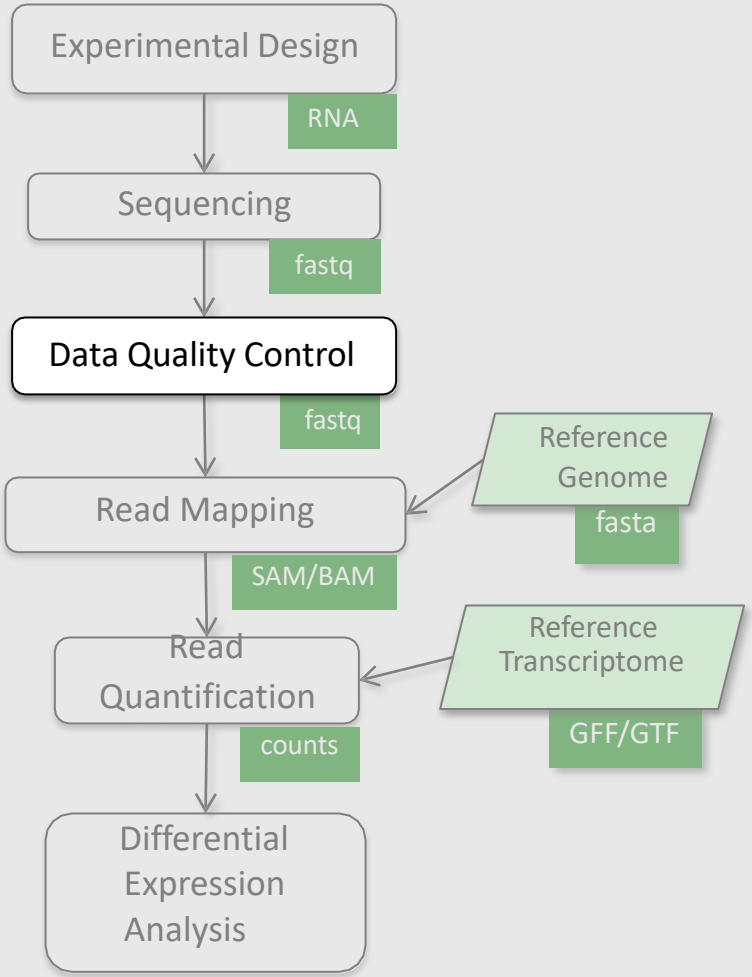
RNA-seq Bioinformatics: Format and QC of short reads

Falko Noé

University of Zurich[UZH]
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



- **Short read file format**
- Quality control of short reads

Properties of sequencing data

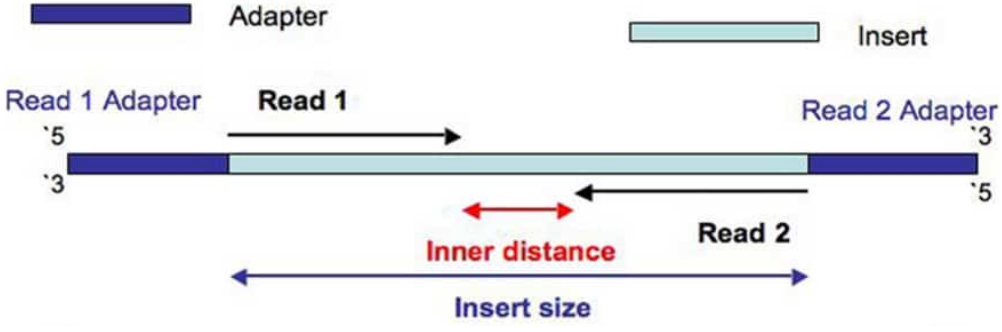
Technology	Read Length	Accuracy	Major error type
Sanger	400 to 900 bp	99.9%	Mismatch
Illumina	50 to 300 bp	98%	Mismatch
ONT	Limited only by the DNA molecules presented	65%-88% (traditionally) 99%-99.9% (current gen)	Indel
PacBio	10 kb to > 40 Kb	99.9999% circular consensus; 87% subread	Indel

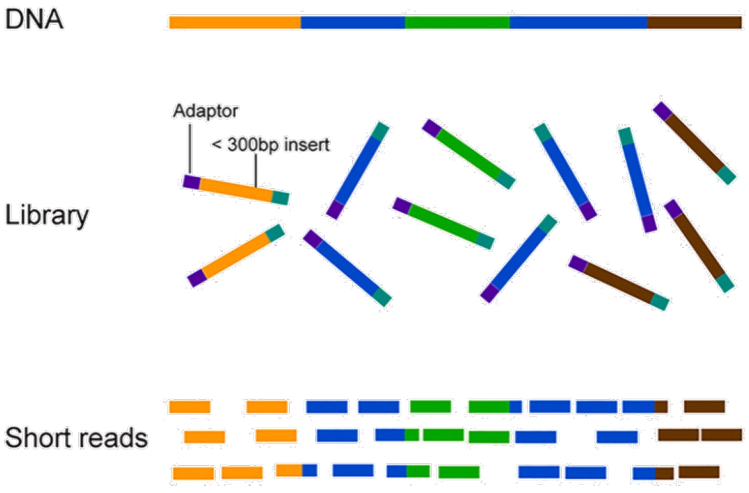
Sequence data – fastq files

- Data delivery (**fastq files**) from sequencing core
  - http://fgcz-gstore.uzh.ch/projects/pXXXX/NextSeq500_20160413_NS38_o2430/
- File names
  - Single end
    - 20160413.A-APC_mut_1_R1.fastq.gz
    - 20160413.A-APC_mut_2_R1.fastq.gz
  - Paired end
    - 20160413.A-APC_mut_1_R1.fastq.gz
    - 20160413.A-APC_mut_1_R2.fastq.gz
    - 20160413.A-APC_mut_2_R1.fastq.gz
    - 20160413.A-APC_mut_2_R2.fastq.gz

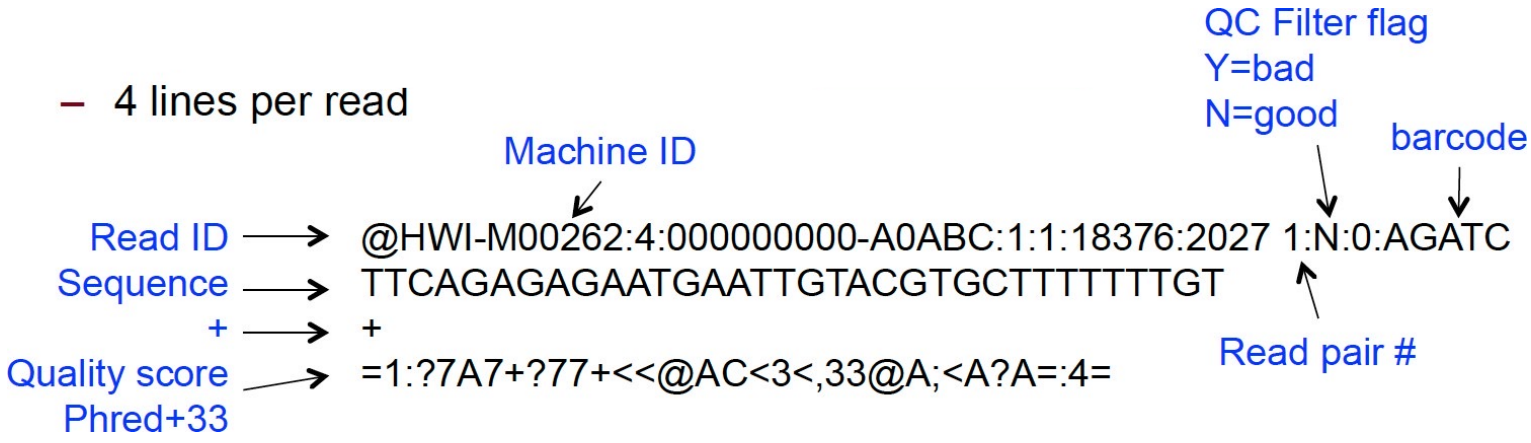
Sequence data – Illumina

- Single-end vs. paired-end reads





Fastq file format



1. Header line for Read (starts with "@" and the sequence ID)
2. Sequence
3. Header line for Qualities (starts with "+")
4. Quality score

Phred scores

- Measure base calling accuracy
  Accuracy of assigning **bases** (nucleobases) to signal peaks

- P
  error probability of a given base call
- Q
  $-10log_{10}P$

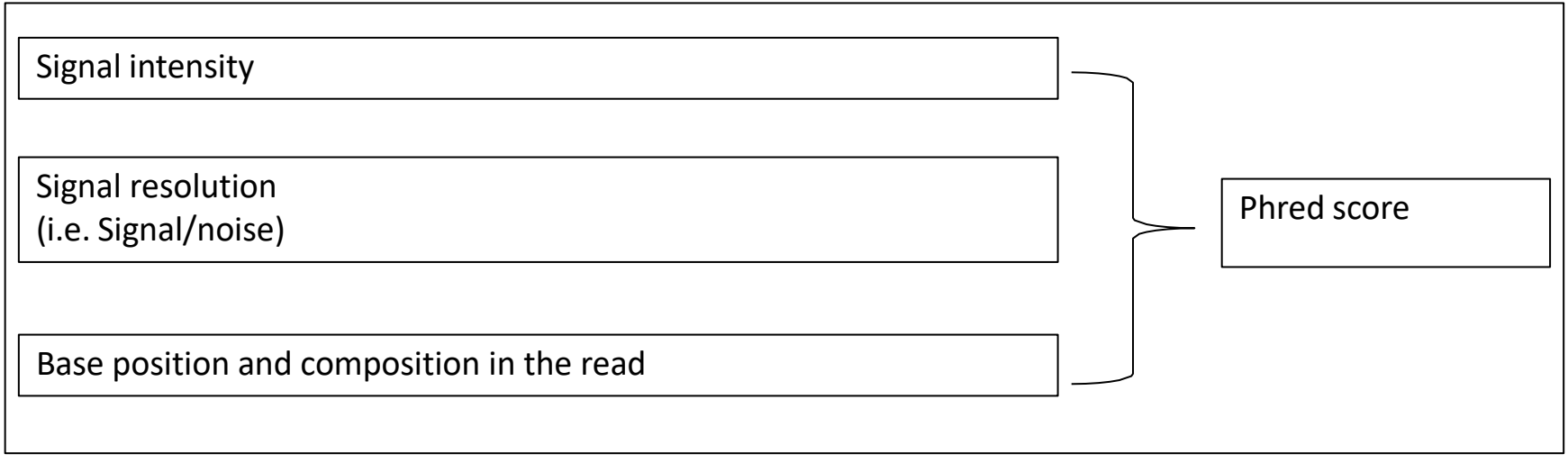
- Assign to each base

- Range from 0-41 for Illumina sequencing

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

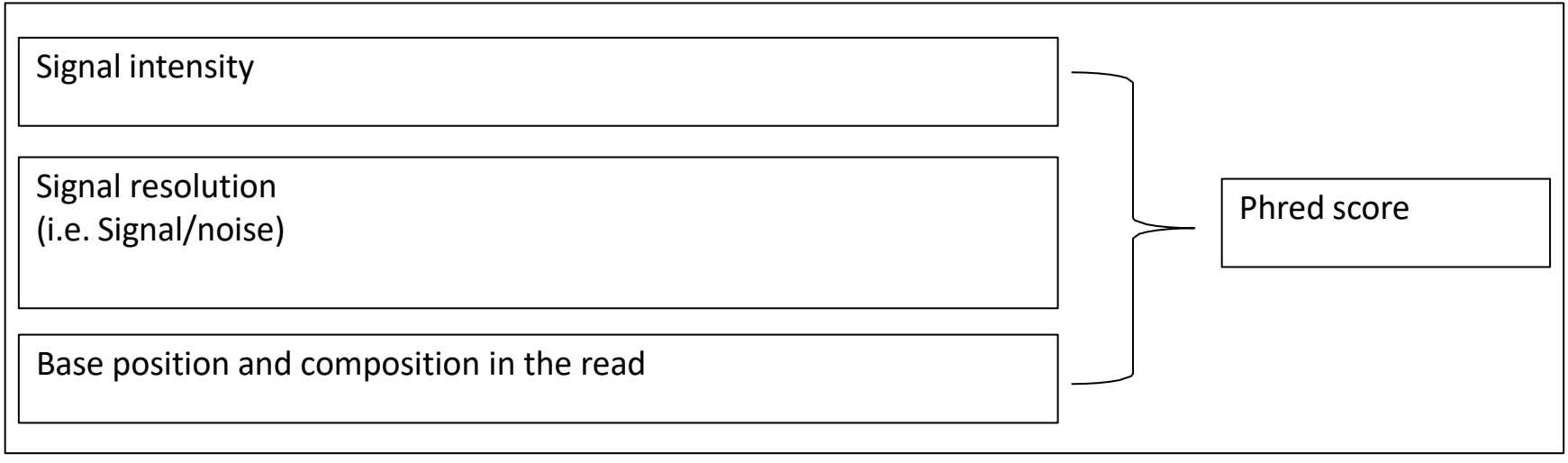
Ewing B, Green P. 1998. Genome Res. 8(3):186-194.

How are Phred scores generated?



Phred algorithm (Ewing and Green, 1998)

How are Phred scores generated?



- Parameters measured during sequencing **real samples**
- Exact sequences are not known

- **Scores assigned** by searching the look- up table

Phred algorithm (Ewing and Green, 1998)

Phred scores can be **ASCII encoded**

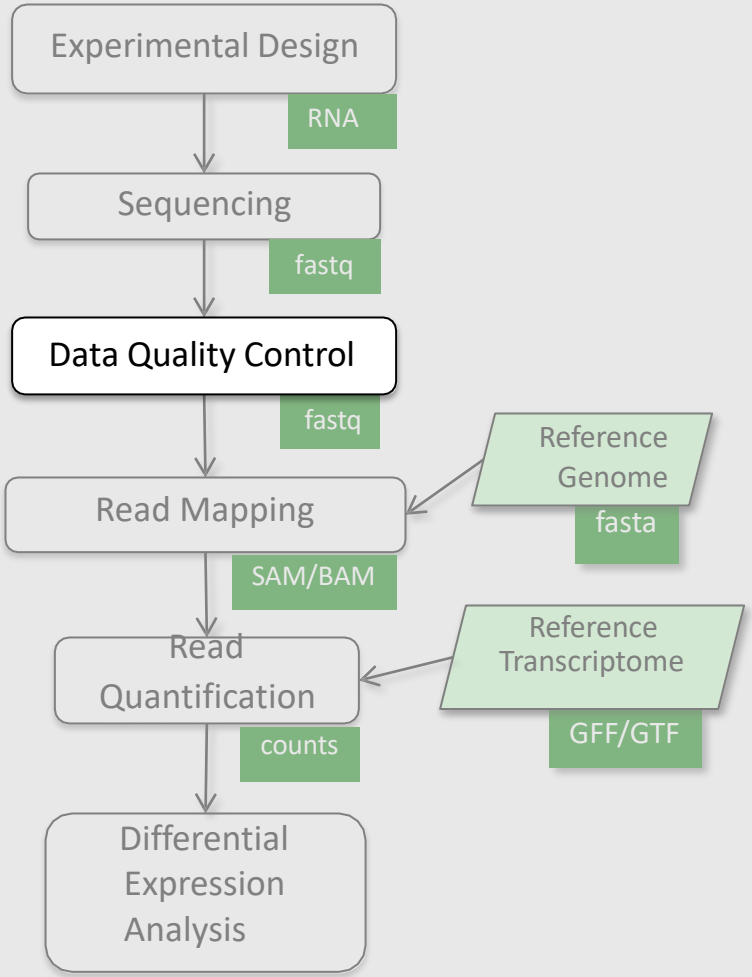
- Add an offset and convert the sum to ASCII

- Current format
  - **Illumina 1.9 (i.e. Sanger format)**
  - Phred scoring: 0-41;
  - Offset: 33
  - 41+33=74 (J)
  - All current sequencers

Million of reads

```
@HWI-ST1034:40:C08PJACXX:2:1101:20681:1994 1:N:0:ATCACG
CTCGNAGACTGGCAACTTGTTCTGGTTTACTGCACCTTCTTTTAAAGGCAGAAAGGCTTTTTGATAAAGAAGTTGTGAAAAGGCTACATGAGCTGCTTTTA
+
CCCF#2ADHHHGHJJJIJJHIIIJJHIJJJJJJJJJJBGIIJJJJJJJJJJJJJJJIJJJJJJJJJIJGHHHHGFFFFFFECEDDDDDCCDCDDDDDDDDD
@HWI-ST1034:40:C08PJACXX:2:1101:1907:2005 1:N:0:ATCACG
CTCACCTTCAACTGTATTCACGCTTGGACCACAGATCTTGGCCTTAGTGCGATATAGGACACAACATTTCTTTCCTCTGGCTACCACCAAGGAACCCTTCA
+
CCCFFFFFHHHHHJGIJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJHIJJIJIJJJIJJJJJJJJHHHHHHFFDFCEDE@@CCCACD@DBDDDDDB=BCCC
@HWI-ST1034:40:C08PJACXX:2:1101:2155:2031 1:N:0:ATCACG
CAATCAATTAACAATATTAGTTACATAAGCACTTCCTTAACCACCCTCTCAAAGTTGGCAAAGAAGAACCCCCTTTCTCAATAGCTTTAACCGCGCTCTC
+
CCCFFFFFHHHHHJJJJJJJIJJJJJJJIJJJJJJJJJJJIIJJJJJJJJJJFHHIIJGIIJHJGHIIJHHHFFFFFFEDDEEEEDCDCDDDDDDDD
@HWI-ST1034:40:C08PJACXX:2:1101:2220:2057 1:N:0:ATCACG
CTGGATCAGGATGCATGGCTGCCTTAGATGCTGCTTACCTGCAGGAGATTGGTGCTCAAGCCGGGAAGACGGACTGACTCCTCATATTTTGCCGCCTA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJIJJIJJJJJJIIJJJJIHIEHDHIJJHIFHGIJJJIJHHDDBBCABB@@BACADDDDDDDDEEDCDCBDDD9
@HWI-ST1034:40:C08PJACXX:2:1101:2466:2116 1:N:0:ATCACG
CTGCCGGCCGCTACATGTGCGCTGAATGCGTCCTCCATCGCGAAGACTACGTGCGCATGCTCGCTCAGCTCTTCCCAGACTACCCAATCCTTGCCAAGTGC
+
CCCFFFFFHHHHHJJJJGHIJJJJJJJJJJHJIJJJJIIIIHFDDDDDDDDBDDDDDBDDDDDDDDDDDDDDDDDDDDDDDDDDDCBDDDDDCDDDCBCA>
@HWI-ST1034:40:C08PJACXX:2:1101:2463:2168 1:N:0:ATCACG
CGTTCATATGCAAAAGAAGCTTCTCAGTCTGCTTTACCACCTCTTAAAGGGGATCAAATGTTGAAGAACATCTTTTTTGAGGTAAAGAACAAATTTGATAT
+
CC@FFFFFHHHGHJJJJJJJJJJJIJJIJJIIJJJJJJEIJIJJJIJJJJJIJIJJJIJJJJJIJJJJJIIHHEEECABEDDDDDDDDDDDDDDDDED
@HWI-ST1034:40:C08PJACXX:2:1101:2378:2207 1:N:0:ATCACG
CACGCGGTGTGGAAAACCCCTTCACATCCATCAATGGCGGCTCGGAGCGATTCAAAATCAAGCATATCCGCTTTGTACAGCACAAGACGATCCGATGCTCC
+
BCCFFFFDHHHHHJJJJJJJJJJGJJJJJJJJJIJJIGIJHEHBD8>6?;ABCDDDDCCDDCCBCCCDEBBBB@CCEEEECCBCDDDCDBBB?BBBDDDDC
```

How accurate are they?
Are they from my samples?



- Short read file format
- **Quality control of short reads**

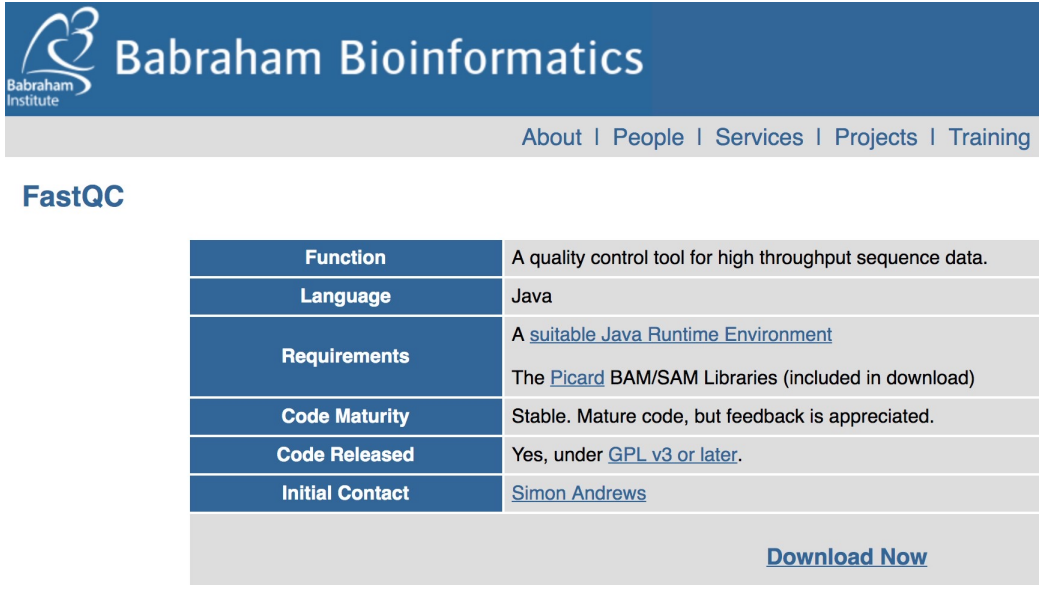
Million of reads

```
@HWI-ST1034:40:C08PJACXX:2:1101:20681:1994 1:N:0:ATCACG
CTCGNAGACTGGCAACTTGTTCTGGTTTACTGCACCTTCTTTTAAAGGCAGAAAGGCTTTTTGATAAAGAAGTTGTGAAAAGGCTACATGAGCTGCTTTTA
+
CCCF#2ADHHHGHJJJIJJHIIIJJHIJJJJJJJJJJBGIIJJJJJJJJJJJJJJJIJJJJJJJJJIJGHHHHGFFFFFFECEDDDDDCCDCDDDDDDDDD
@HWI-ST1034:40:C08PJACXX:2:1101:1907:2005 1:N:0:ATCACG
CTCACCTTCAACTGTATTCACGCTTGGACCACAGATCTTGGCCTTAGTGCGATATAGGACACAACATTTCTTTCCTCTGGCTACCACCAAGGAACCCTTCA
+
CCCFFFFFHHHHHJGIJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJHIJJIJIJJJIJJJJJJJHHHHHHFFDDDDDB@@CCCACD@DBDDDDDB@BCCC
@HWI-ST1034:40:C08PJACXX:2:1101:2155:2031 1:N:0:ATCACG
CAATCAATTAACAATATTAGTTACATAAGCACTTCCTTAACCACCCTCTCAAAGTTGGCAAAGAAGAACCCCCTTTCTCAAAAGCTTAACCGCGCTCTC
+
CCCFFFFFHHHHHJJJJJJJIJJJJJJJIJJJJJJJJJJJIIJJJJJJJJ@FHHIIJGIIJHJGHIIJHHHFFFFFFEDDEEEEDCDCDDDDDDDD
@HWI-ST1034:40:C08PJACXX:2:1101:2220:2057 1:N:0:ATCACG
CTGGATCAGGATGCATGGCTGCCTTAGATGCATGGATTACCTGCAGGAGATTGGGCCAAGCCGGGAAGACGGACTGACTCCTCATATTTTGCCGCCTA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJIJJIJJJJJJIIJJJJJIHIEHDHIJJHIFHGIJJJIJHHDDBBCABB@@BACADDDDDDDDEEDCDCBDDD9
@HWI-ST1034:40:C08PJACXX:2:1101:2466:2116 1:N:0:ATCACG
CTGCCGGCCGCTACATGTGCGCTGAATGCGTCCTCCATCGCGAAGACTACGTGCGCATGCTCGCTCAGCTCTTCCCAGACTACCCAATCCTTGCCAAGTGC
+
CCCFFFFFHHHHHJJJJGHIJJJJJJJJJJHJIJJJJIIIIHFDDDDDDDDBDDDDDBDDDDDDDDDDDDDDDDDDDDDDDDDDDCBDDDDDCDDDCBCA>
@HWI-ST1034:40:C08PJACXX:2:1101:2463:2168 1:N:0:ATCACG
CGTTCATATGCAAAAGAAGCTTCTCAGTCTGCTTTACCACCTCTTAAAGGGGATCAAATGTTGAAGAACATCTTTTTTGAGGTAAAGAACAAATTTGATAT
+
CC@FFFFFHHHGHJJJJJJJJJJJIJJIJJIIJJJJJJEIJIJJJIJJJJJIJIJJJJIJJJJIJJJJIJJJJIIHHEEECABEDDDDDDDDDDDDDDDED
@HWI-ST1034:40:C08PJACXX:2:1101:2378:2207 1:N:0:ATCACG
CACGCGGTGTGGAAAACCCCTTCACATCCATCAATGGCGGCTCGGAGCGATTCAAAATCAAGCATATCCGCTTTGTACAGCACAAGACGATCCGATGCTCC
+
BCCFFFFDHHHHHJJJJJJJJJJGJJJJJJJJJIJJIGIJHEHBD8>6?;ABCDDDDCCDDCCBCCCDEBBBB@CCEEEECCBCDDDCDBBB?BBBDDDDC
```

How accurate are they? - FastQC
Are they from my samples? - fastqScreen

FastQC

- Can be embeded in workflows as a data analysis module
- Can also be run as an independent application with GUI using a laptop
  - http://www.youtube.com/watch?v=bz93ReOv87Y



Babraham Bioinformatics

About | People | Services | Projects | Training

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

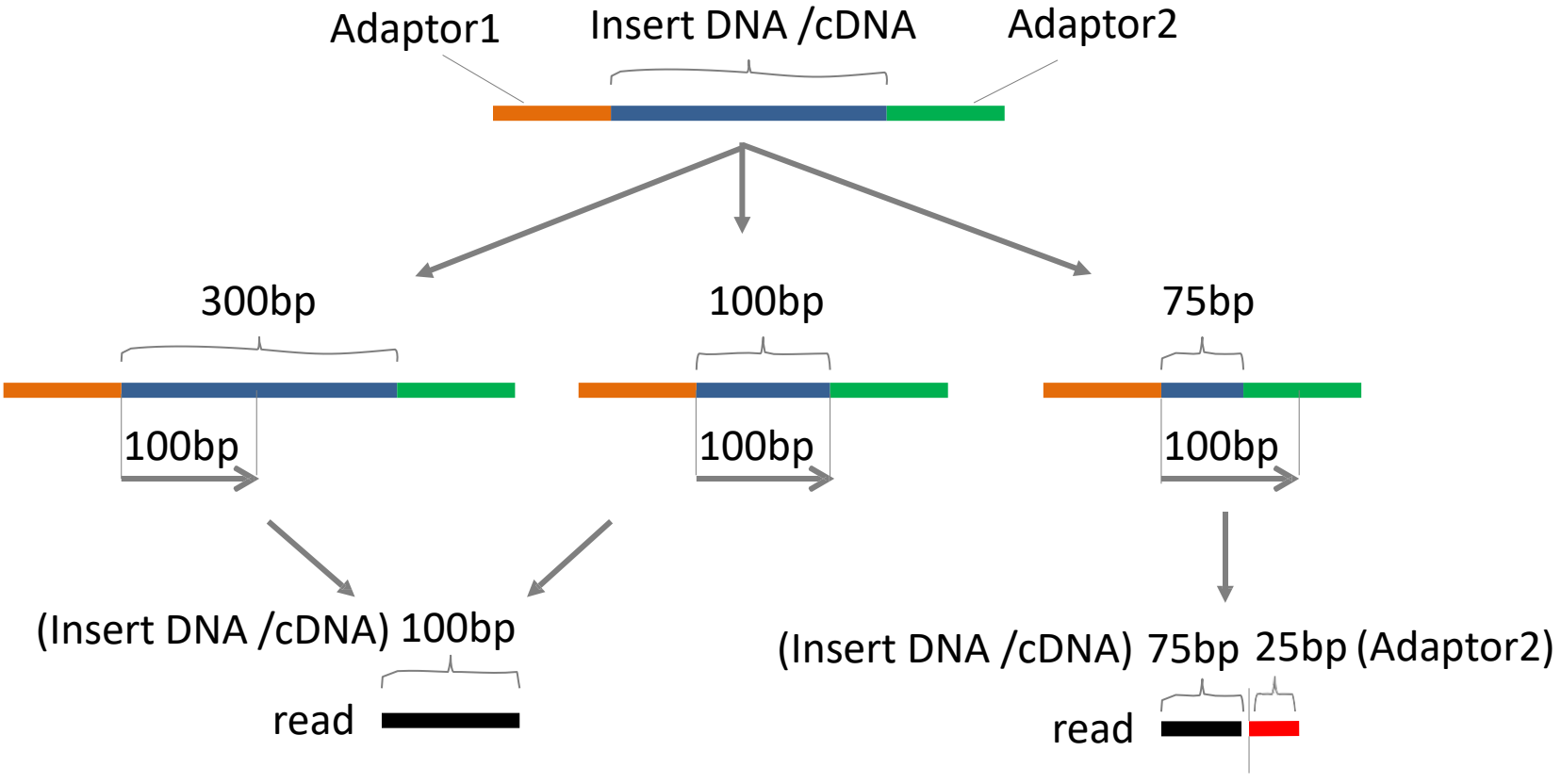
Download Now

Boxplots, histograms, heatmaps

- Fastqc uses boxplots and histograms for the analysis of phred scores and GC content

- Heatmaps can also be used. They are more visual and colors can be used to represent different information

Different scenarios

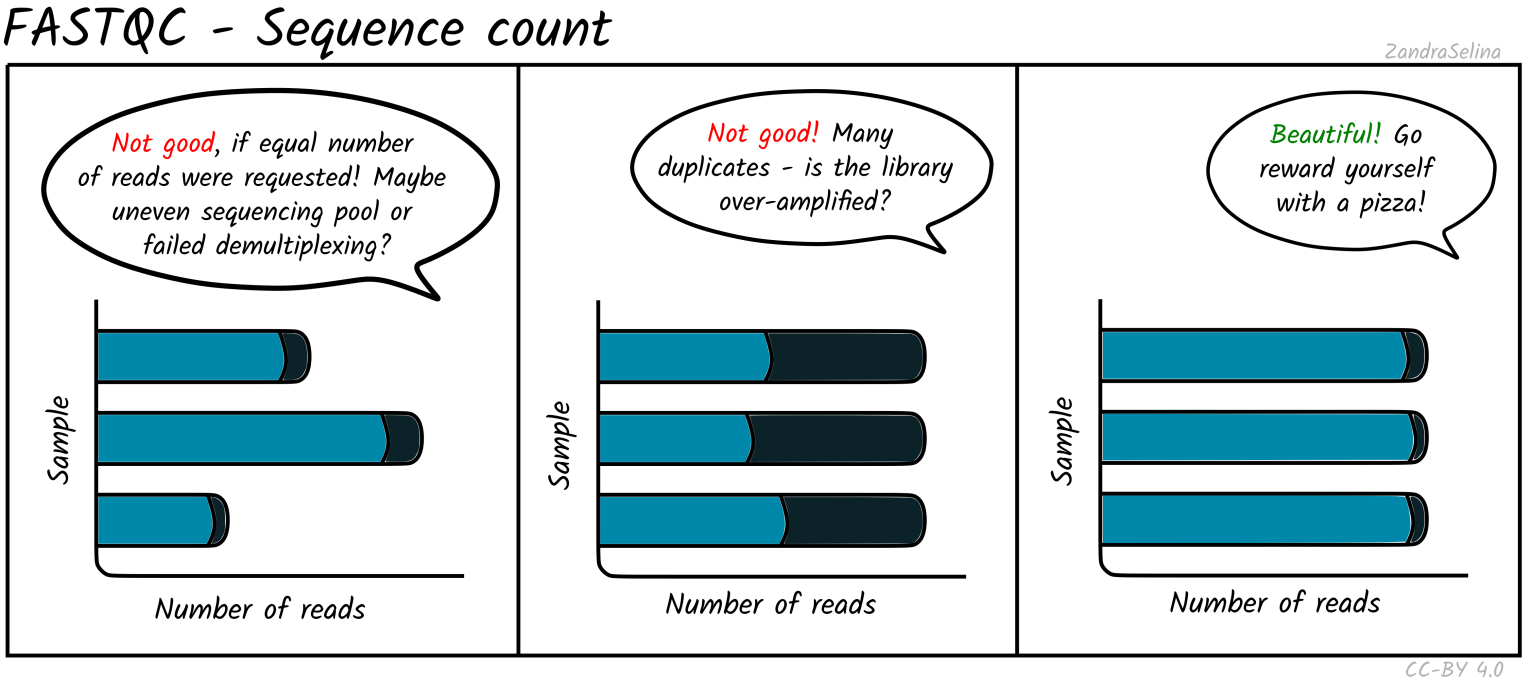


Bias and errors

- Library construction could introduce bias
  - Fragmentation, ligation, amplification
  - GC bias
  - Over-amplification
  - Contamination
- Sequencing errors
  - Chemical, optical, computational

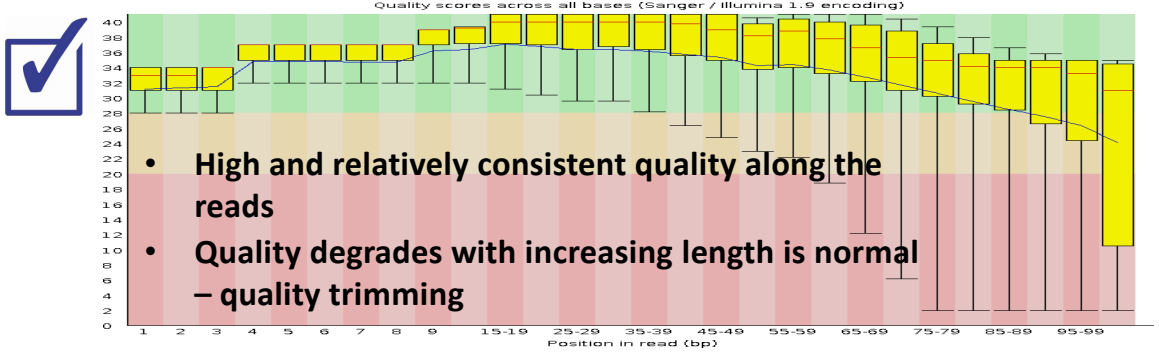
FastQC

- Sequence counts and basic statistics
- Do I have enough read counts/coverage?



Per base sequence quality - FastQC

- Range of quality values across all bases at each position

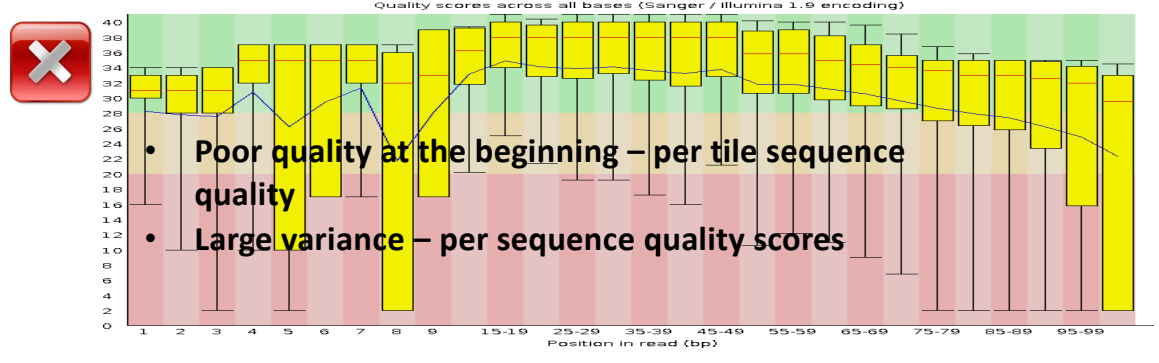


Green: >Q28, good

Orange: >Q20, reasonable

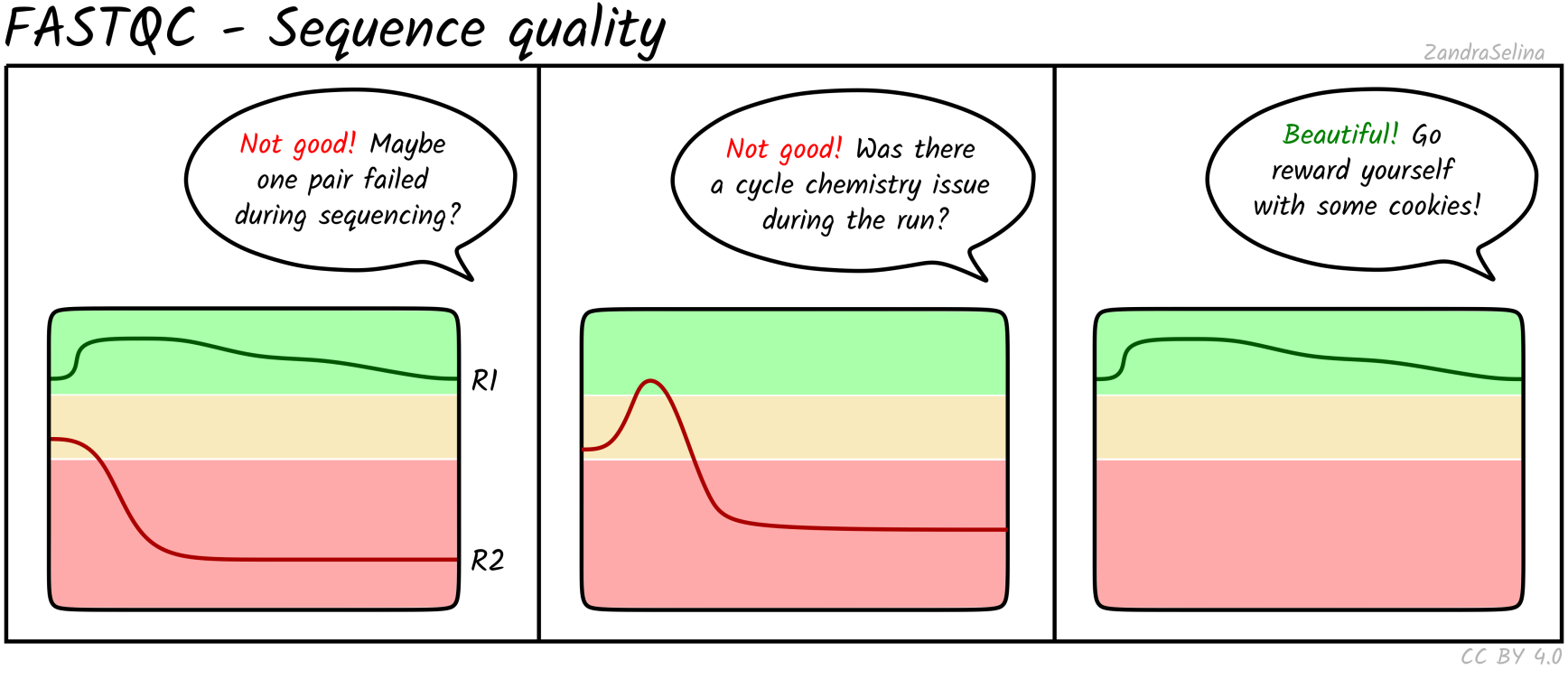
Red:<Q20, poor

Median > Q25



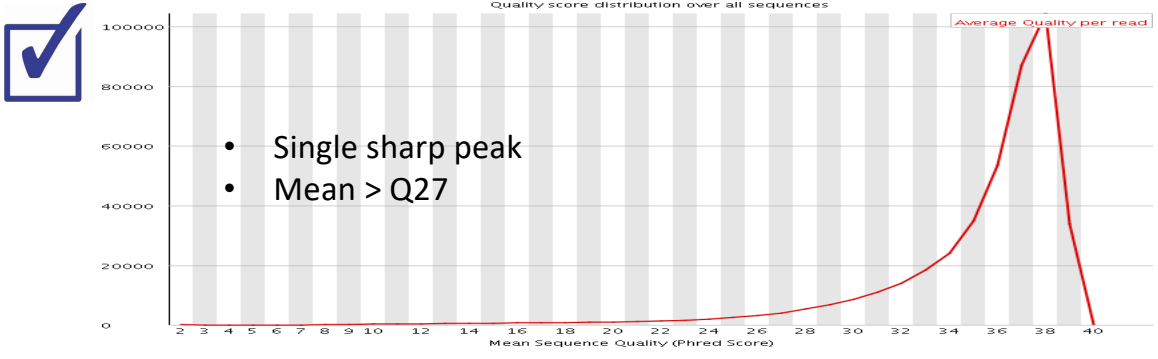
Median < Q20

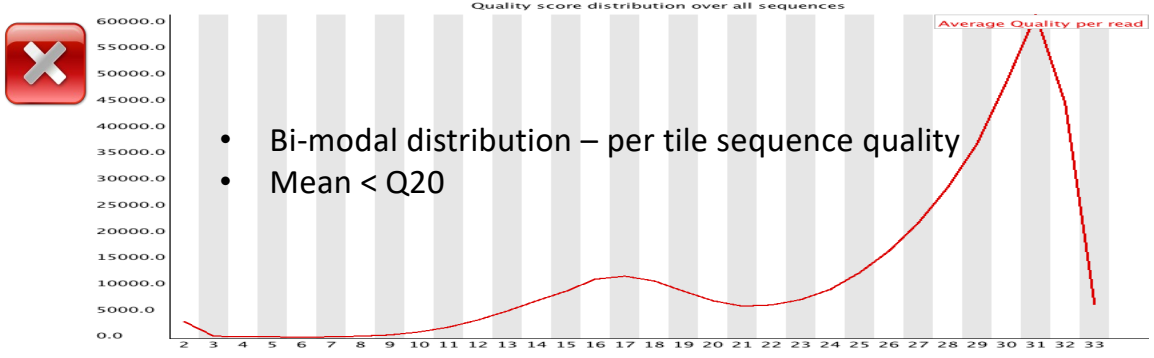
Per base sequence quality - FastQC



Per sequence quality scores - FastQC

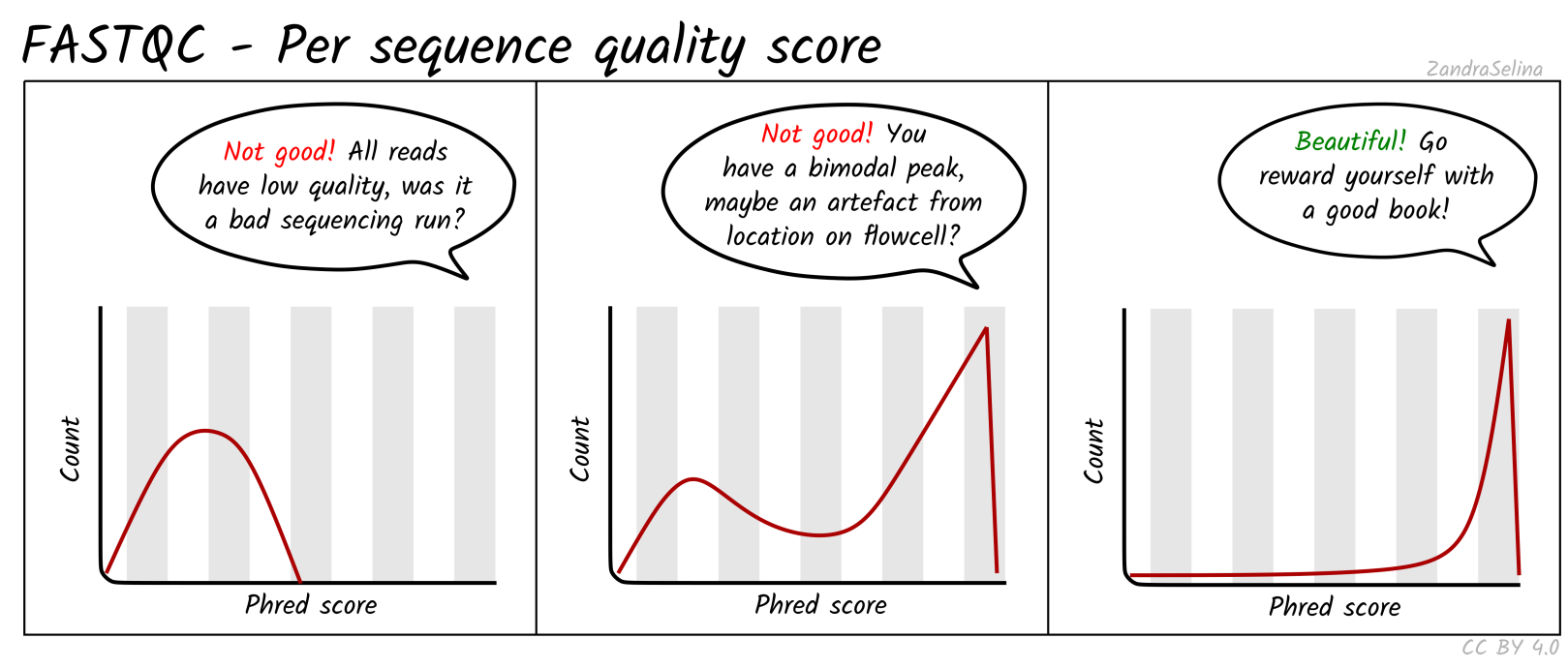
- Subset of sequences with universally low quality values





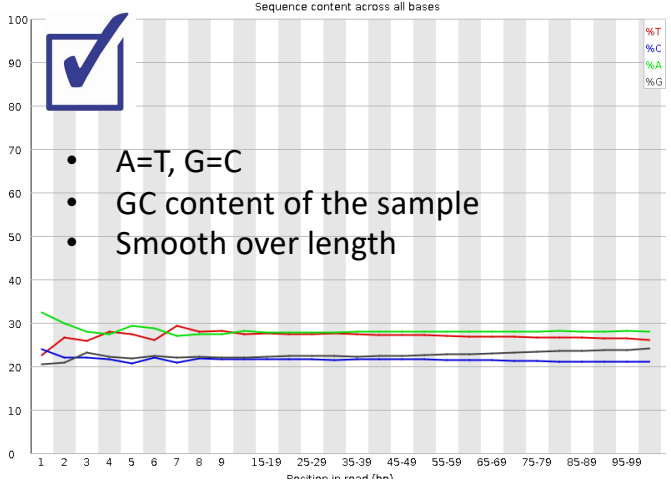
Per sequence quality scores - FastQC

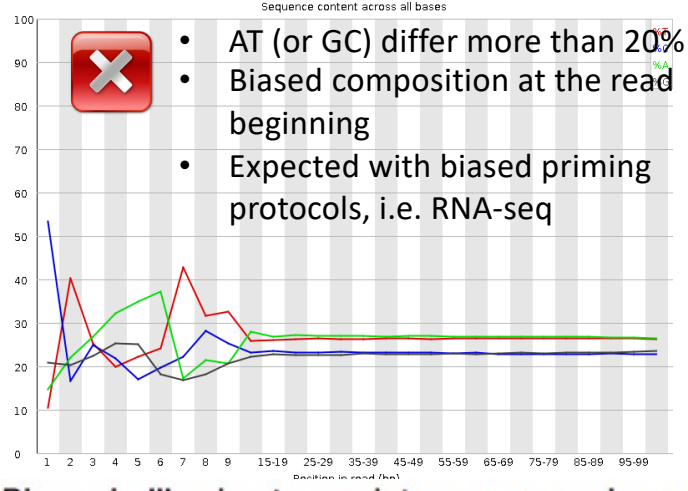
- Subset of sequences with universally low quality values

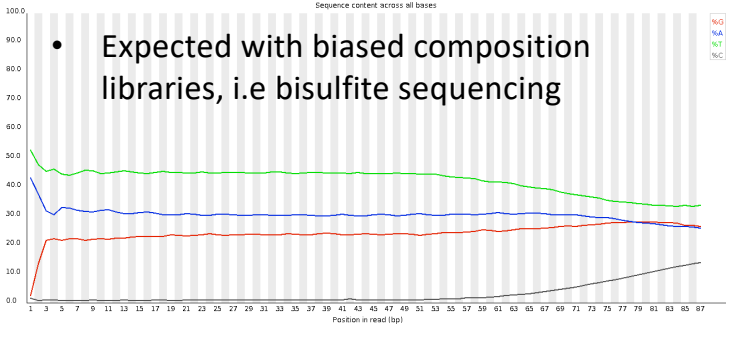


Per base sequence content - FastQC

- The portion of A, T, G, and C at each position







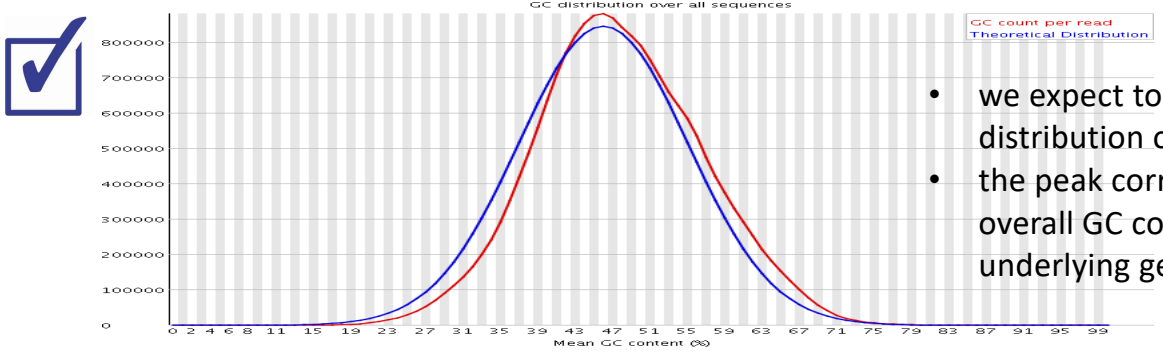
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

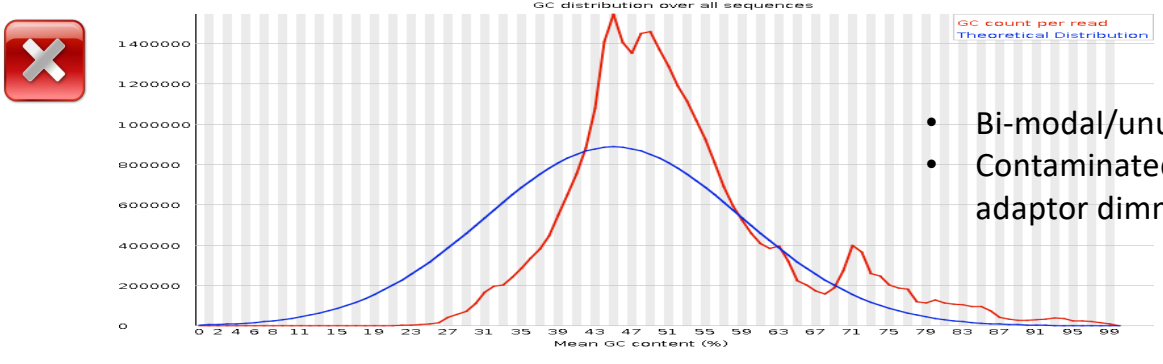
Treatment of DNA with bisulfite converts cytosine to uracil, but leaves methylated cytosine unaffected. Therefore, DNA that has been treated with bisulfite retains only methylated cytosines.

Per sequence GC content - FastQC

- Distribution of average GC in all reads



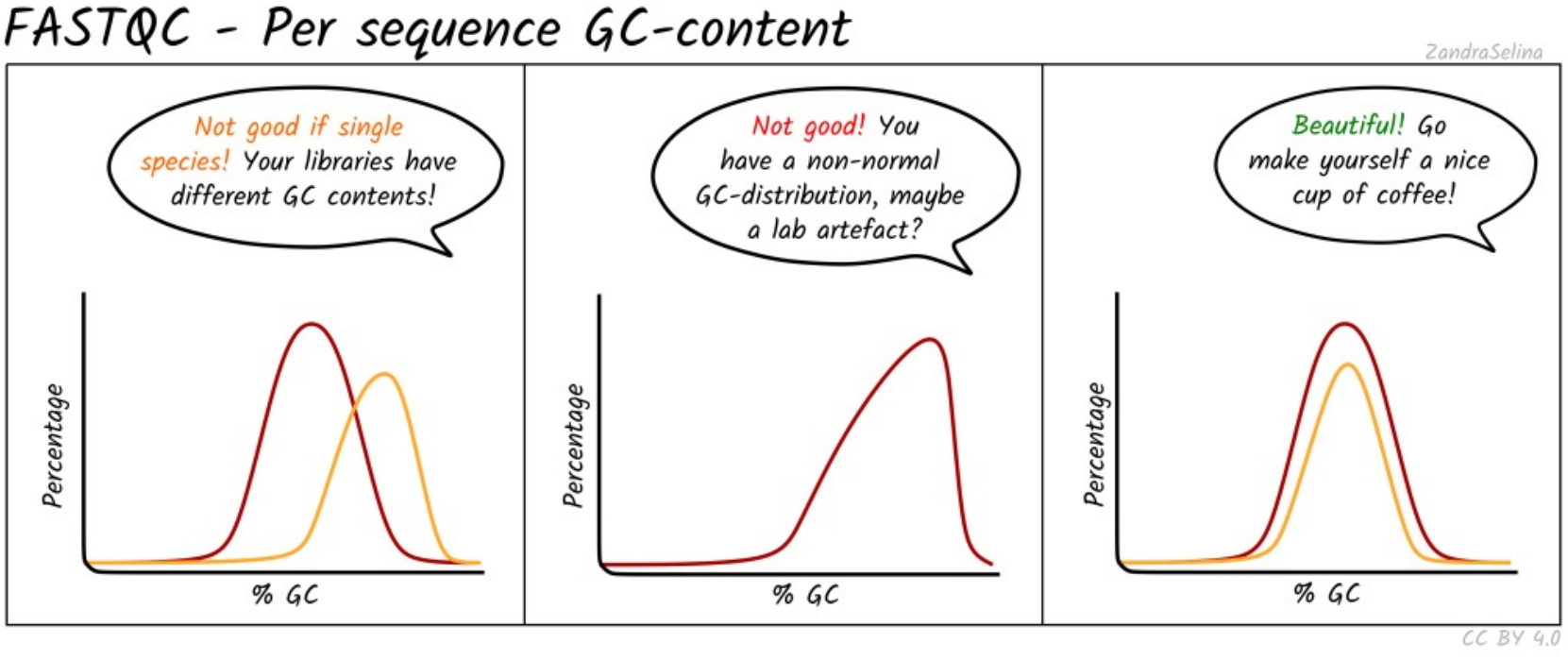
- we expect to see a roughly normal distribution of GC content
- the peak corresponds to the overall GC content of the underlying genome



- Bi-modal/unusual distribution
- Contaminated/biased subset, i.e. adaptor dimmers, rRNA etc

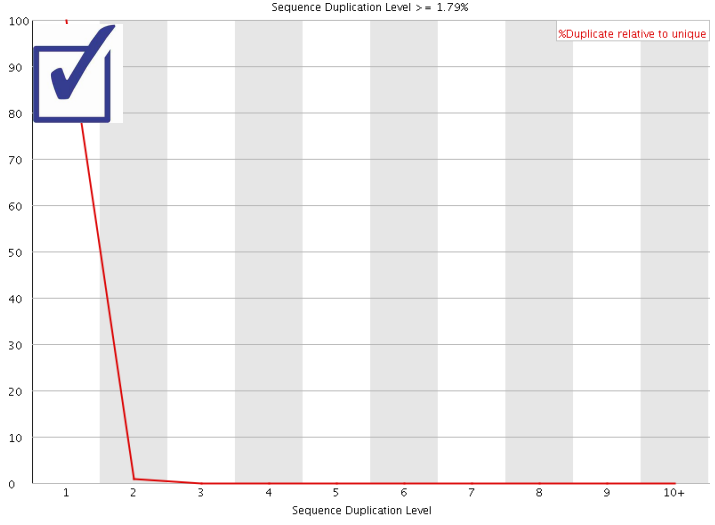
Per sequence GC content - FastQC

- Distribution of average GC in all reads

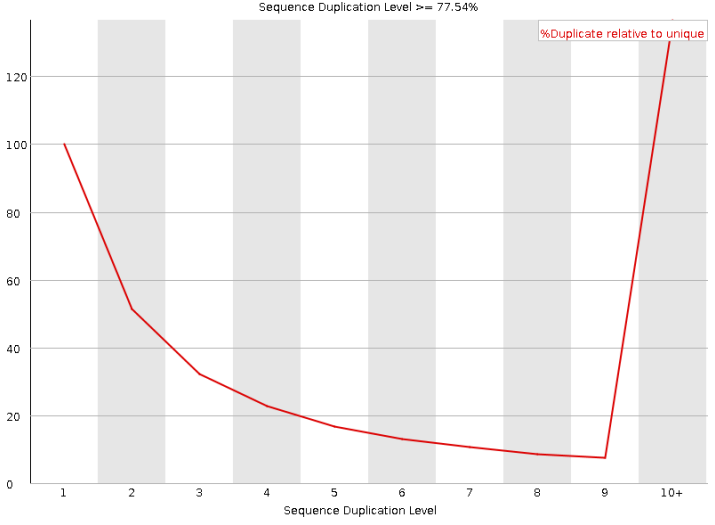


Sequence duplication - FastQC

- Relative number of sequences with different degrees of duplication



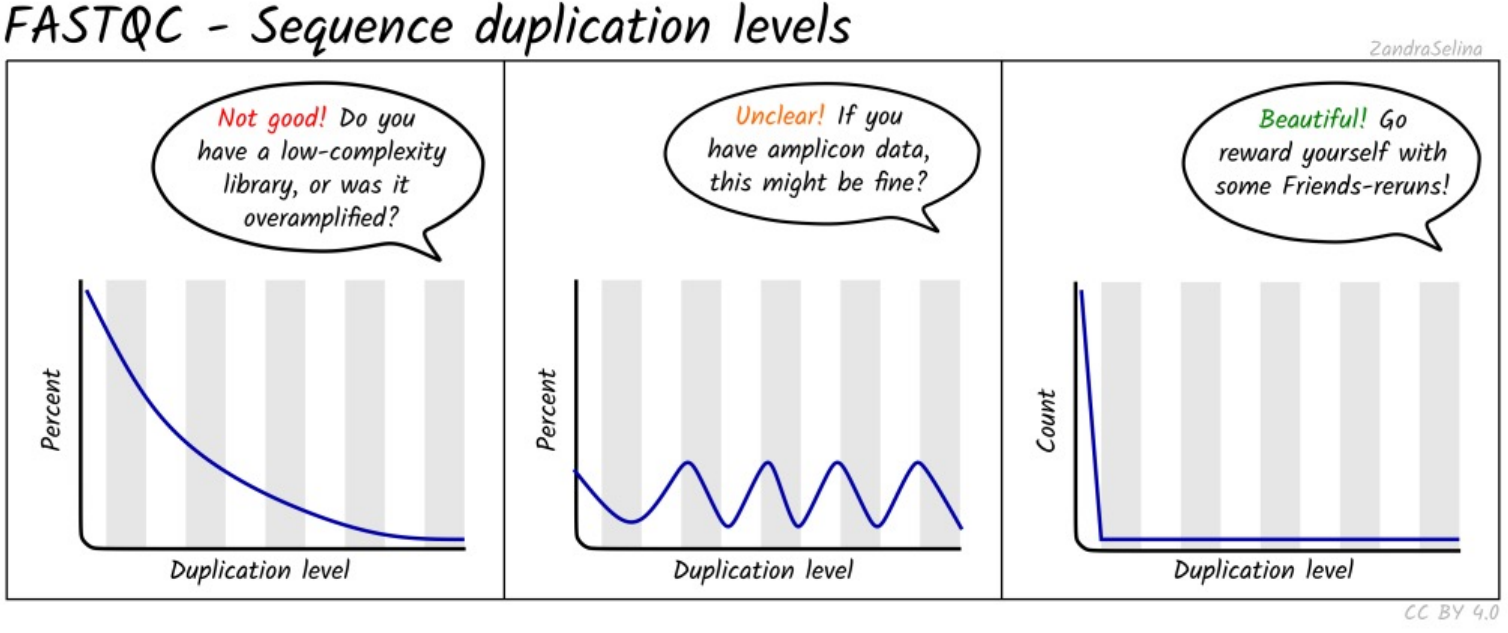
- Low level duplication is expected for a diverse library



- High level duplication: enrichment bias, saturated sequencing depth
- Normal for RNA-seq (high sequencing depth) and ChIP-seq (enriched libraries)

Sequence duplication - FastQC

- Relative number of sequences with different degrees of duplication



Overrepresented sequences - FastQC

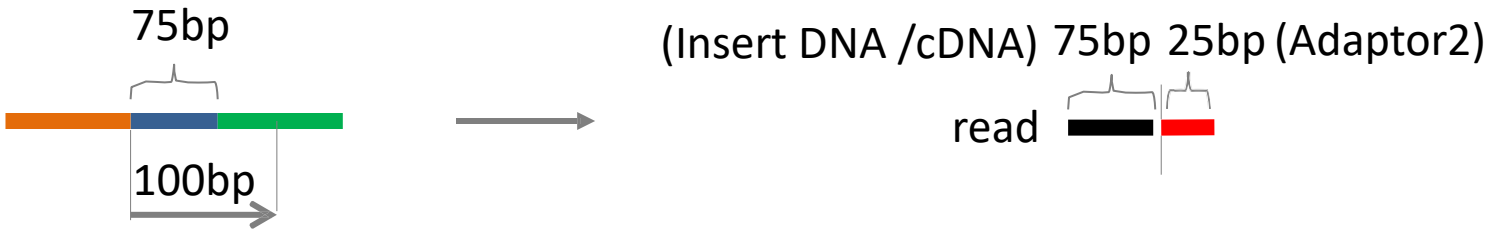
- Sequences make up >0.1 % of the total
- Compare those with a contamination database for finding contamination (i.e. adaptor dimmers)

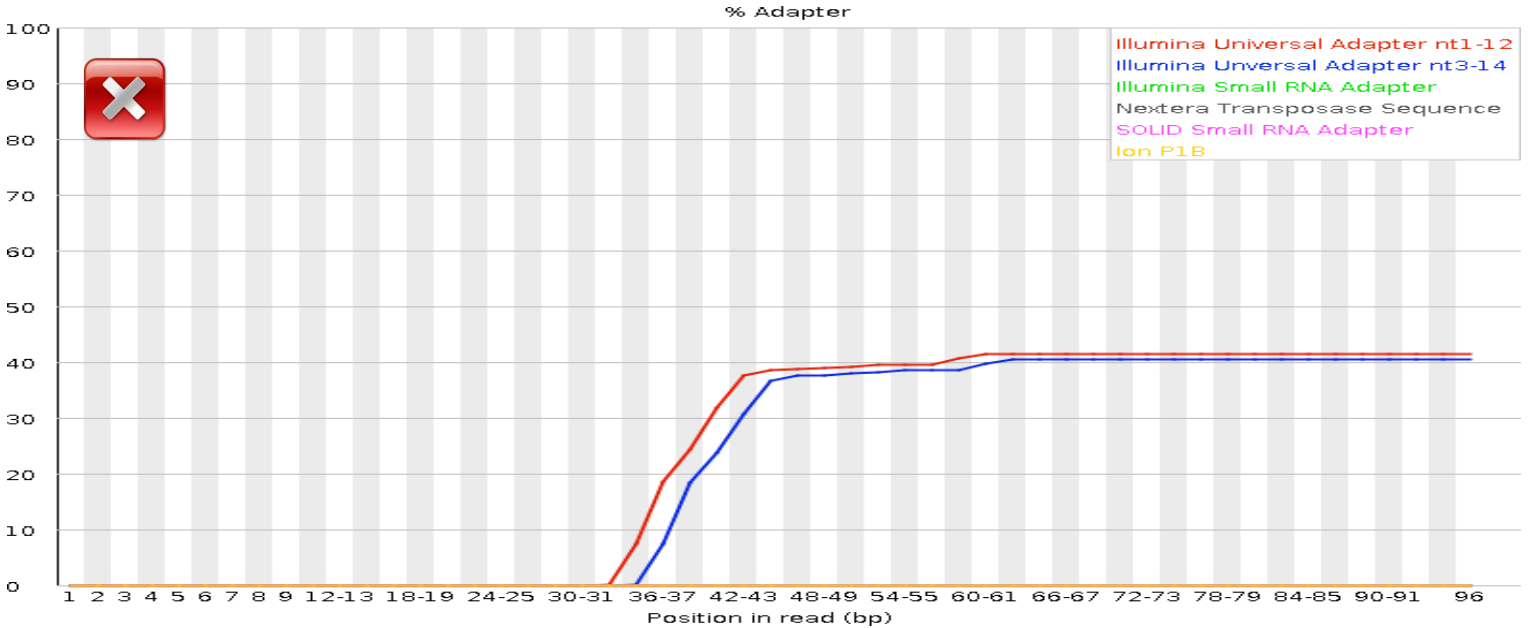
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTC	75874	1.5613887498682963	TruSeq Adapter, Index 7 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTC	7636	0.15713900010536297	TruSeq Adapter, Index 2 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTC	7539	0.1551428656095248	TruSeq Adapter, Index 5 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTC	5117	0.10530123933199874	TruSeq Adapter, Index 6 (100% over 50bp)

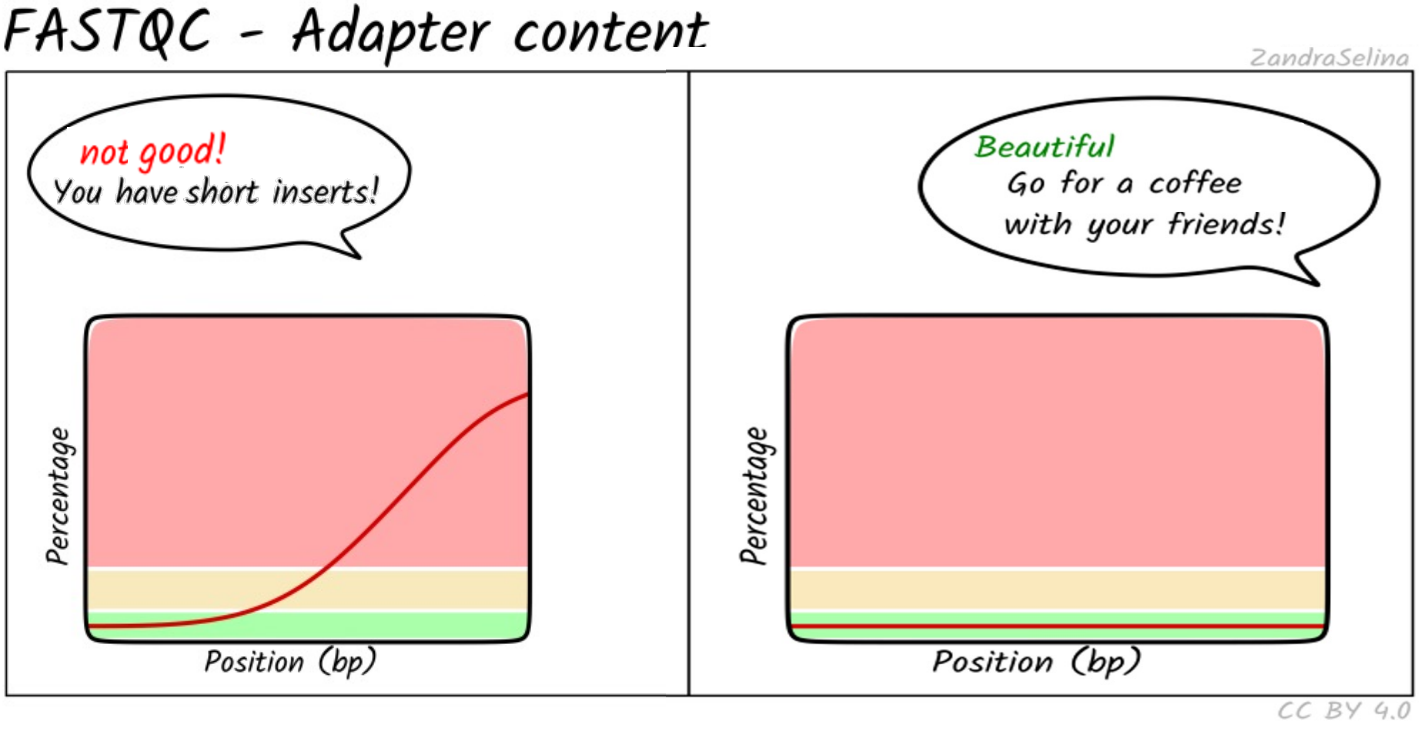
- Can be normal and biologically meaningful
  - highly expressed transcripts
  - high copy number repeats
  - Less diverse library (amplicons)

Adapter Content - FastQC



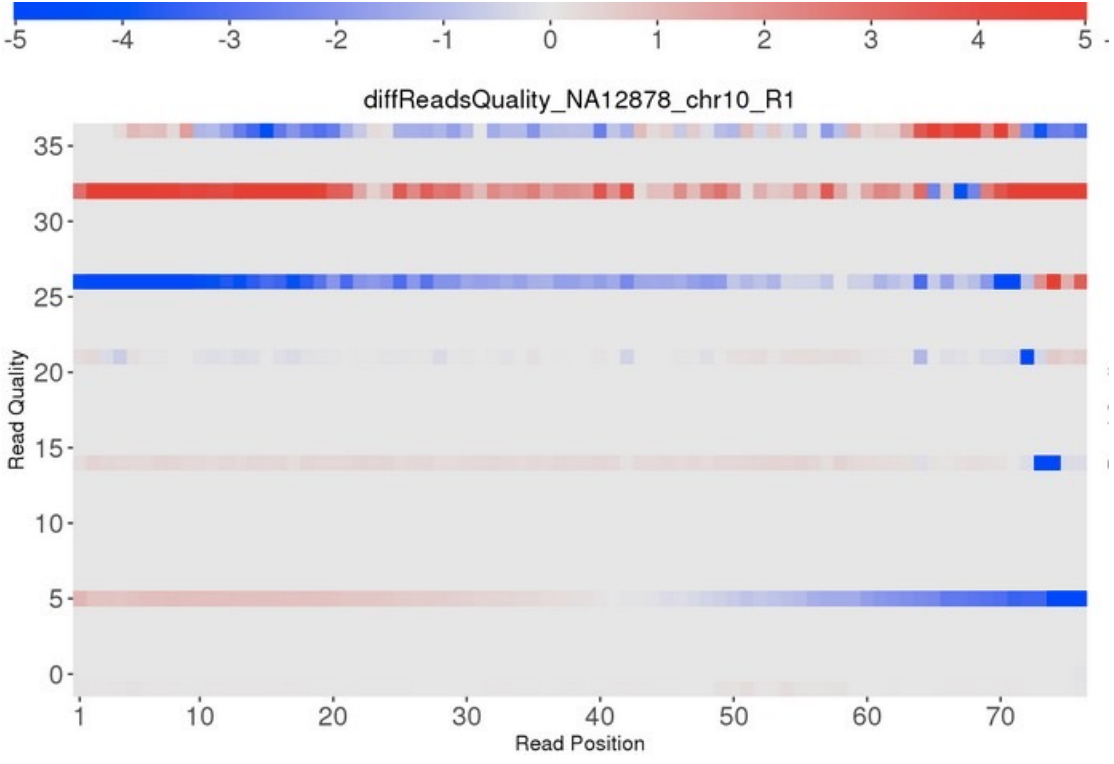


Adapter Content - FastQC



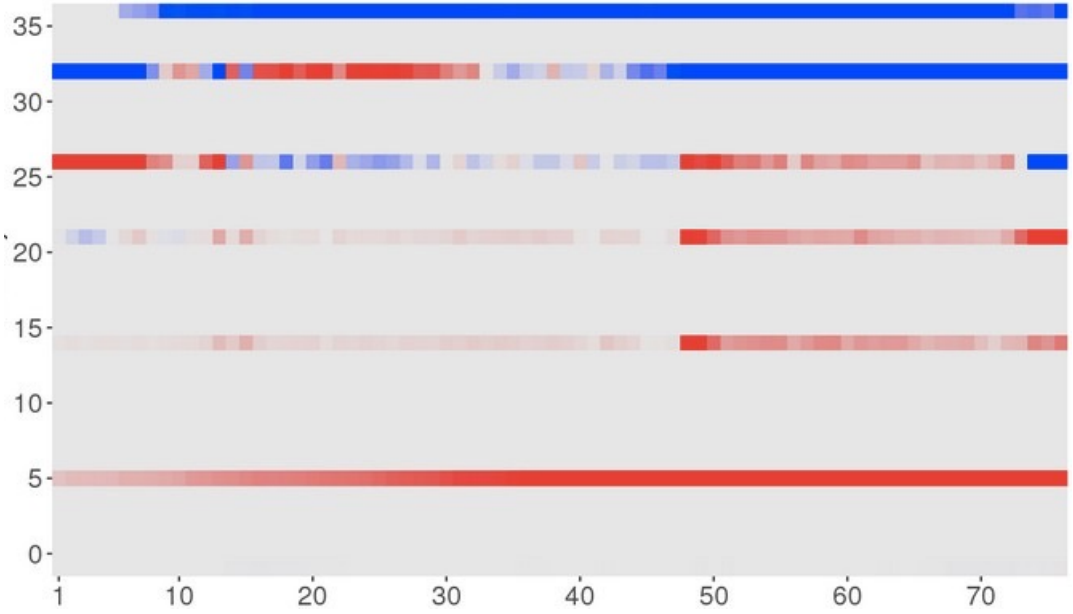
Comparative heatmap of per base phred score

- Reads in one sample /Average of reads in all samples
- Sample with better quality than average



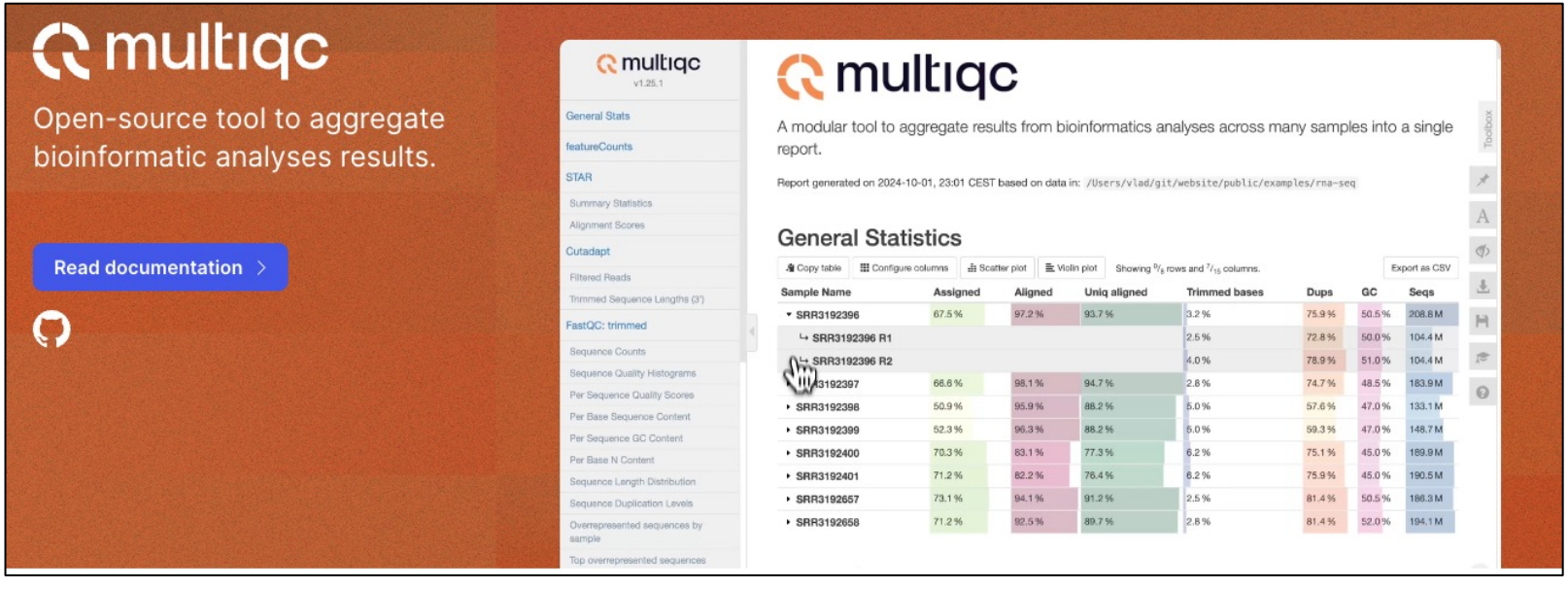
Comparative heatmap of per base phred score

- Sample with lower quality than average



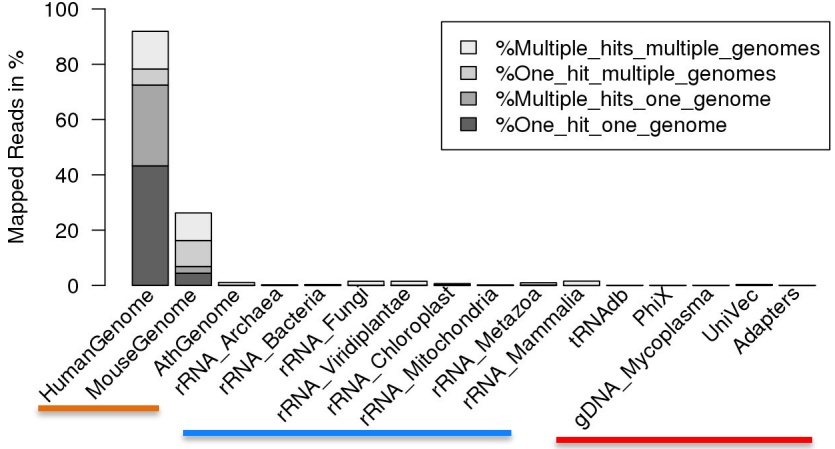
MultiQC

- https://seqera.io/multiqc/
- Aggregates bioinformatics analyses results from different simples and sources, including FastQC



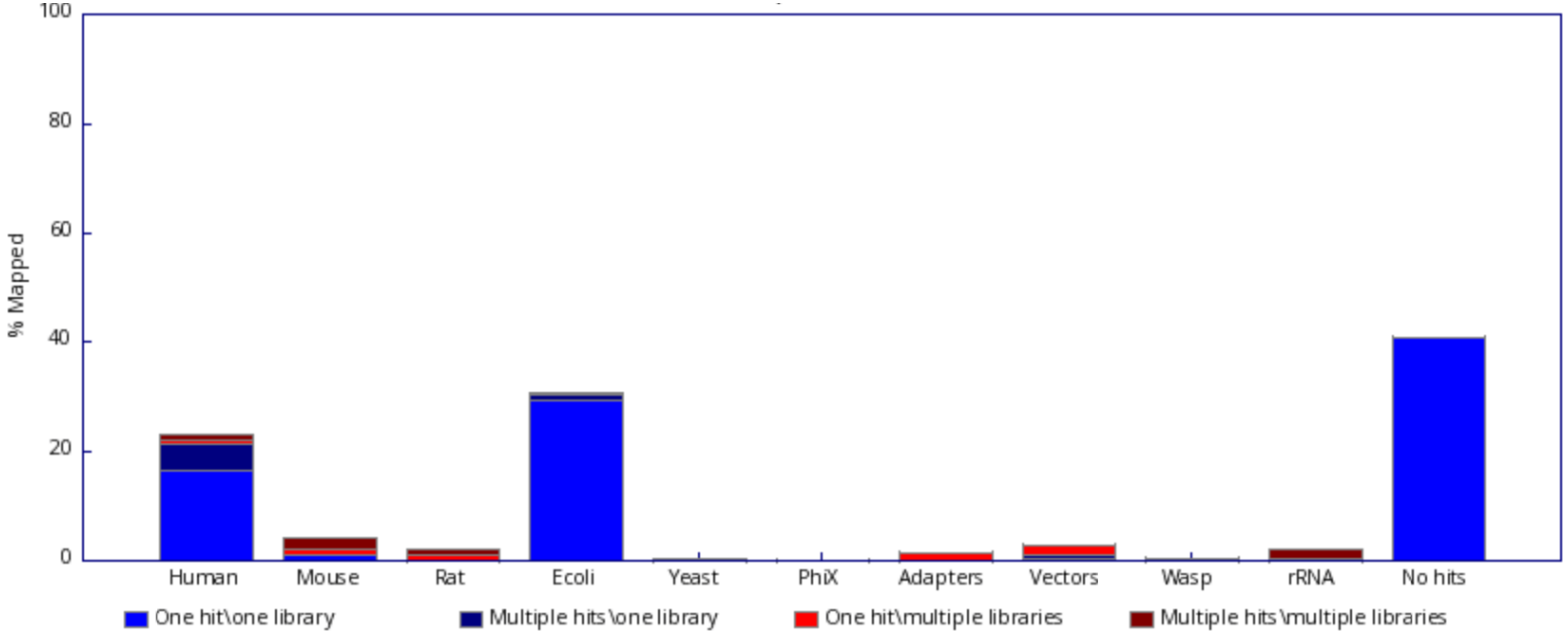
Fastqscreen – check for sample contamination

- Compare sequencing reads to databases of known sequences
- Report top matches
- In clonal sample, uniquely mapped reads should come from only a signal organism



Frequently sequenced organisms rRNA genes (Silva) Frequent contamination

Contamination Check



Sequencing data pre-processing tasks

- Trimming: remove bases from read end(s)
  - Adaptor sequence
  - Low quality bases
- Filtering: remove reads
  - Low quality reads
  - Contaminating sequences
  - Low complexity reads (repeats)
  - Short (<20bp) reads – they slow down mapping software

Tools for pre-processing sequencing data

- PRINSEQ
  - http://prinseq.sourceforge.net/
  - Quality/hard trimming, quality filtering, reformat, ...
- Trimmomatic
  - http://www.usadellab.org/cms/?page=trimmomatic
  - Adaptor trimming, quality trimming &filtering, ...
- fastp*
  - https://github.com/OpenGene/fastp
  - Adaptor trimming, quality trimming, deduplication, parallel processing
- FlexBar (FAR)
  - http://sourceforge.net/projects/theflexibleadap/
  - Flexible barcode detection and adapter removal

Summary

- Always generate quality plots for all data sets
- Interpretation of the plots need knowledge about the samples and libraries
- Trim and/or filter data if needed
  - Always trim and filter away low quality data for variant analysis

Questions ?!

Raw QC Clicker Questions