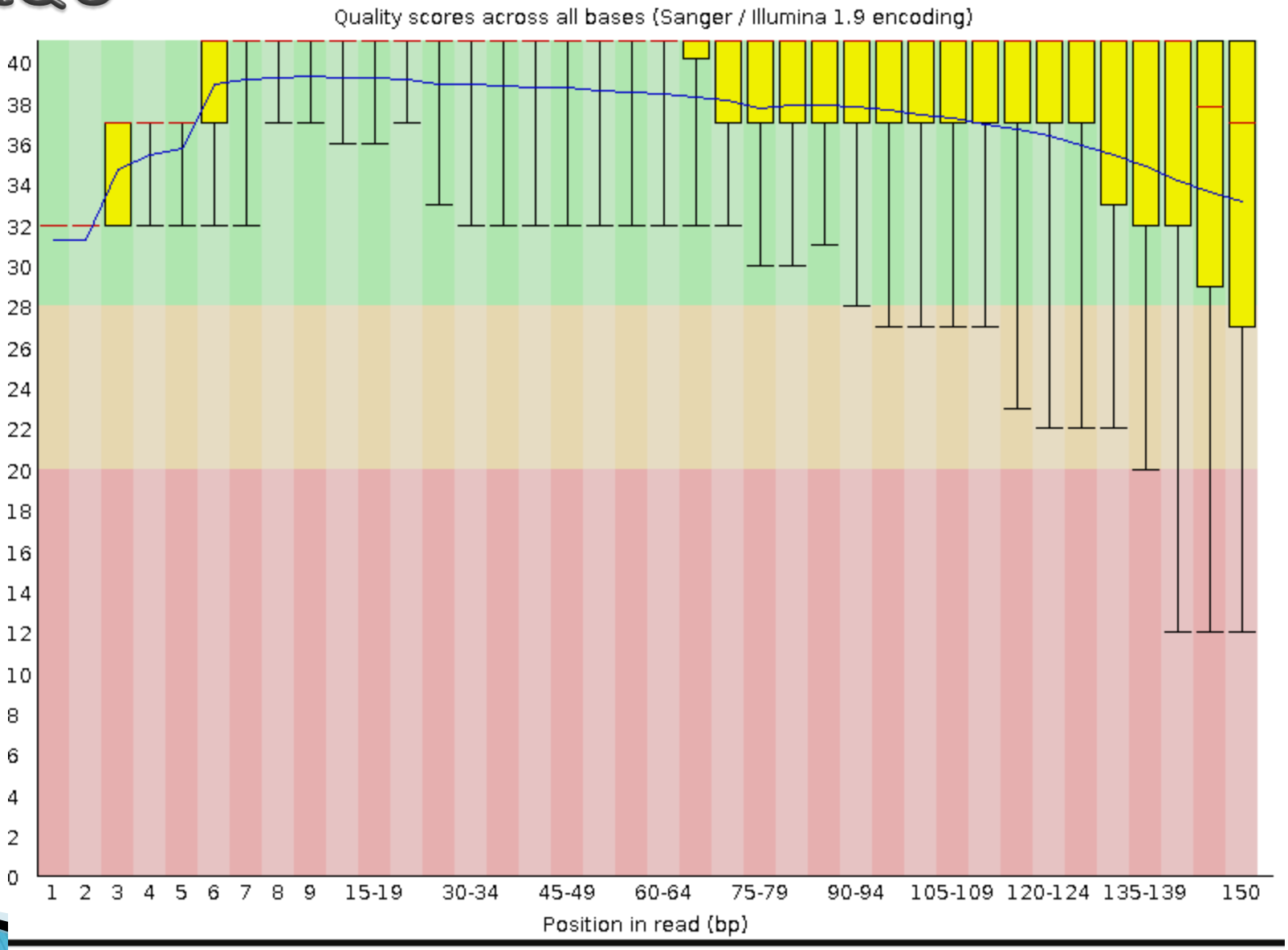
FastQC

The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

FastQC



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

Notes

- It is normal with all Illumina sequencers for the median quality score to start out lower over the first 5-7 bases and to then rise.
- The average quality score will steadily drop over the length of the read.
- With paired end reads the average quality scores for read 1 will almost always be higher than for read 2.

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

Thanks!

// | ?