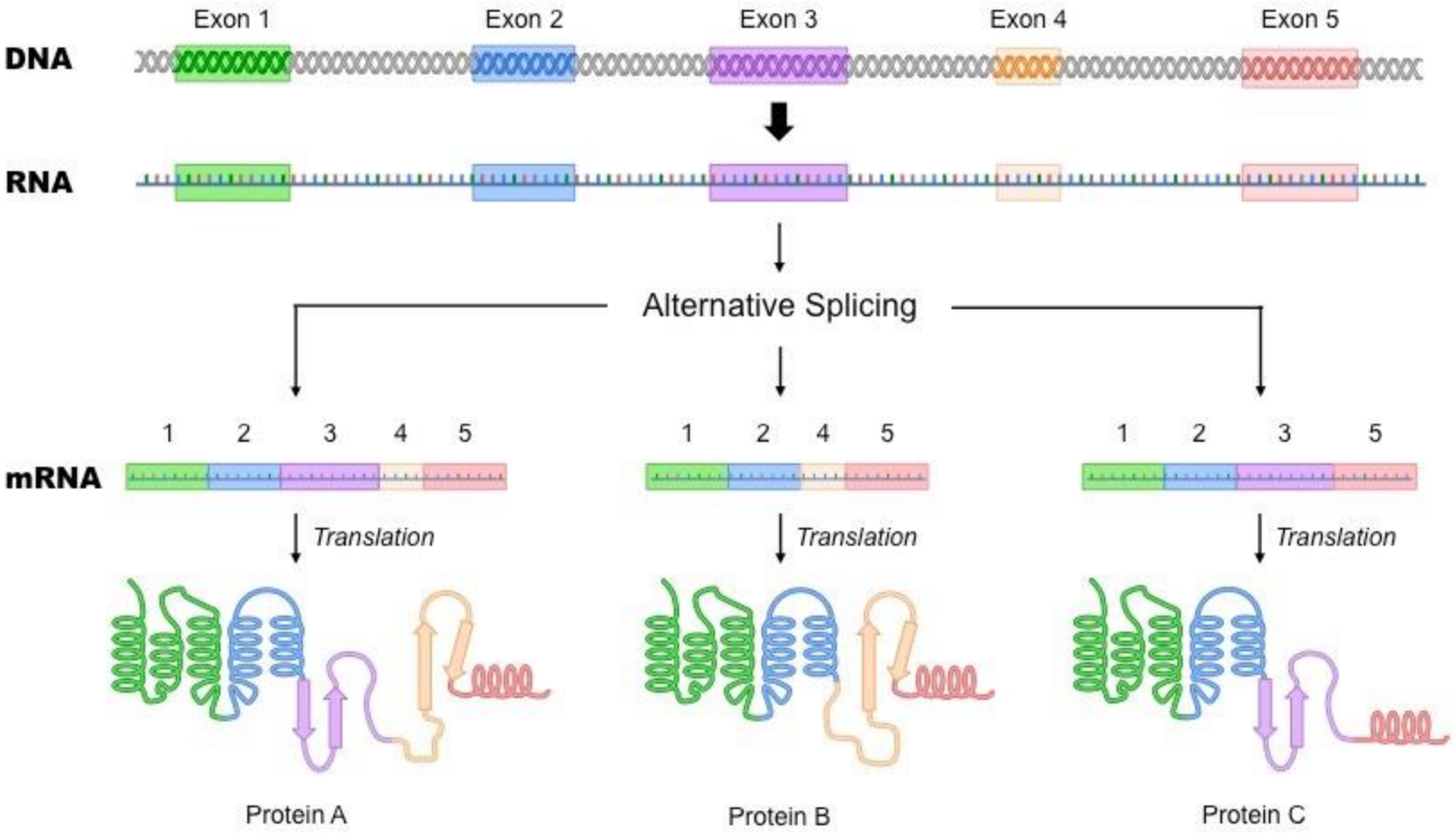
Informatics on High-throughput Sequencing Data

(Summer Course 2020)

Day 10

Alternative Splicing



SRA

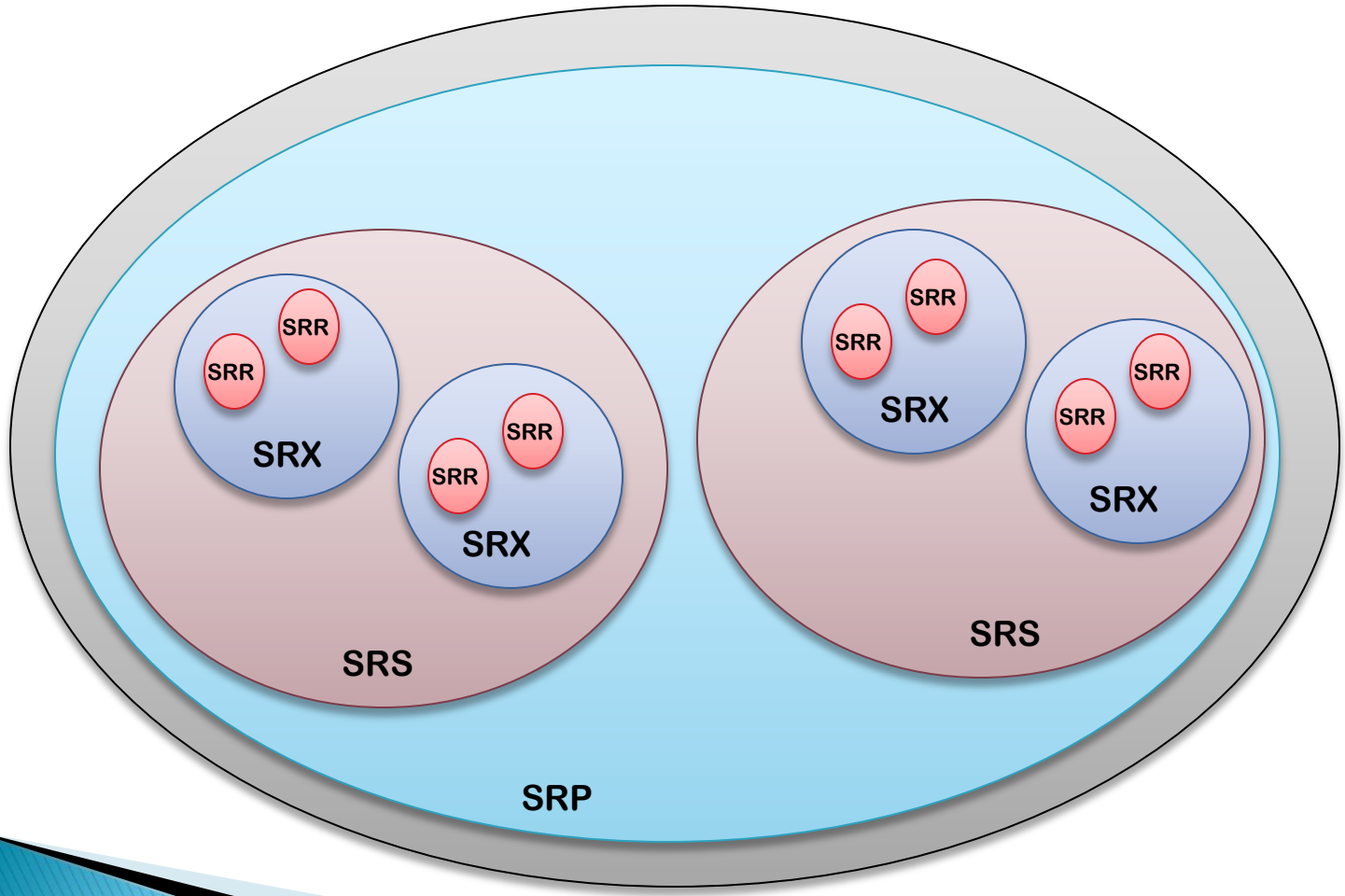
- **SRA is the largest publicly available repository of high throughput sequencing data .**

Metadata	Description
Study (SRP)	A study is a set of experiments and has an overall goal.
Experiment (SRX)	An experiment is a consistent set of laboratory operations on input material with an expected result.
Sample (SRS)	An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
Run (SRR)	Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.
Submission	A submission is a package of metadata and/or data objects and a directive for what to do with those objects.

A Study (SRP) has one or more samples; a sample (SRS) has one or more experiments (SRX); an experiment has one or more runs (SRR).

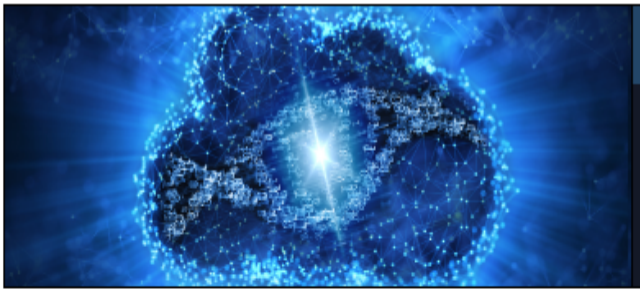
https://linsalrob.github.io/ComputationalGenomicsManual/Databases/SRA.html

SRA



SRA submission Accession

SRA



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Announcement

NIH Request for Information (RFI) on SRA data format changes and plans.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

SRA

NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

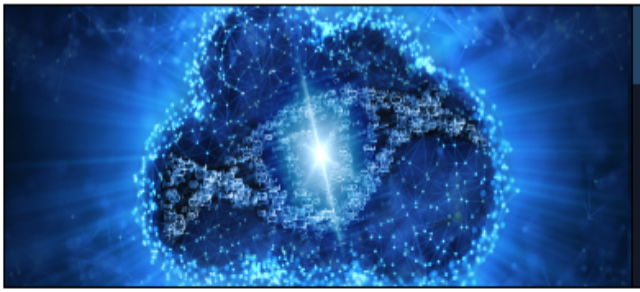
SRA Toolkit

Compiled binaries/install scripts of June 29, 2020, version 2.10.8:

- CentOS Linux 64 bit architecture - non-sudo tar archive
- Ubuntu Linux 64 bit architecture - non-sudo tar archive
- Cloud - apt-get install script - for Debian and Ubuntu - requires sudo permissions
- Cloud - yum install script - for for CentOS - requires sudo permissions
- MacOS 64 bit architecture
- MS Windows 64 bit architecture
- md5 checksums

- `sudo apt-get install sra-toolkit`
- `wget download link`
- `tar -xzvf`
- `cd bin`
- `./vdb-config -i`

SRA



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Announcement

NIH Request for Information (RFI) on SRA data format changes and plans.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

SRA

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide

Protected Data Usage Guide

Frequently Used Tools:

fastq-dump: Convert SRA data into fastq format

prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data

sam-dump: Convert SRA data to sam format

sra-pileup: Generate pileup statistics on aligned SRA data

vdb-config: Display and modify VDB configuration information

vdb-decrypt: Decrypt non-SRA dbGaP data ("phenotype data")

- `./prefetch  SRX1074313`

SRA

Open Access | Published: 01 September 2015

High-resolution analysis of the human T-cell receptor repertoire

Eliana Ruggiero, Jan P. Nicolay, Raffaele Fronza, Anne Arens, Anna Paruzynski, Ali Nowrouzi, Gökçe Ürenden, Christina Lulay, Sven Schneider, Sergij Goerdt, Hanno Glimm, Peter H. Krammer, Manfred Schmidt & Christof von Kalle

Nature Communications **6**, Article number: 8081 (2015) | Cite this article

Additional information

Accession codes: The TCR-sequencing data generated in this paper have been deposited in the SRA database under the accession code SRP059581.

How to cite this article: Ruggiero, E. *et al.* High-resolution analysis of the human T-cell receptor repertoire. *Nat. Commun.* 6:8081 doi: 10.1038/ncomms9081 (2015).

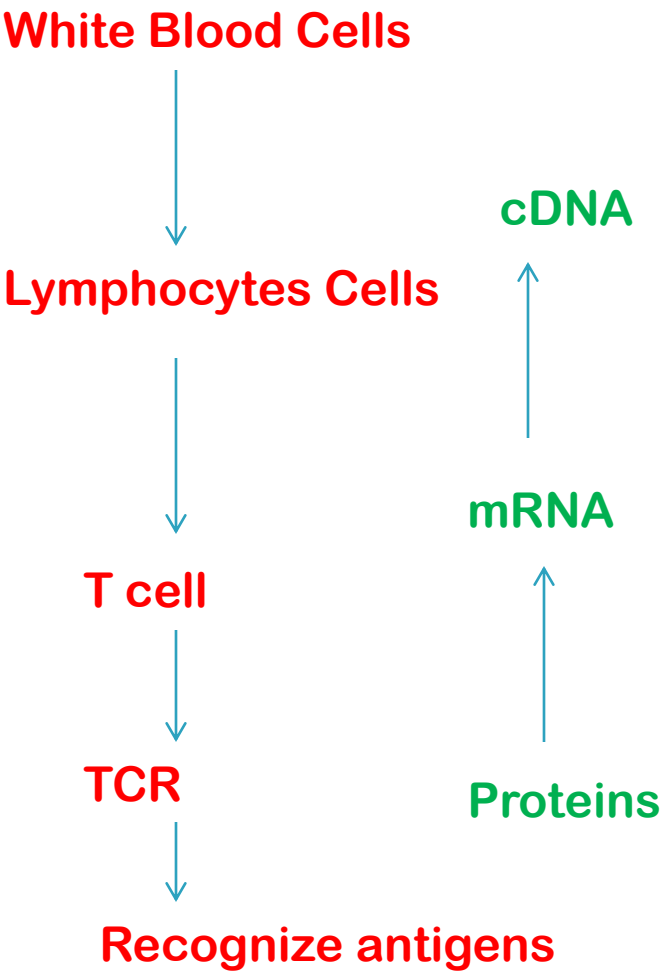
Accession codes

Accessions

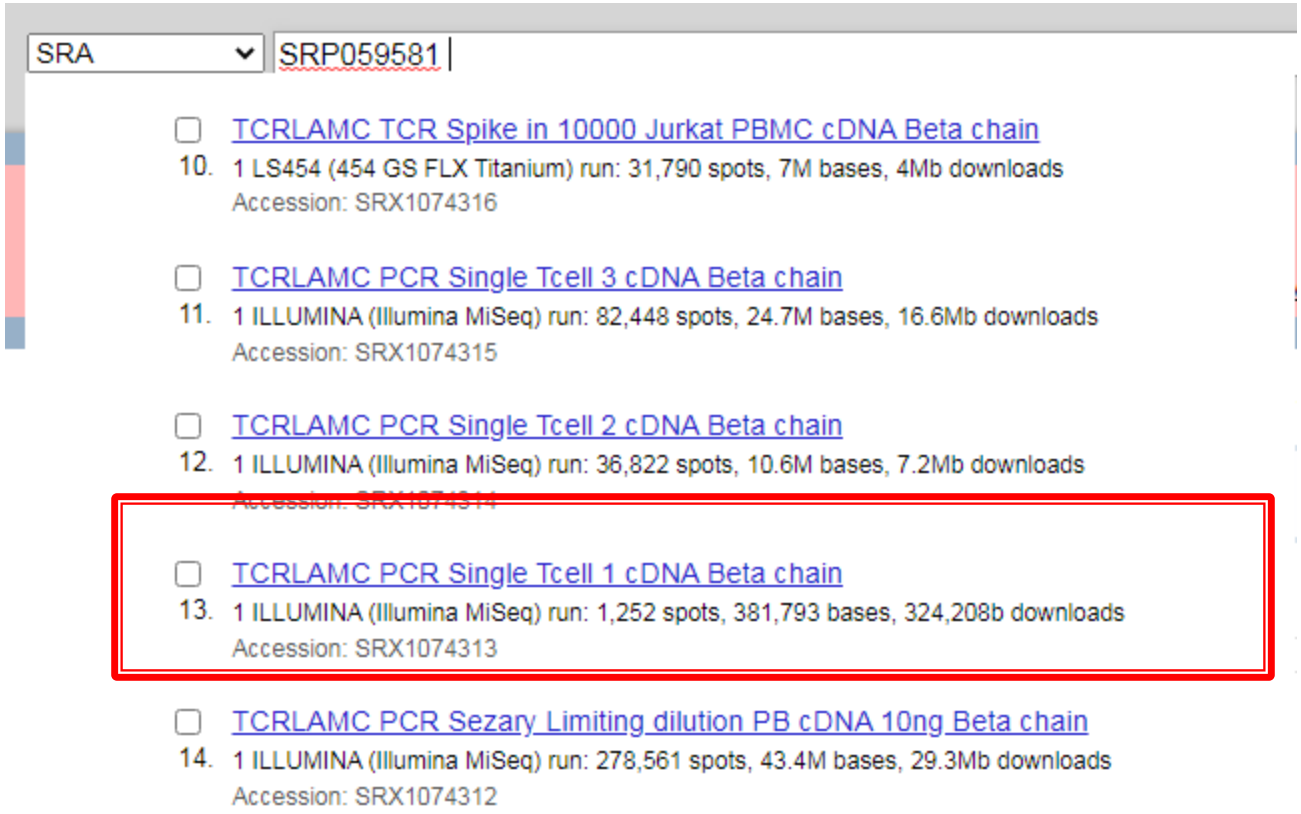
Sequence Read Archive

SRP059581

Note


White Blood Cells
Lymphocytes Cells
T cell
TCR
Recognize antigens
cDNA
mRNA
Proteins

SRA


SRA
SRP059581
TCRLAMC TCR Spike in 10000 Jurkat PBMC cDNA Beta chain
10. 1 LS454 (454 GS FLX Titanium) run: 31,790 spots, 7M bases, 4Mb downloads
Accession: SRX1074316
TCRLAMC PCR Single Tcell 3 cDNA Beta chain
11. 1 ILLUMINA (Illumina MiSeq) run: 82,448 spots, 24.7M bases, 16.6Mb downloads
Accession: SRX1074315
TCRLAMC PCR Single Tcell 2 cDNA Beta chain
12. 1 ILLUMINA (Illumina MiSeq) run: 36,822 spots, 10.6M bases, 7.2Mb downloads
TCRLAMC PCR Single Tcell 1 cDNA Beta chain
13. 1 ILLUMINA (Illumina MiSeq) run: 1,252 spots, 381,793 bases, 324,208b downloads
Accession: SRX1074313
TCRLAMC PCR Sezary Limiting dilution PB cDNA 10ng Beta chain
14. 1 ILLUMINA (Illumina MiSeq) run: 278,561 spots, 43.4M bases, 29.3Mb downloads
Accession: SRX1074312

SRA

SRX1074313: **TCRLAMC PCR Single Tcell 1 cDNA Beta chain**
1 ILLUMINA (Illumina MiSeq) run: 1,252 spots, 381,793 bases, 324,208b downloads

Submitted by: DKFZ

Study: TCR ligation anchored-magnetically captured PCR (TCR-LA-MC PCR) for TCR a-and ß-chain diversity dissection
PRJNA287162 • SRP059581 • All experiments • All runs
show Abstract

Sample: SingleTCell_1_cDNA
SAMN03797456 • SRS973392 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: TRANSCRIPTOMIC
Selection: PCR
Layout: SINGLE

Runs: 1 run, 1,252 spots, 381,793 bases, 324,208b

Run	# of Spots	# of Bases	Size	Published
SRR2079548	1,252	381,793	324,208b	2015-07-02

SRA

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide

Protected Data Usage Guide

Frequently Used Tools:

fastq-dump: Convert SRA data into fastq format

prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data

sam-dump: Convert SRA data to sam format

sra-pileup: Generate pileup statistics on aligned SRA data

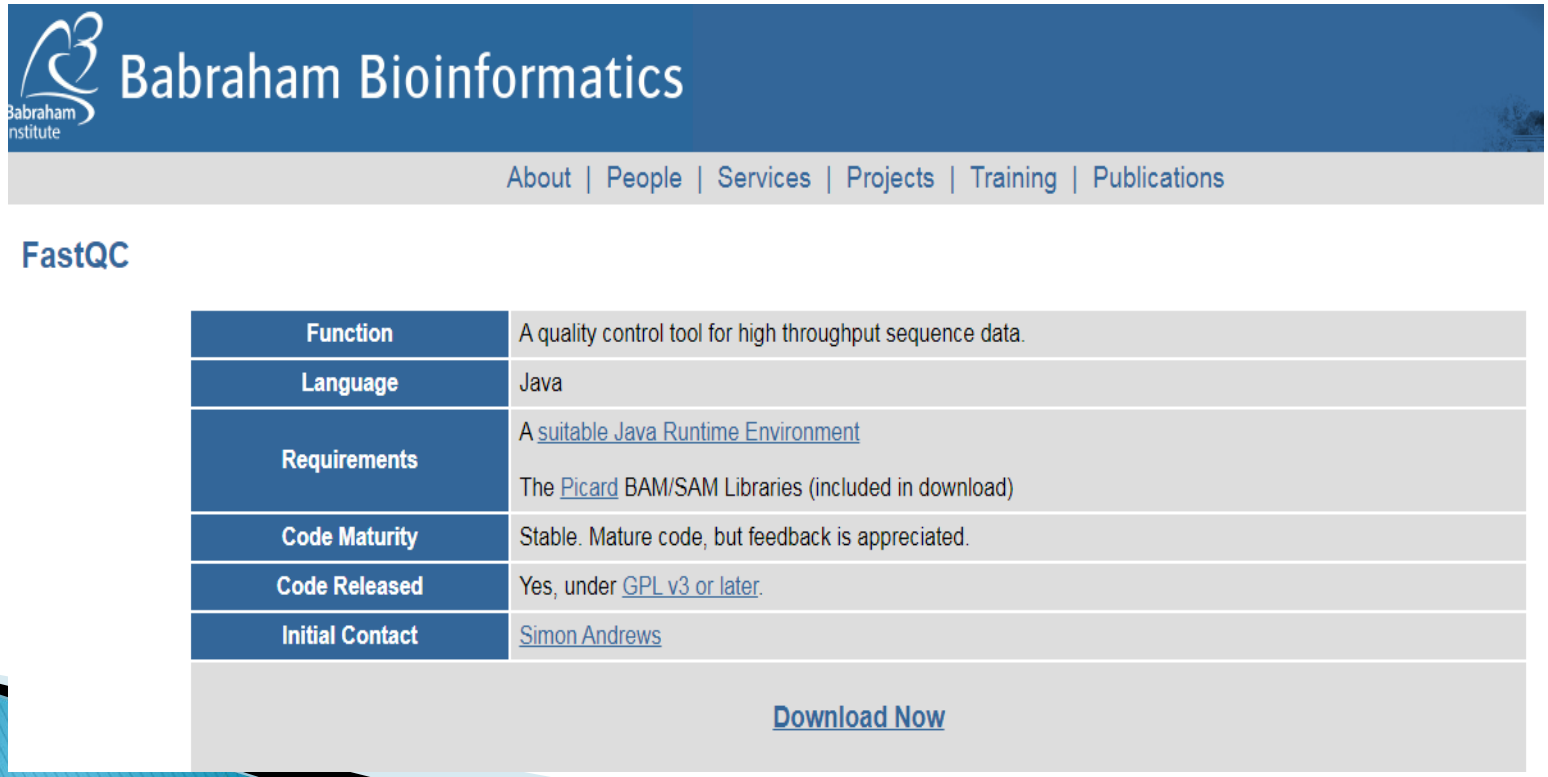
vdb-config: Display and modify VDB configuration information

vdb-decrypt: Decrypt non-SRA dbGaP data ("phenotype data")

- `cd SRR2079547`
- `./fastq-dump file.sra`

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

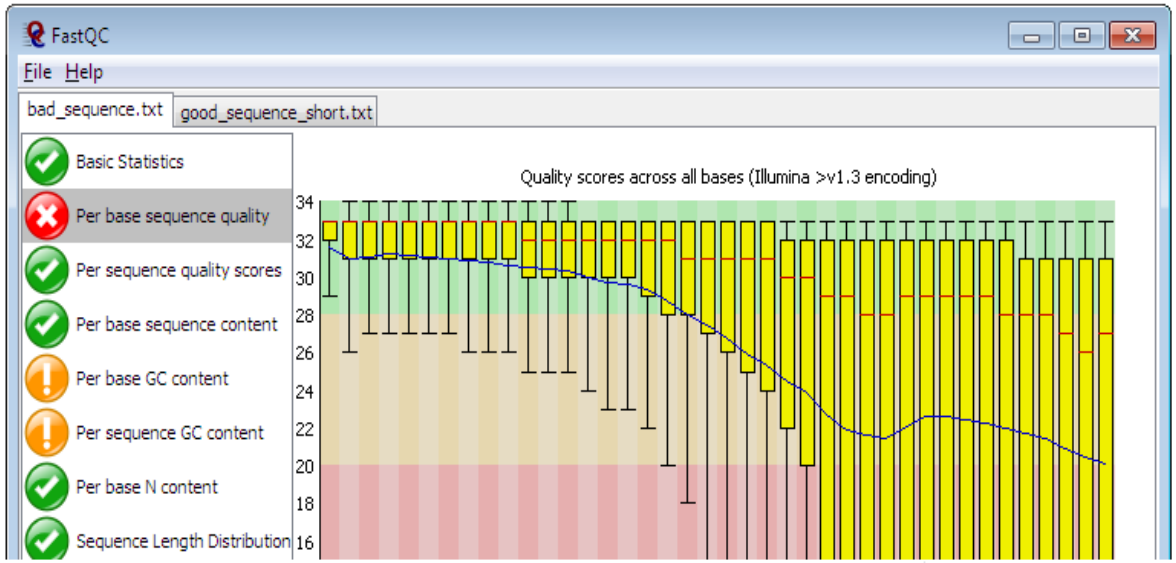
master 3 branches 4 tag

s-andrews Merge pull request #28 fror

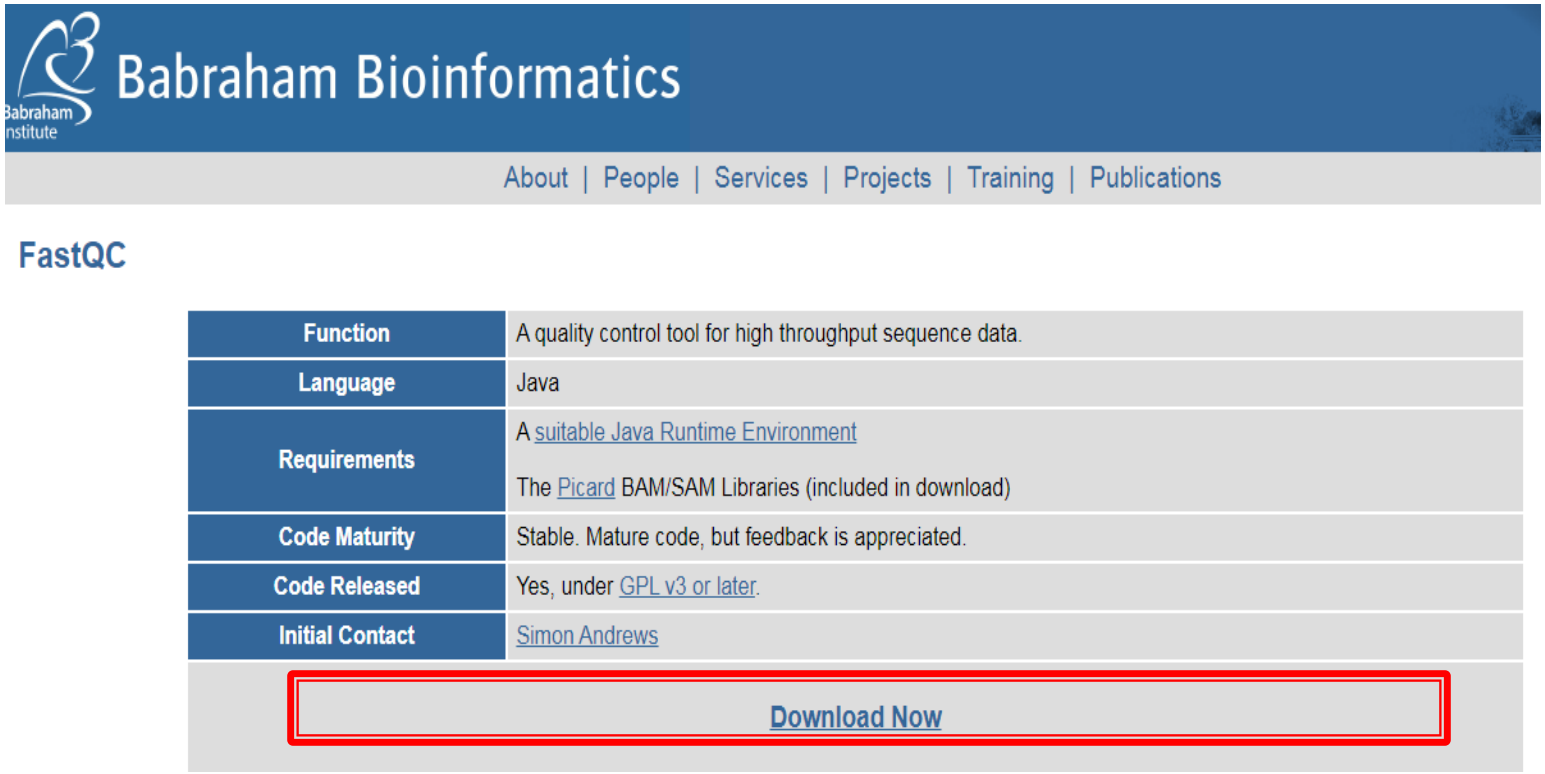
- .settings
- Configuration
- Help
- Templates
- net/sourceforge/iharder/base64
- org/apache/commons/math3
- uk/ac/babraham/FastQC

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
```

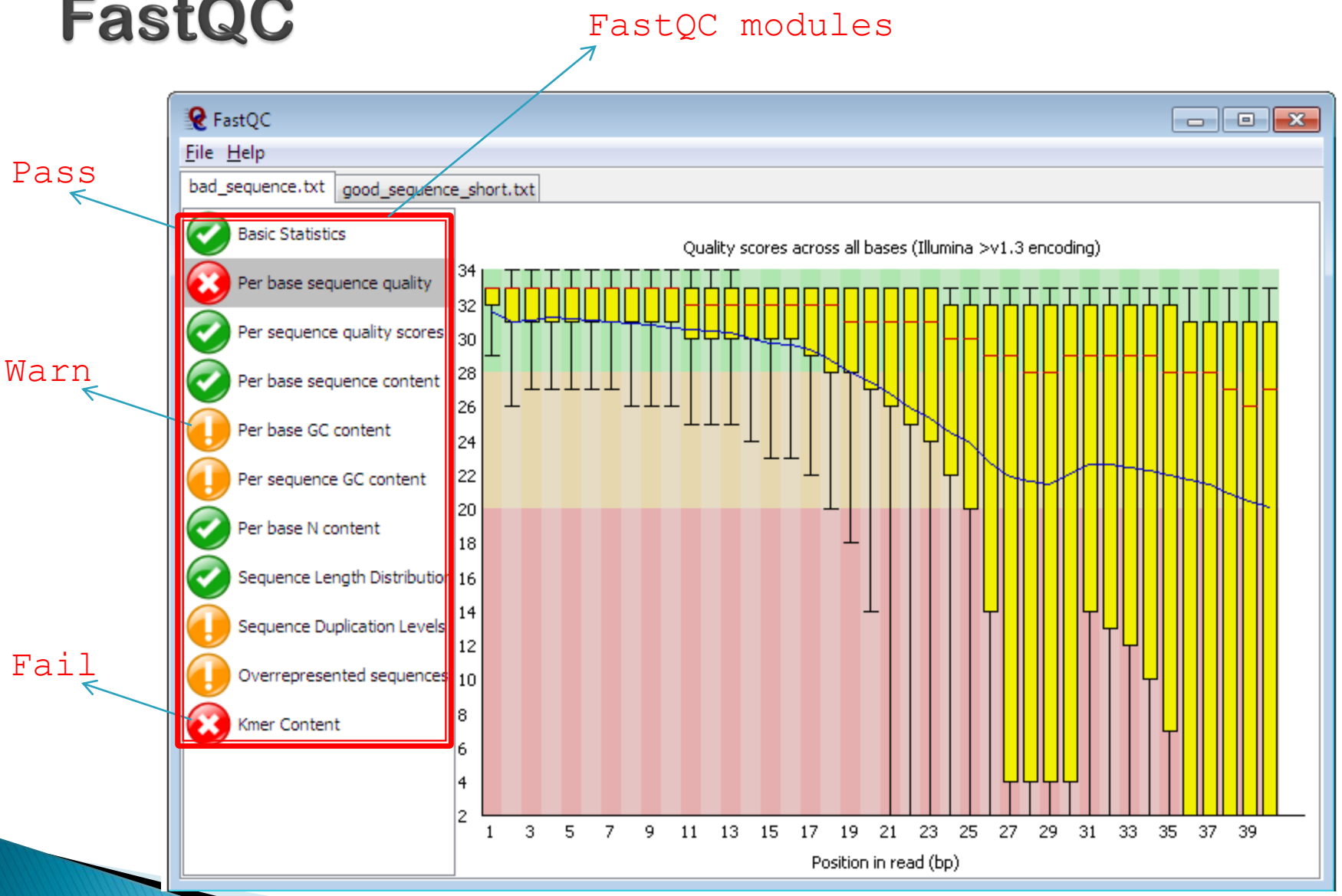
FastQC

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

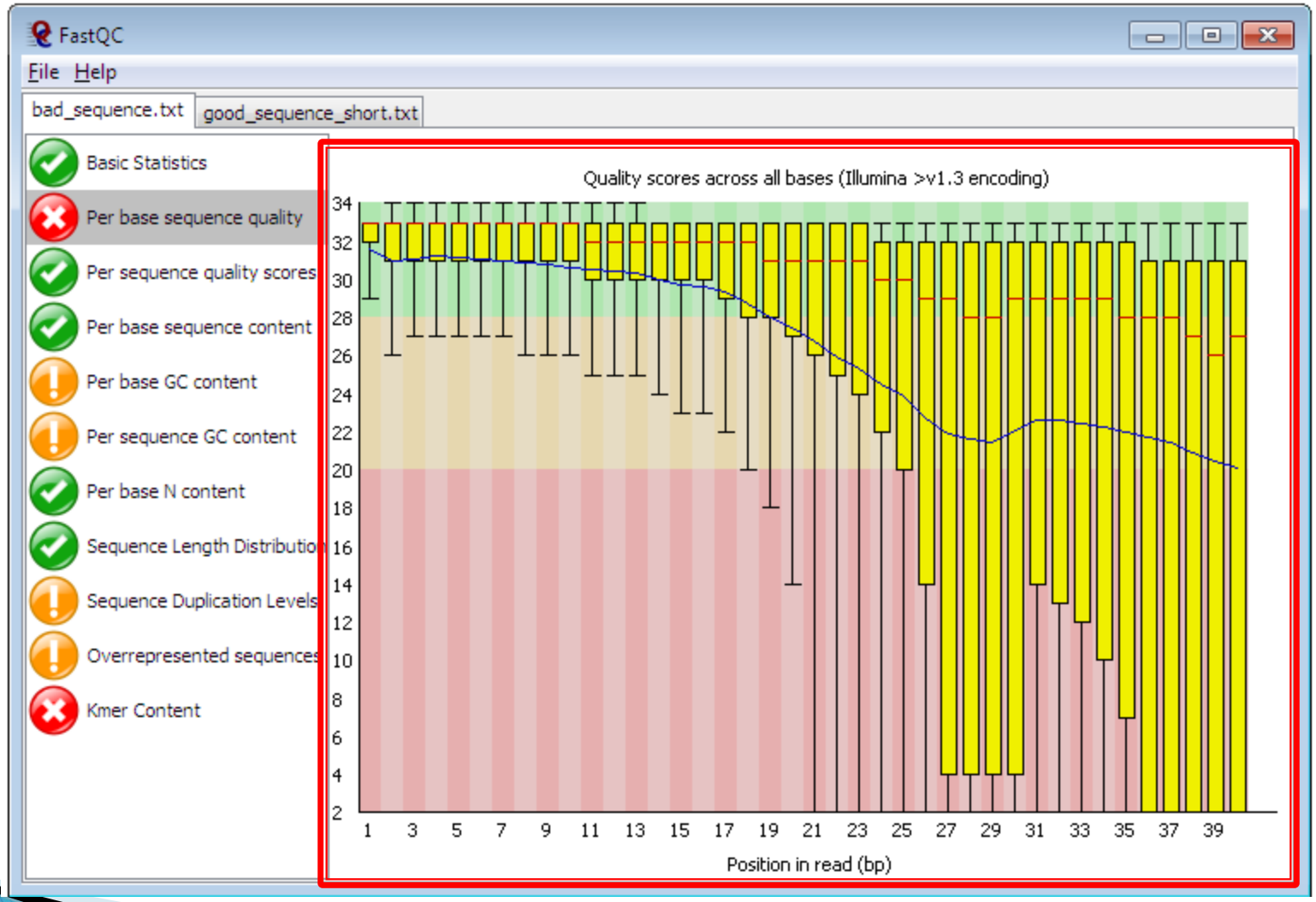
- `wget download link fastqc`
- `chmod 755 fastqc`
- `./fastqc`

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

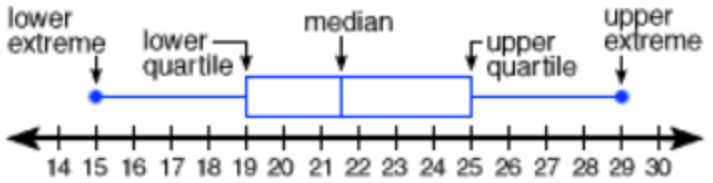
Box and whisker plot



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**

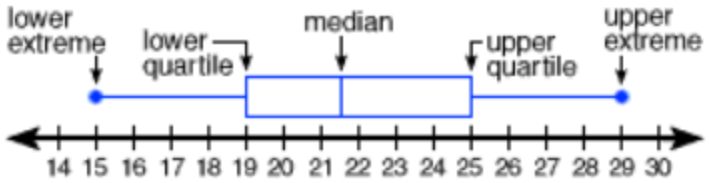
   There are six numbers below the median, namely: 6, 7, 13, 17, 20, 25.
   Q1 = the median of these six items
   = (6 + 1) ÷ 2= 3.5^{th} value
   = (third + fourth observations) ÷ 2
   = (13 + 17) ÷ 2
   = **15**

   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:

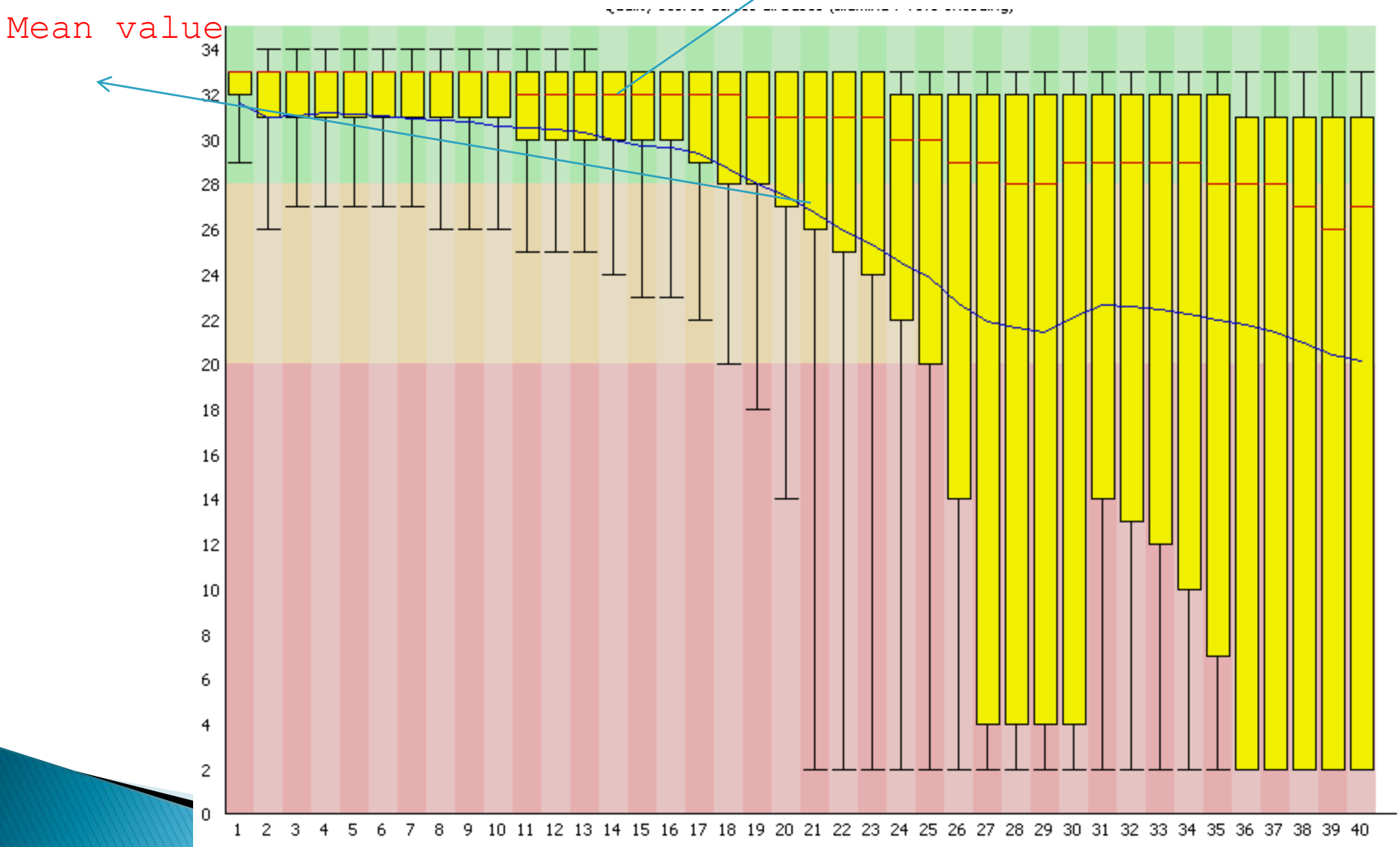
10, 11, 12, 14, 14, 16, 17, 17, 22, 23, 24

resulting:

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

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

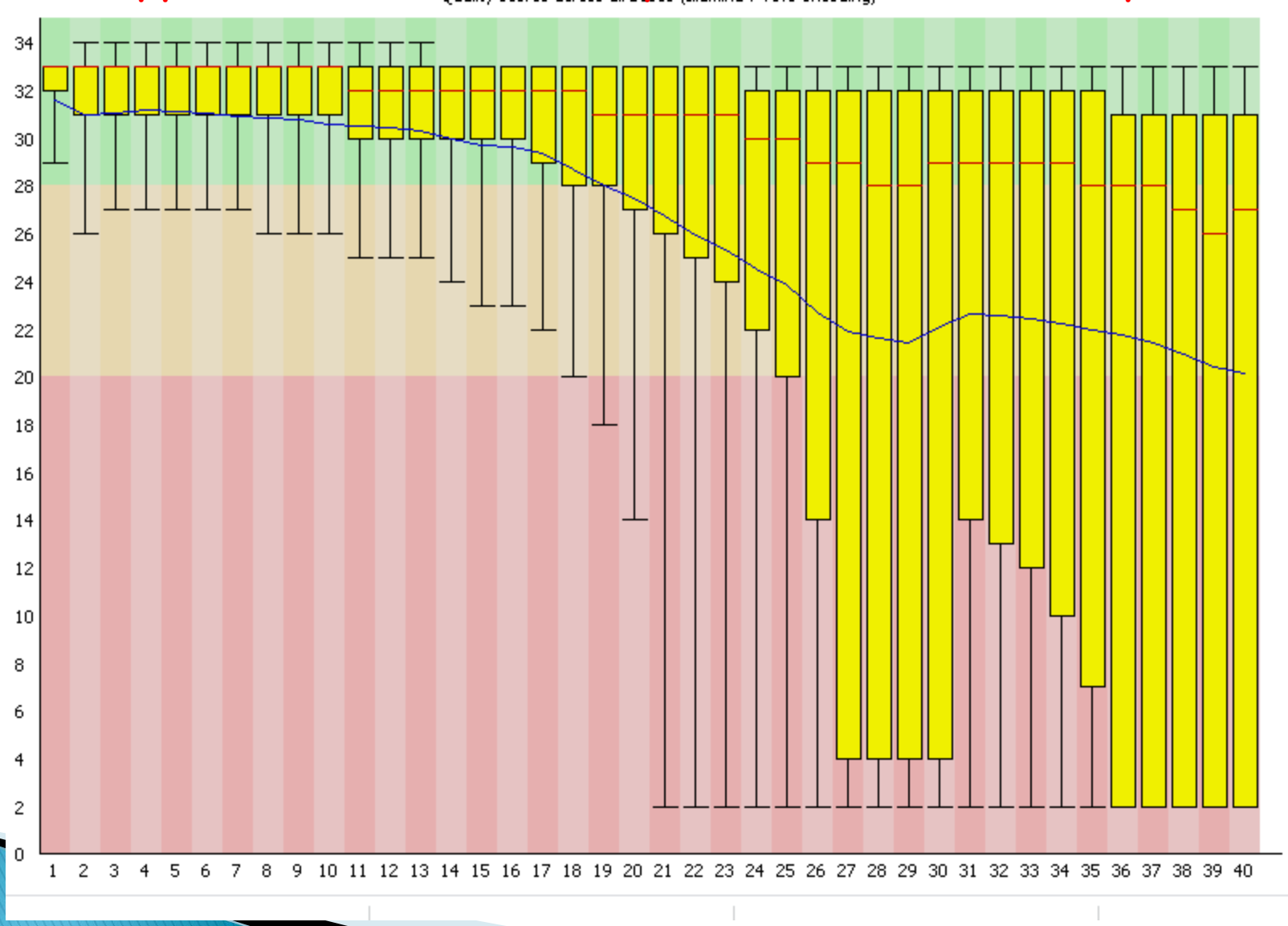
FastQC



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

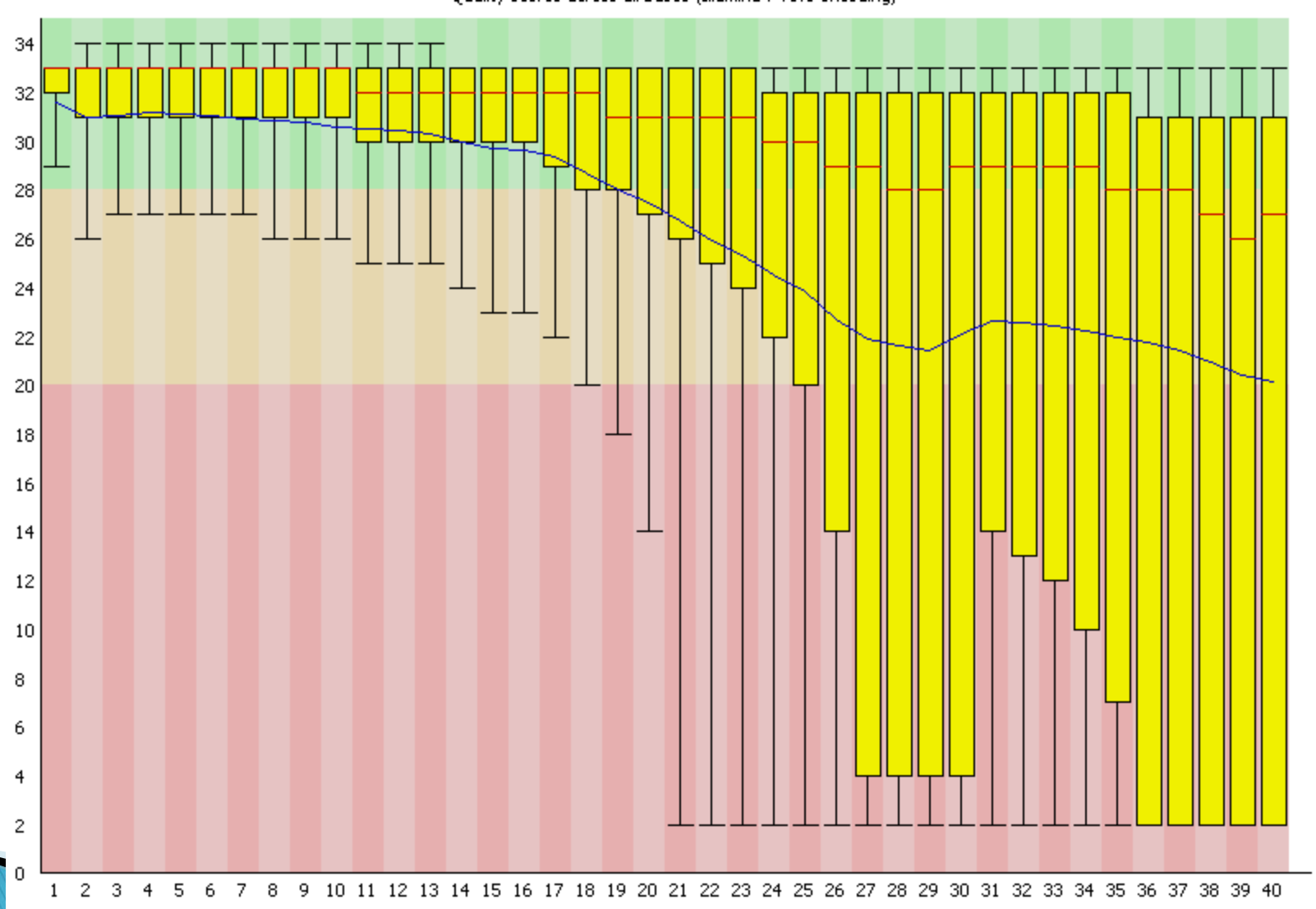
FastQC

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



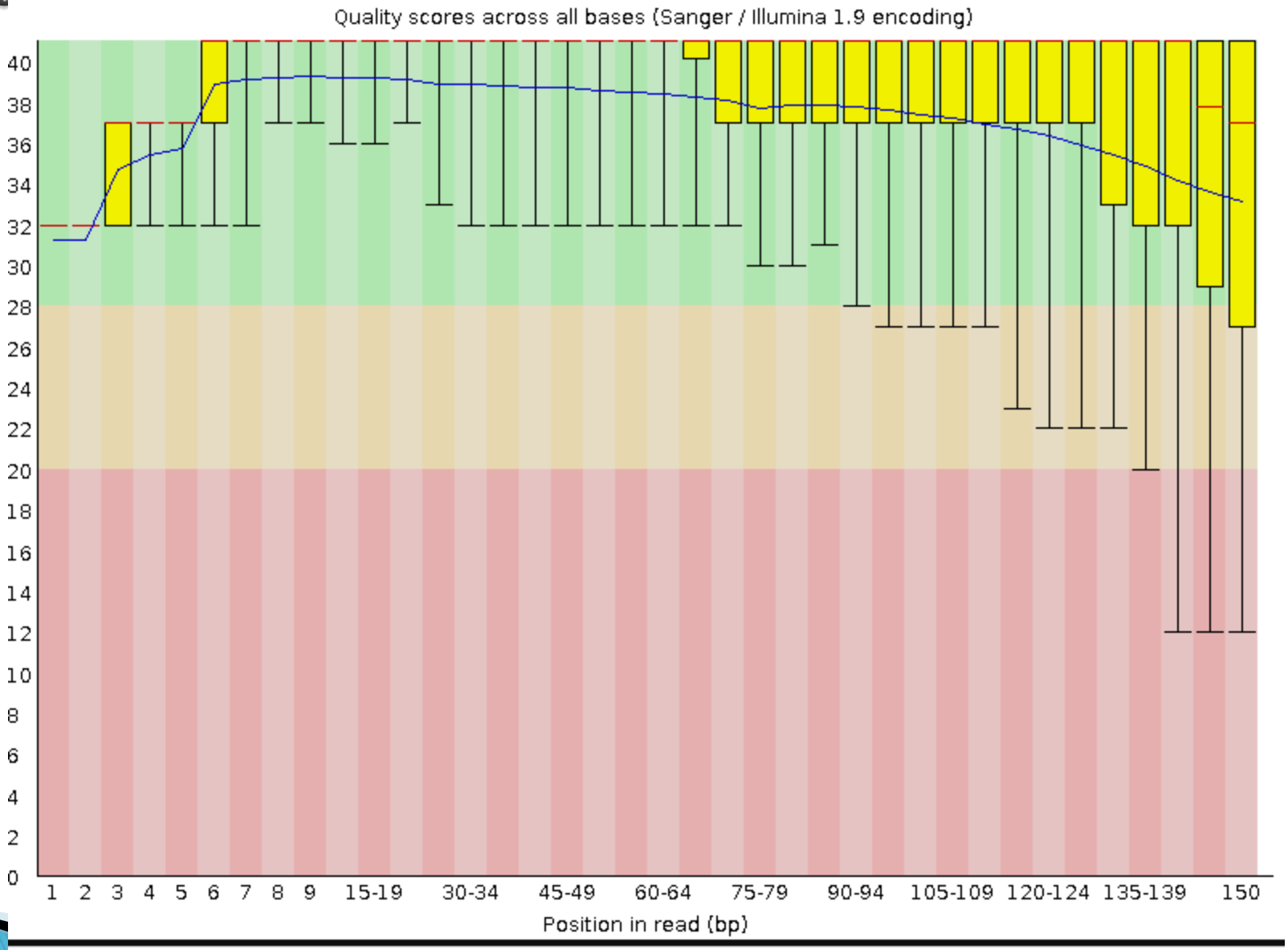
FastQC

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

Thanks!

// | ?