From raw images to preprocessed data

Alex Sánchez

GRUP DE RECERCA EN ESTADÍSTICA I BIOINFORMÀTICA

Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

UNIVERSITAT DE BARCELONA

Vall d'Hebron Institut de Recerca

Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

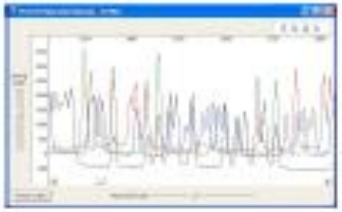
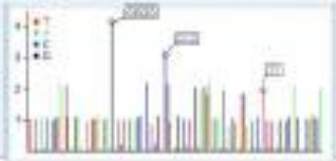
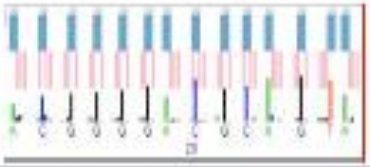
UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Outline

- What do sequencing technologies produce
  - NGS data formats. Sequence files formats
- What problems may there be?
  - Error sources and error models
- Detecting errors
  - Quality control of NGS data
- Adjusting errors
  - Post-processing of sequence data
- Conclusions and perspectives

Data and data formats

One input, many possible outputs …

Technology	Raw Data	Primary Data	Data Formats
CE (Sanger)Sequencing by Synthesis Chain termination, electrophoretic separation with fluorescent detection			Binary files; one / sample AB1, SCF common ESD, ZTR, RCF, SRF, fasta, quality values, others
454 Sequencing by SynthesisPyrosequenicng, multiple rounds of light detection			Binary and text files; one group / plate section, "flow space," SFF, fasta, quality values, SRF
Illumina GA Sequencing by Synthesis Reversible terminators, multiple rounds of fluorescent detection			Binary and text files; one group / plate section, fasta, fastq, quality values, SRF
SOLiD Sequencing by Ligation Dibase encoding, multiple rounds of extension and fluorescent detection		CSFasta >Myseq1 G012310223…	Binary and text files; one group / plate section, "color space," csfasta, quality values, SRF

Typical Sequence Data Formats(I)

- All Sequence formats are **ASCII** text containing sequence ID, Quality Scores, Annotation details, comments, and other descriptions about sequence

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

- Formats are designed to **hold sequence** data and other information about sequence

- **FastA** format (everybody knows about it)
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).

Why so many formats?

- **Created based on the information required for each step of analysis**

- **Efficient Data & time management**

Types of sequence file formats

- **Raw Sequence files**
- **Co-ordinate files**
- **Parameter files**
- **Annotation files**
- **Metadata files**

- Raw sequence data formats
  - SFF
  - Fasta, csfasta
  - Qual file
  - Fastq

- **Each Data formats vary in the information they contain**

Read output formats

- **454**
- **Solexa/Illumina**
- **SOLiD**

454 output formats

Standard flowgram *format*

Read name

.sff

Flowed nucleotide*

Flow no.

Flow value (light signal strength)

Quality scores

oedipus.imr.no - PuTTY

FR3GPDN01CVFQV	786	a	0.10	
FR3GPDN01CVFQV	787	c	0.09	
FR3GPDN01CVFQV	788	g	0.08	
FR3GPDN01CVFQV	789	t	0.10	
FR3GPDN01CVFQV	790	a	0.10	
FR3GPDN01CVFQV	791	c	0.09	
FR3GPDN01CVFQV	792	g	0.08	
FR3GPDN01CVFQV	793	t	0.10	
FR3GPDN01CVFQV	794	a	0.10	
FR3GPDN01CVFQV	795	c	0.09	
FR3GPDN01CVFQV	796	g	0.07	
FR3GPDN01CVFQV	797	t	0.14	
FR3GPDN01CVFQV	798	a	0.11	
FR3GPDN01CVFQV	799	c	0.08	
FR3GPDN01CVFQV	800	g	0.09	
FR3GPDN01A5TLK	1	c	0.57	37
FR3GPDN01A5TLK	2	a	0.00	
FR3GPDN01A5TLK	3	c	1.01	37
FR3GPDN01A5TLK	4	g	0.04	
FR3GPDN01A5TLK	5	t	0.02	
FR3GPDN01A5TLK	6	a	1.02	37
FR3GPDN01A5TLK	7	c	0.00	
FR3GPDN01A5TLK	8	g	2.10	37,37
FR3GPDN01A5TLK	9	T	2.91	37,37,37
FR3GPDN01A5TLK	10	A	0.01	
FR3GPDN01A5TLK	11	C	0.00	
FR3GPDN01A5TLK	12	G	1.01	37
FR3GPDN01A5TLK	13	T	2.15	37,37
FR3GPDN01A5TLK	14	A	0.00	
FR3GPDN01A5TLK	15	C	0.00	
FR3GPDN01A5TLK	16	G	1.01	37
FR3GPDN01A5TLK	17	T	0.36	
FR3GPDN01A5TLK	18	A	0.94	37
FR3GPDN01A5TLK	19	C	0.00	
FR3GPDN01A5TLK	20	G	1.01	37
FR3GPDN01A5TLK	21	T	0.10	
FR3GPDN01A5TLK	22	A	1.07	37
FR3GPDN01A5TLK	23	C	0.00	
FR3GPDN01A5TLK	24	G	0.16	
FR3GPDN01A5TLK	25	T	0.88	39
FR3GPDN01A5TLK	26	A	0.19	
FR3GPDN01A5TLK	27	C	0.00	
FR3GPDN01A5TLK	28	G	1.91	39,39
FR3GPDN01A5TLK	29	T	1.07	39
FR3GPDN01A5TLK	30	A	0.04	

.fna

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
TAACAATCGAGGCGAAGTCCCGTGAGAAGCTGTTTACTTCTCATGATCACACAGGCGCTG
GCTCCTCAGGCAAACAGGTACGTCTACGATAGGTTCCATGAAAAGTCCAAGTTTGGCCGA
GCTCTGGCTCCTTTTGACGCACAGTGGAACTTCCTTGTTCACGGAAATTGCA
```

.qual

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
28 35 28 27 34 27 26 25 25 28 31 24 26 27 32 25 27 27 32 28 6  28 27 27 27 27 33 26
27 26 27 27 34 30 10 27 25 34 27 28 22 28 27 26 26 27 27 26 27 25 22 23 28 27 18 20
23 27 27 29 21 25 25 34 26 27 24 25 32 24 22 33 28 7  25 20 30 22 28 27 24 25 28 28
28 27 28 26 27 25 23 33 25 35 28 34 27 27 25 28 38 34 21 8  25 27 34 27 31 23 22 36
32 17 29 21 32 24 24 27 28 19 27 28 26 34 28 23 25 35 28 38 34 21 8  26 26 27 25 27
21 28 28 27 27 34 27 34 27 25 30 21 34 26 33 25 26 35 28 20 28 25 34 27 37 33 15 33
25 23 28 25
```

Illumina output formats

.seq.txt

```
File: s_1_0001_seq.txt
1      1      137    689    AACATAATGTGTTCACTGAGAACACATTGCACTCAA
1      1      87     649    TATTGCAACTTGTTTAATTTTTCATGCCATTATCA
1      1      121    642    TACATGATTTGCATTTGGTAAATAGCTACTTTTTAT
1      1      6      591    C...T.......T.......................
```

.prb.txt

```
40 -40 -40 -40     40 -40 -40 -40     -40 40 -40 -40     40 -40 -40 -40
-40 -40 -40 40     40 -40 -40 -40     40 -40 -40 -40     -40 -40 -40 40
-40 -40 40 -40     -40 -40 -40 40     -40 -40 40 -40     -40 -40 -40 40
-40 -40 -40 40     -40 40 -40 -40     40 -40 -40 -40     -40 40 -40 -40
-40 -40 -40 40     -40 -40 40 -40     40 -40 -40 -40     -40 -40 40 -40
40 -40 -40 -40     40 -40 -40 -40     -40 40 -40 -40     40 -40 -40 -40
-40 40 -40 -40     40 -40 -40 -40     -40 -40 -40 40     -40 -40 -40 40
-40 -40 40 -40     -40 40 -40 -40     40 -40 -40 -40     -40 40 -40 -40
-40 -40 -40 40     -40 40 -40 -40     40 -40 -40 -40     37 -37 -40 -40
```

Illumina FASTQ

(ASCII – 64 is Illumina score)

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Qseq

(ASCII – 64 is Phred score)

```
HWUSI-EAS521    2    1    26    0
        .GGCAGCGGGCAGGGCGAGCCAATGCGTGT
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2    1    26    0
        .GGGAAATAGCTTTCACGCCTTTAGATAATTT
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2    1    26    0
        .ATCTTTAACAGACCAAGACTGGGCCACAAG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2    1    26    0
        .GGCTGGGTATGAGTCAGGGGGCTCCAGAG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Illumina single line format

```
>1-1-137-689    AACATAATGTTCACTGAGAACACATTGCACTCAA    U0
>1-1-87-649     TATTGCAACTTGTTTAATTTTTCATGCCATTATCA    U1
>1-1-121-642    TACATGATTTGCATTTGGTAAATAGCTACTTTTAT    U0
```

SCARF

Solexa Compact ASCII Read *Format*

```
HWI-EAS102_3:6:1:897:791:AATGTCAATCTGAGTT...TTT:40 40 40 40 40..
HWI-EAS102_3:6:1:930:291:AATGTACTTTTTCTAA...CTA:40 29 14 17 16..
HWI-EAS102_3:6:1:944:665:AATCGATCCCCTTCCC...TTC:40 34 33 40 40..
```

Typical Sequence Data formats (2)

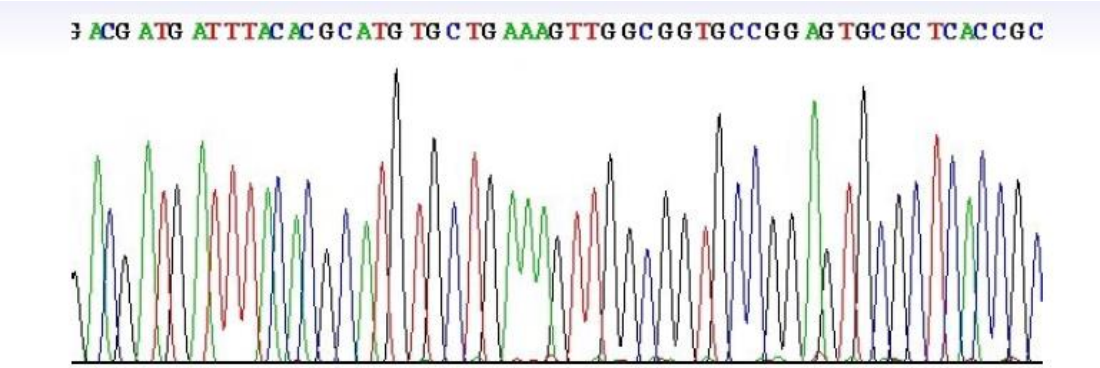
- **FastQ** format (http://maq.sourceforge.net/fastq.shtml)
  - First is the sequence (like Fasta but starting with "@")
  - Then "+" and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
    - Different quality encode variants
- Nearly all downstream analysis take **FastQ** as input sequence

The fastq format

- A FASTQ file normally uses four lines per sequence.
  - Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
  - Line 2 is the raw sequence letters.
  - Line 3 begins with a '+' character and is*optionally* followed by the same sequence identifier (and any description) again.
  - Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.
    - Different encodings are in use
    - Sanger format can encode a Phred quality score from 0 to 93 using ASCII 33 to 126

```
@Seq description
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Phred Scores



- Sequencing systems use to assign quality scores to each peak
- Phred scores provide log(10)-transformed error probability values: If *p* is probability that the base call is wrong the Phred score is

$$Q = .10 \cdot \log_{10} p$$

  - score = 20 corresponds to a 1% error rate
  - score = 30 corresponds to a 0.1% error rate
  - score = 40 corresponds to a 0.01% error rate
- The base calling (A, T, G or C) is performed based on Phred scores.
- Ambiguous positions with Phred scores <= 20 are labeled with N.

SOLiD output format(s)

CSFASTA

color-space *sequence* reads in a fasta *format*

```
@ERR000451.1 VAB_S0103_20080915_542_14_17_70_F3
T3302323020310210322333002030023300 1
+
T%245719<.6353&:%0#%&%2(--27*%&%,
```

- These reads can be **retained and analyzed in color-space** by software
- The **Format Conversion Tool** offers options for cleaning of the CSFASTA files

Common (“standard”) format for read alignments: **Alignment/Assembly Format**

SAM is a plain-textual format of the alignments (in a flavour that is similar but different to GFF or BED). It is luckily extensible.

BAM is a dedicated binary format including the compressed SAM. It enables fast access to data without having to “unzip” the whole file.

For the typically large data, BAM is currently the most recommended and most “standard” format.

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

Sequencers & Sequence Assembly Packages

454

Solexa/Illumina

SOLiD

Maq
BWA
Bowtie
RMAP
Eland
SOAP
SOAP2
MOSAIK
SOCS
PatMaN
ZOOM
PerM
RazerS
segemehl
MPSCAN
BFAST
Lastz
BLAT

Celera
Newbler
Velvet
Euler
SOAP denovo

Formats for Genome/Gene annotation

BED format
GFF format
BioXSD (XML)

```
##gff-version 3
ctg123 . operon          1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  Parent=mrna0001
ctg123 . mRNA           10000 15000  .  +  .  ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon           10000 12000  .  +  .  Parent=mrna0002
ctg123 . exon           14000 15000  .  +  .  Parent=mrna0002
```

Points to remember on Data Formats

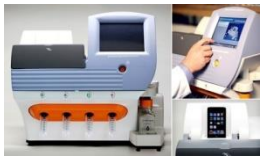
- **For base-call data, “standard” FASTQ (Sanger, Phred)**
- **For read alignments, SAM/BAM/MAQ format**
- **For annotation results (e.g. GFF or BED format)**

All platforms have errors

Illumina

SoLID/ABI-Life

Roche 454



Ion Torrent

1. Removal of low quality bases/ Low complexity regions
2. Removal of adaptor sequences
3. Homopolymer-associated base call errors (3 or more identical DNA bases) causes higher number of (artificial) frameshifts

Illumina artefacts

- under represented GC rich regions
  - PCR
  - Sequencing
- GGC/GCC motif is associated with low quality and mismatches
- Low quality reads < 20% phred score

Need for QC & Preprocessing

QC analysis of sequence data is extremely important for meaningful downstream analysis

- **To analyze problems in quality scores/ statistics of sequencing data**
- **To check whether further analysis with sequence is possible**
- **To remove redundancy (filtering)**
- **To remove low quality reads from analysis**
- **To remove adapter contamination**

Highly efficient and fast processing tools are required to handle large volume of datasets

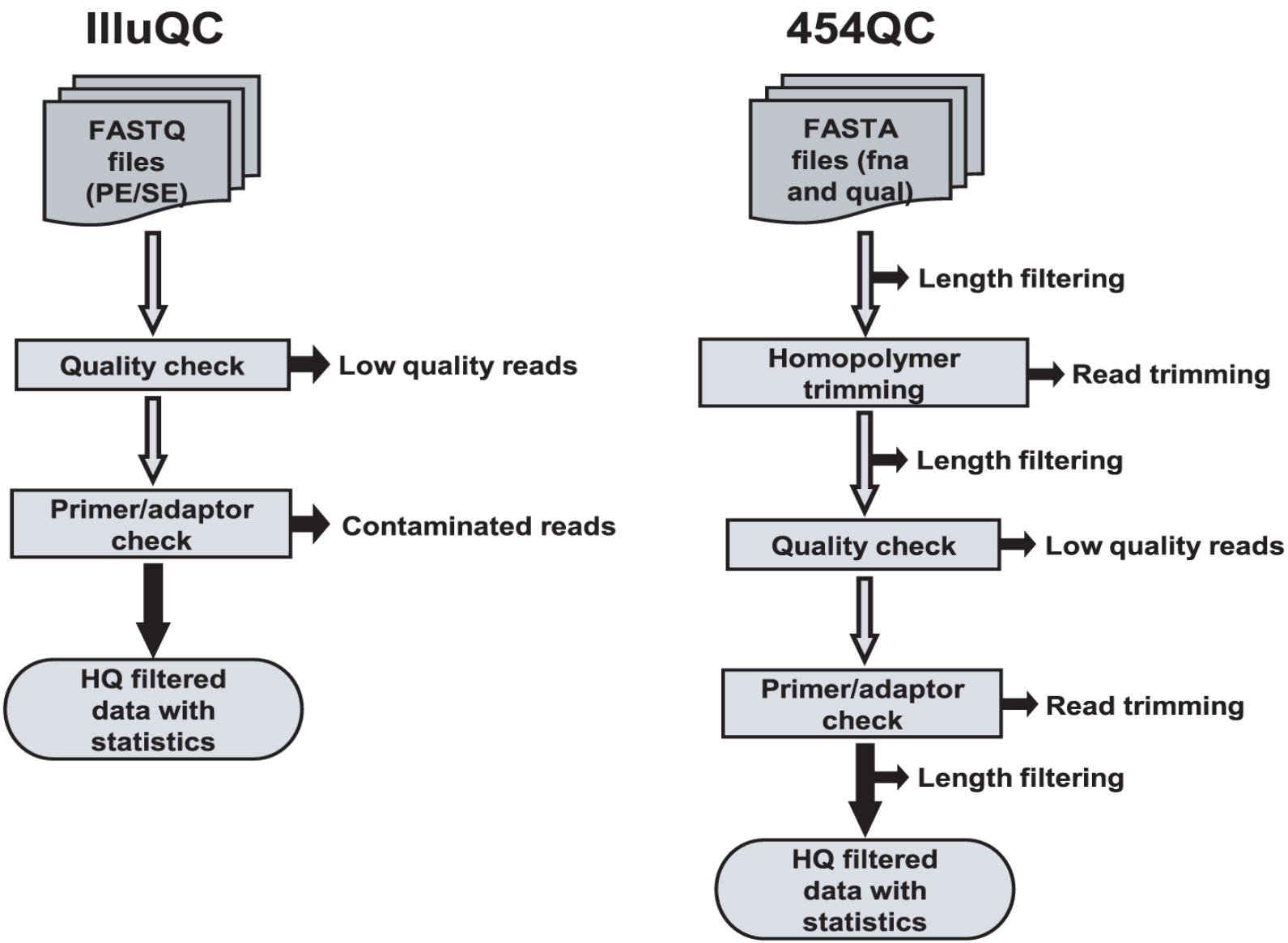
Need for QC & Preprocessing

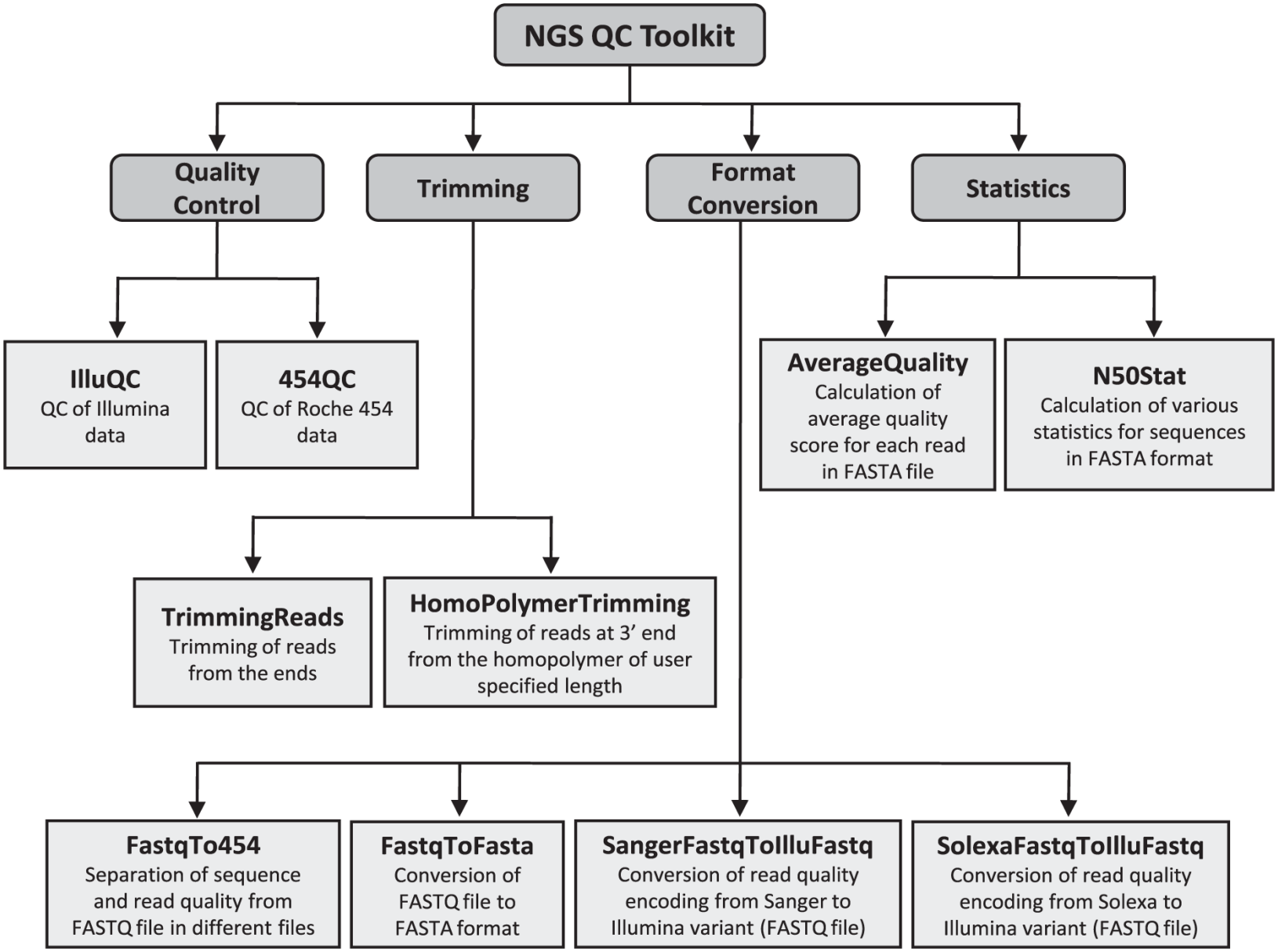
- The quality of data is very important for various downstream analyses, such as sequence assembly, single nucleotide polymorphisms identification
- Most of the programs available for downstream analyses do not provide the utility for quality check and filtering of NGS data before processing

NGS QC Toolkit & FastQC

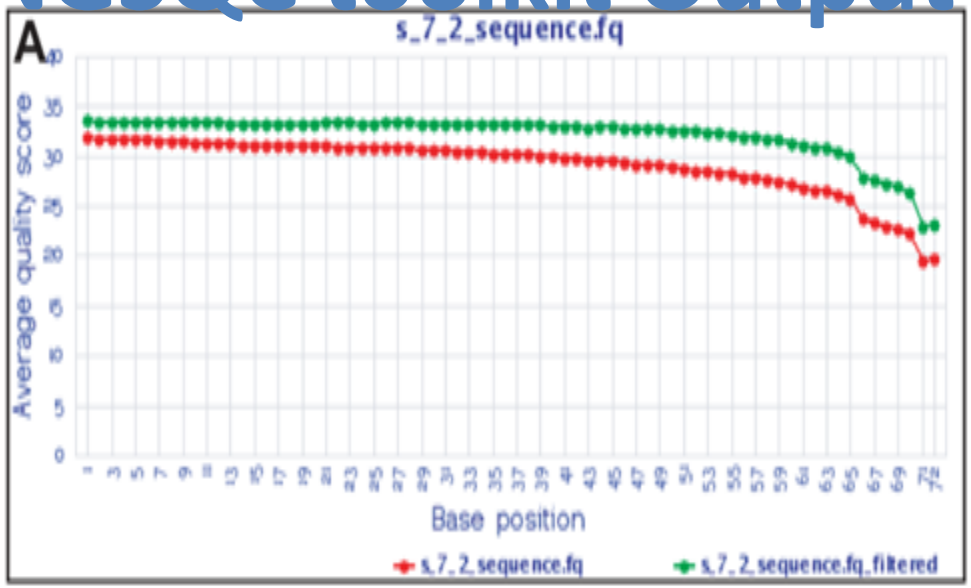
- NGS QC Toolkit is for **quality check** and **filtering of high-quality read**
- This toolkit is a standalone and open source application **freely available** at **http://www.nipgr.res.in/ngsqctoolkit.html**
- Application have been implemented in **Perl programming language**
- QC of sequencing data generated using **Roche 454 and Illumina platforms**
- Additional tools to aid QC : (sequence **format converter** and **trimming tools**) and analysis (**statistics tools**)

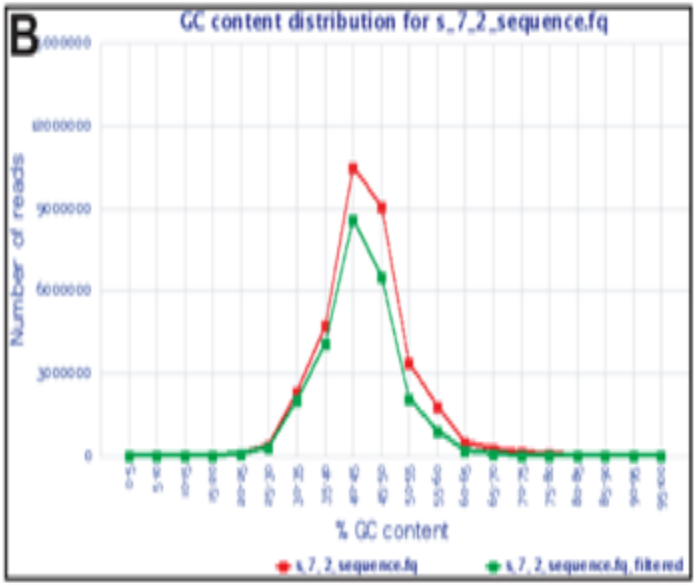
FastQC can be used only for preliminary analysis

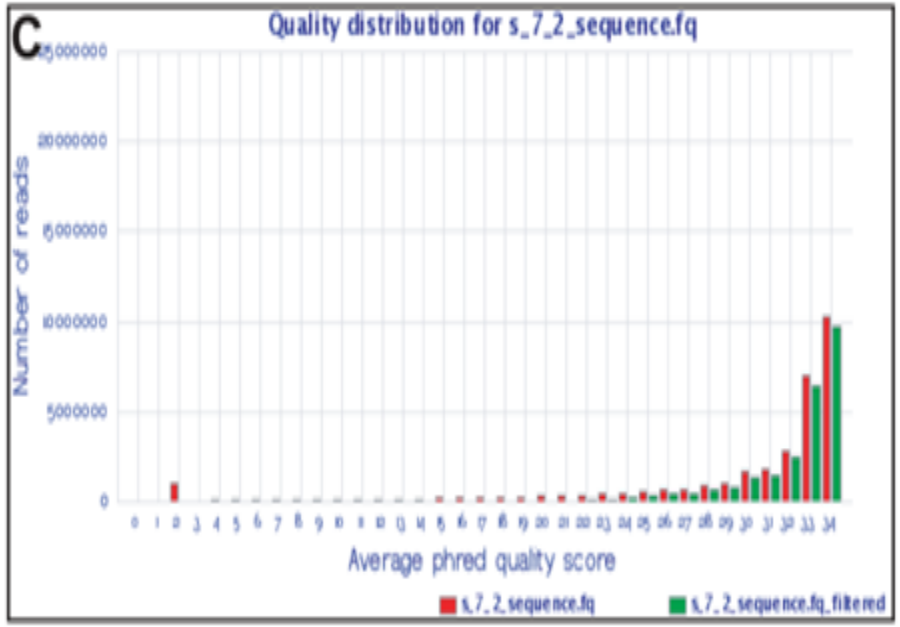

IlluQC
FASTQ files (PE/SE)
Quality check
Low quality reads
Primer/adaptor check
Contaminated reads
HQ filtered data with statistics
454QC
FASTA files (fna and qual)
Length filtering
Homopolymer trimming
Read trimming
Length filtering
Quality check
Low quality reads
Primer/adaptor check
Read trimming
Length filtering
HQ filtered data with statistics

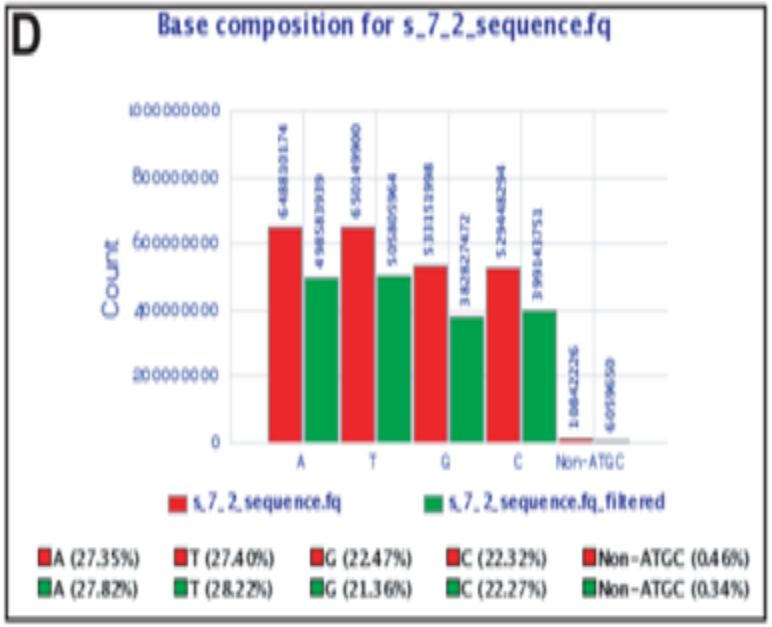

NGS QC Toolkit
Quality Control
Trimming
Format Conversion
Statistics
IlluQC
QC of Illumina data
454QC
QC of Roche 454 data
AverageQuality
Calculation of average quality score for each read in FASTA file
N50Stat
Calculation of various statistics for sequences in FASTA format
TrimmingReads
Trimming of reads from the ends
HomoPolymerTrimming
Trimming of reads at 3' end from the homopolymer of user specified length
FastqTo454
Separation of sequence and read quality from FASTQ file in different files
FastqToFasta
Conversion of FASTQ file to FASTA format
SangerFastqToIlluFastq
Conversion of read quality encoding from Sanger to Illumina variant (FASTQ file)
SolexaFastqToIlluFastq
Conversion of read quality encoding from Solexa to Illumina variant (FASTQ file)

NGSQC toolkit Output

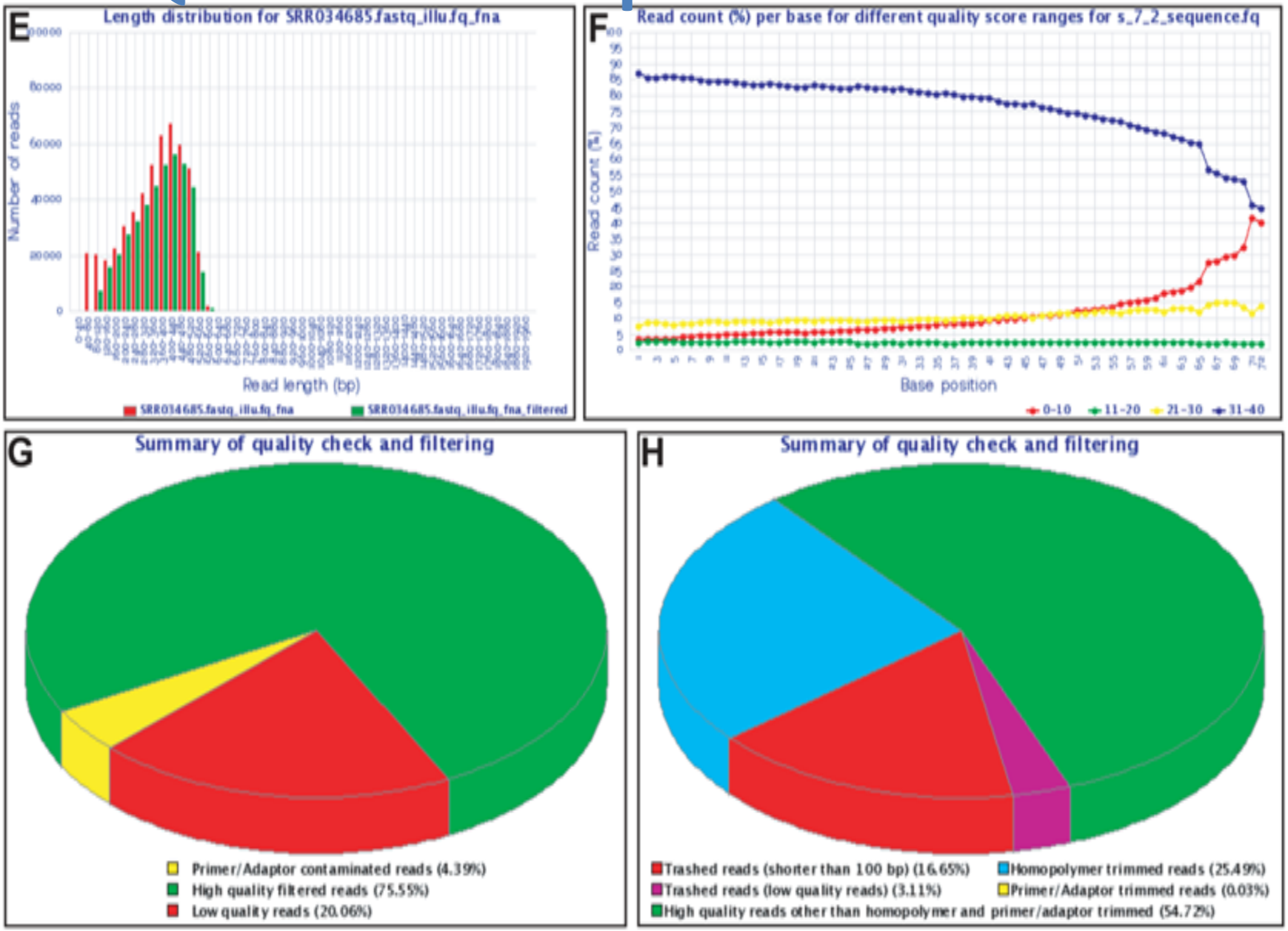








NGSQC toolkit Output



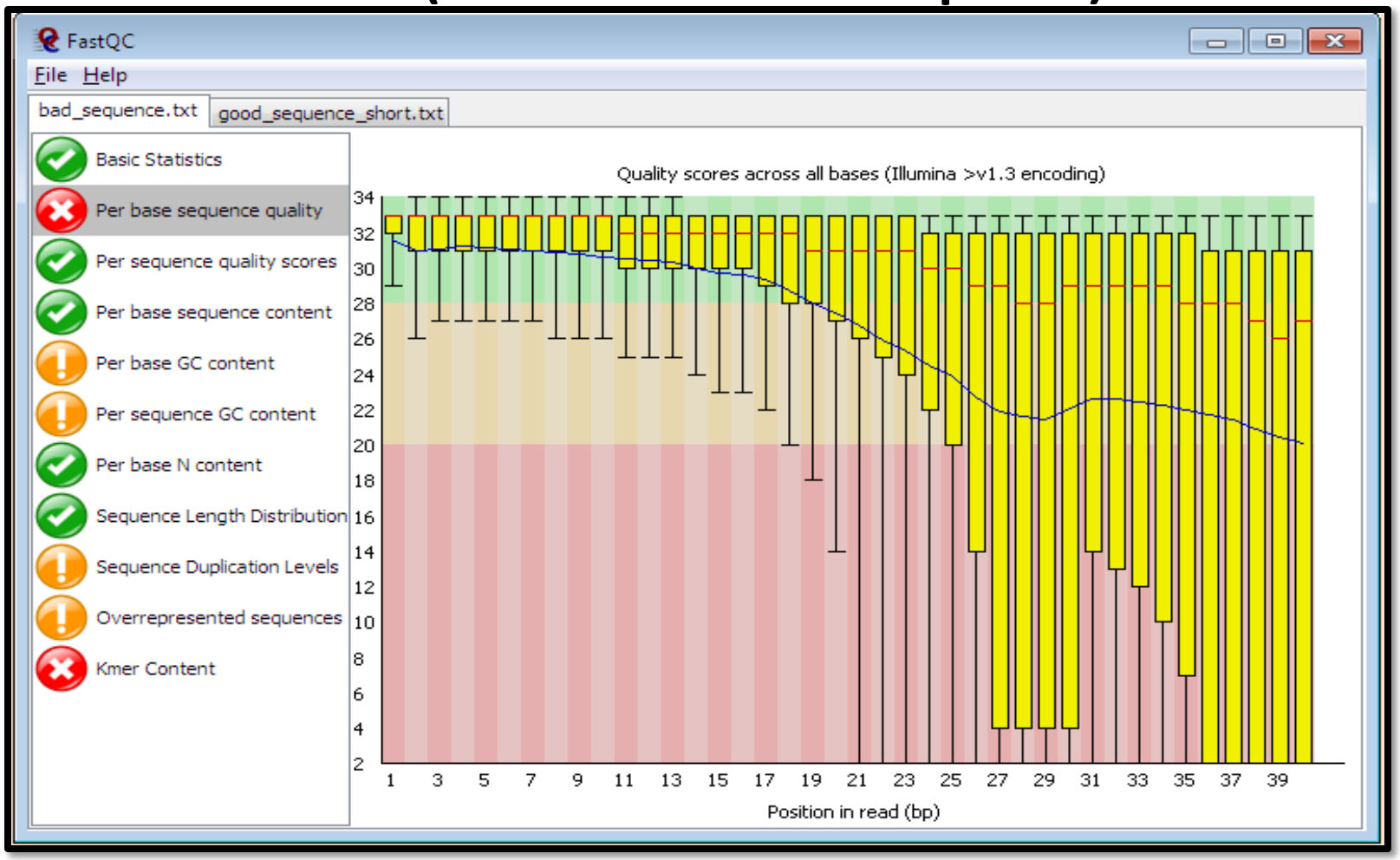
Comparison - QC tools

Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17[1]	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12[1]	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ[2]	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No[3]	No	Yes	Yes	No	Yes[4]	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No[5]	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No[6]	Yes	No	Yes	No	No	Yes
Compatability with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No[7]	No	Yes	Yes	No[7]	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png	-	BLAST, NCBI nr database

FastQC

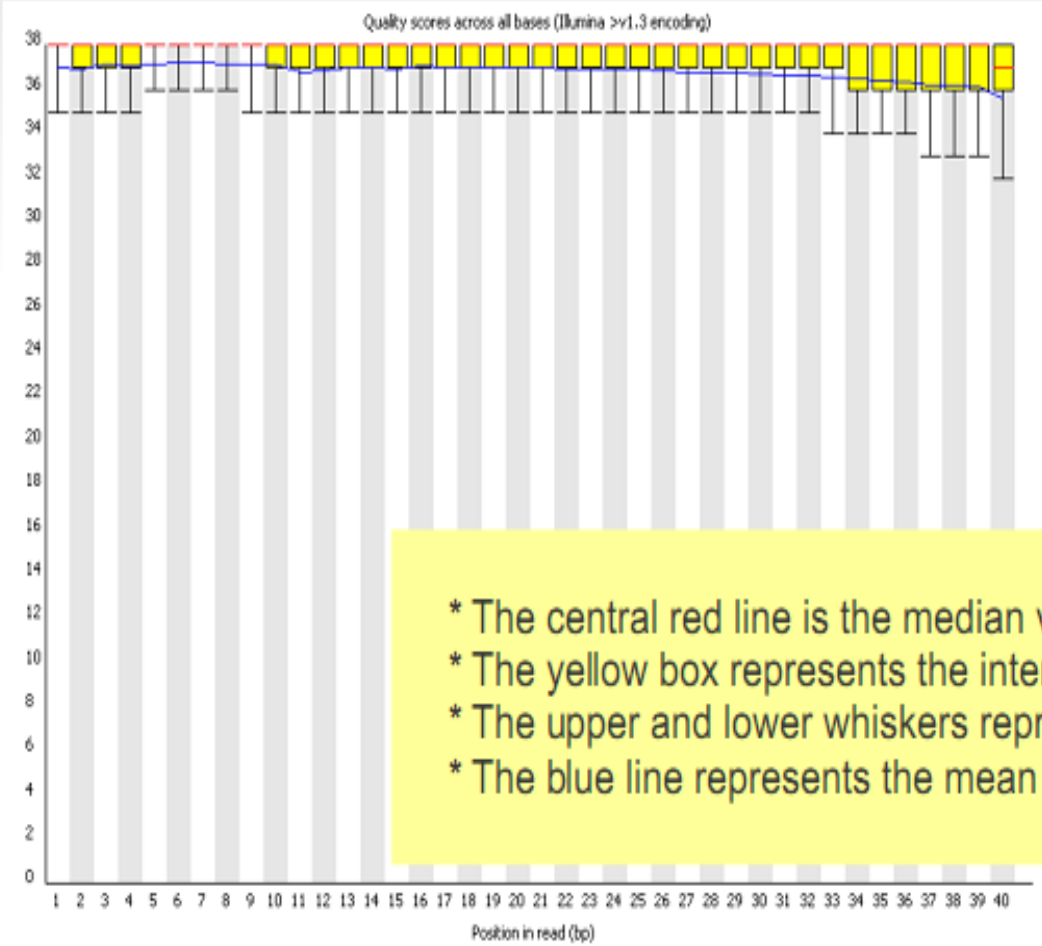
- Basic statistics
- Quality- Per base position
- Per Sequence Quality Distribution
- Nucleotide content per position
- Per sequence GC distribution
- Per base GC distribution
- Per base N content
- Length Distribution
- Overrepresented/ duplicated sequences
- K-mer content

FastQC (Box-Whisker plot)



Y axis- Quality Score
X axis- Base position

2. Quality- Per base position



Good data

- Consistent
- High quality along the read

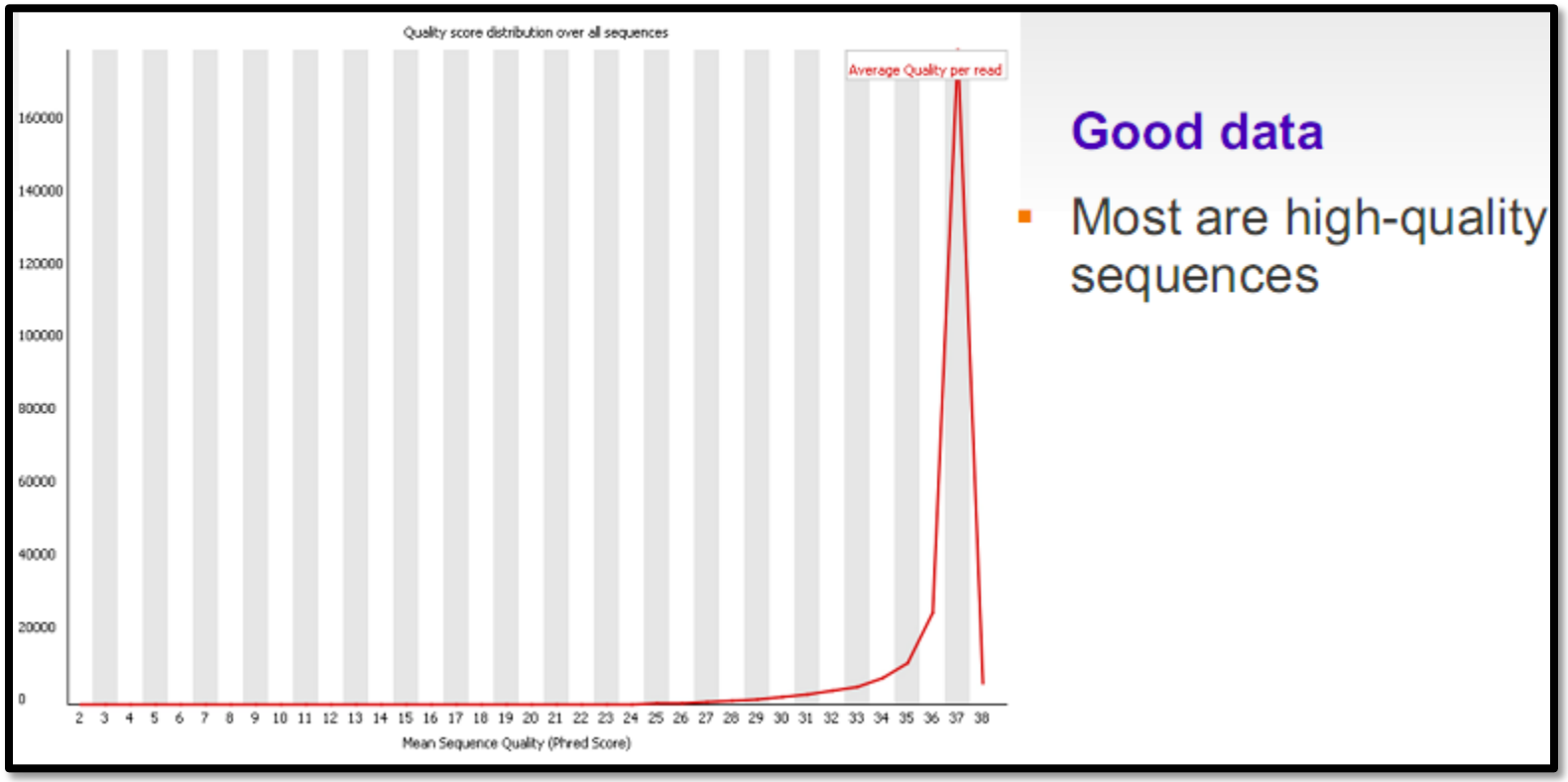
* The central red line is the median value
* The yellow box represents the inter-quartile range (25-75%)
* The upper and lower whiskers represent the 10% and 90% points
* The blue line represents the mean quality

2. Quality- Per base position

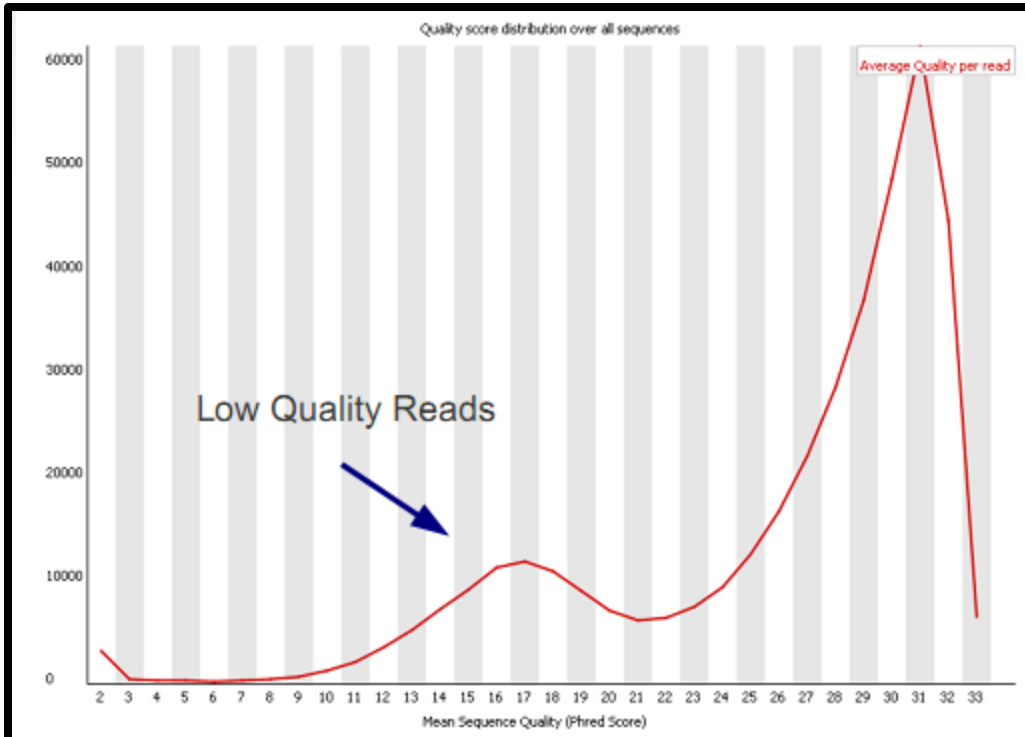
Bad data

- High variance
- Quality decrease with length

3.Per Sequence Quality Distribution



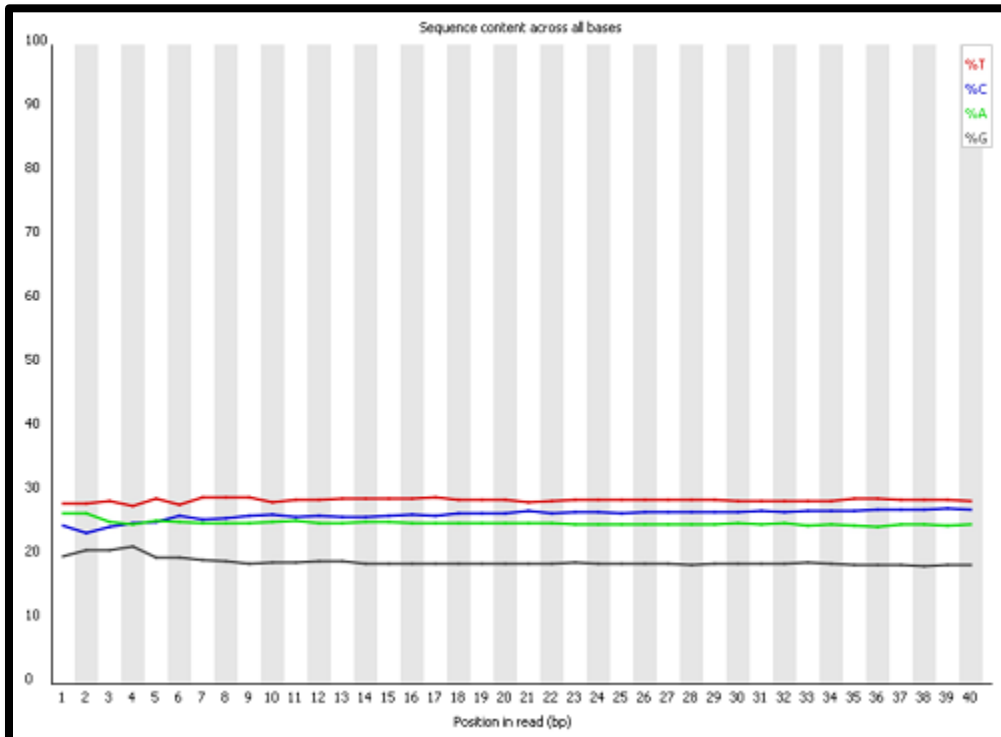
3. Per Sequence Quality Distribution



Bad data

- Not uniform distribution

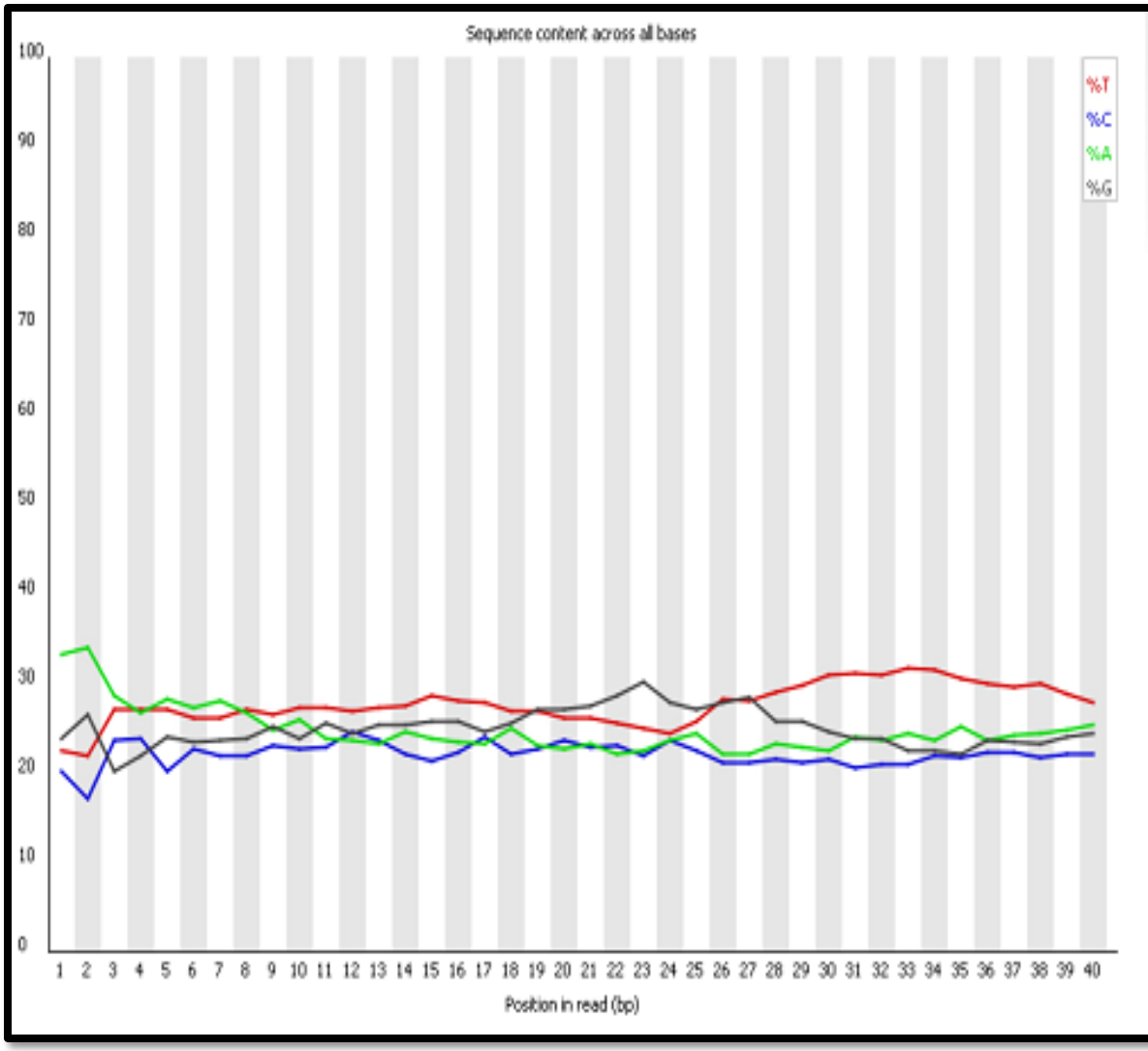
4.Nucleotide content per position



Good data

- Smooth over length
- Organism dependent (GC)

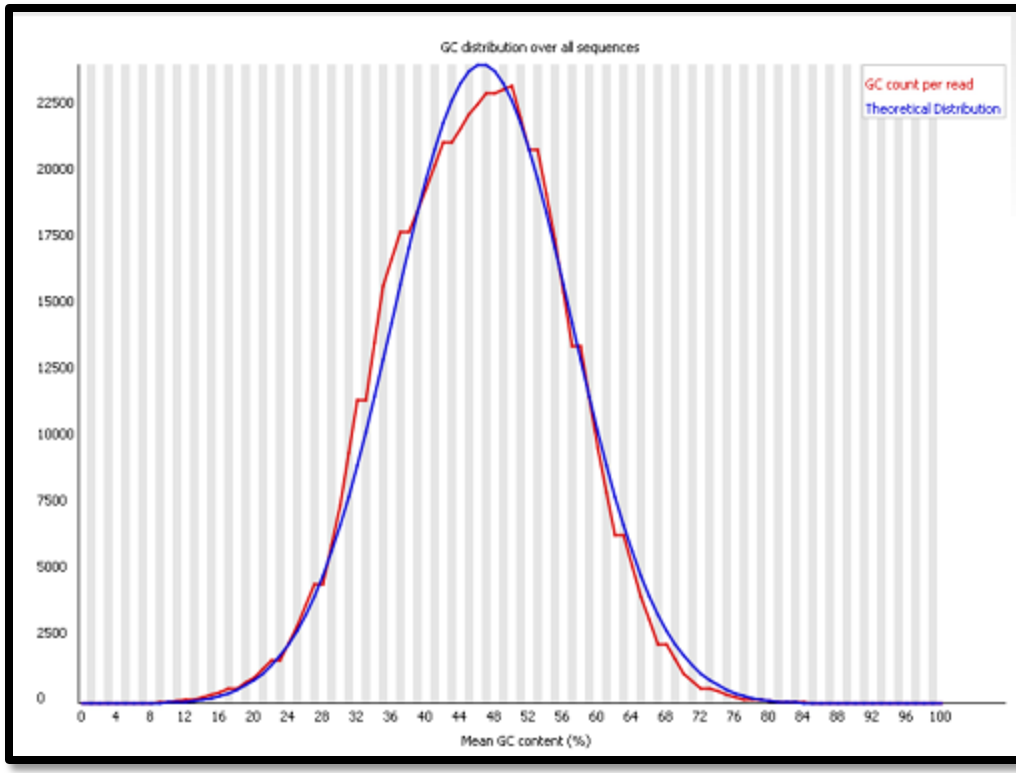
4. Nucleotide content per position



Bad data

- Sequence position bias

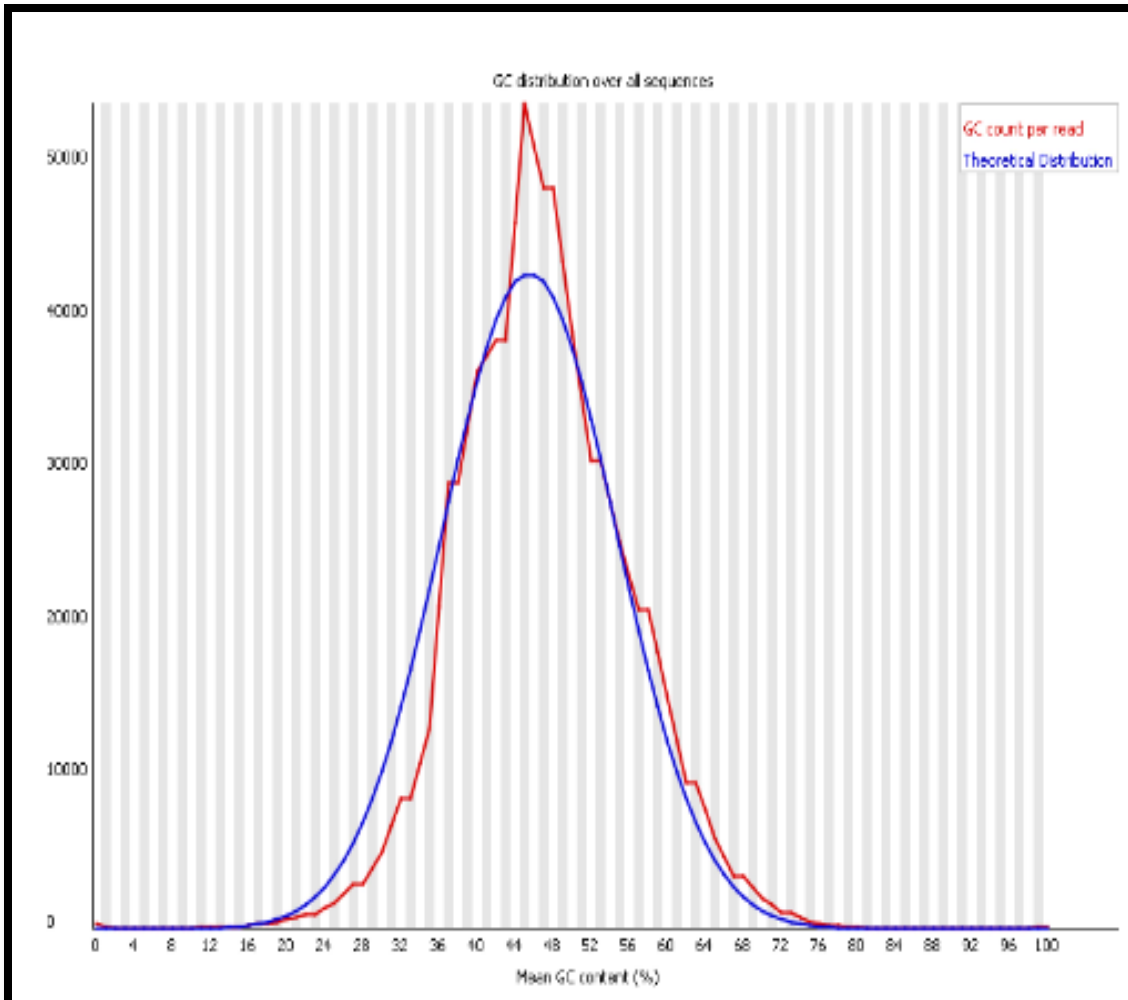
5.Per sequence GC distribution



Good data

- Fits with the expected
- Organism dependent

5.Per sequence GC distribution

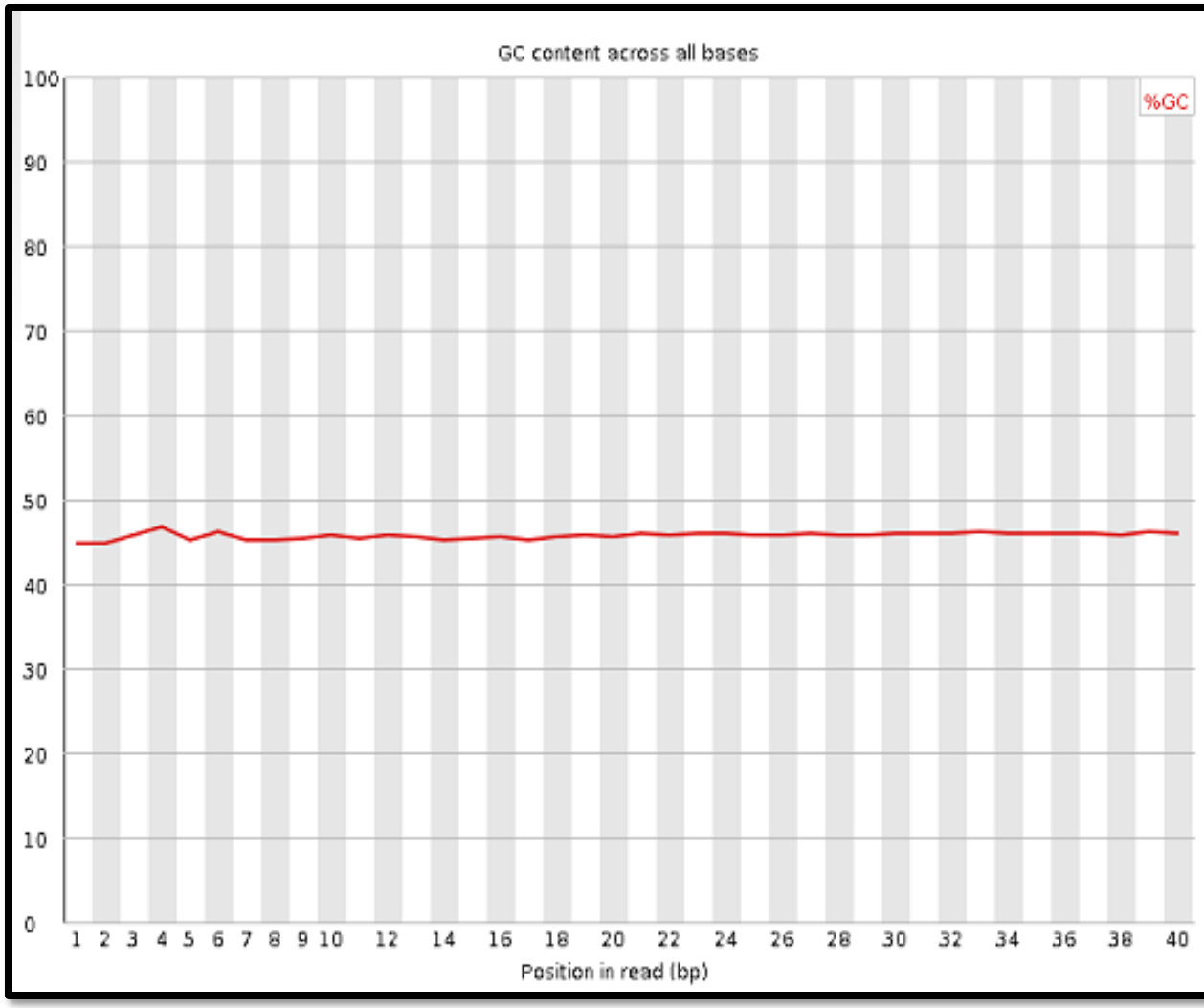


Bad data

- It does not fit with expected
- Organism dependent

Library contamination?

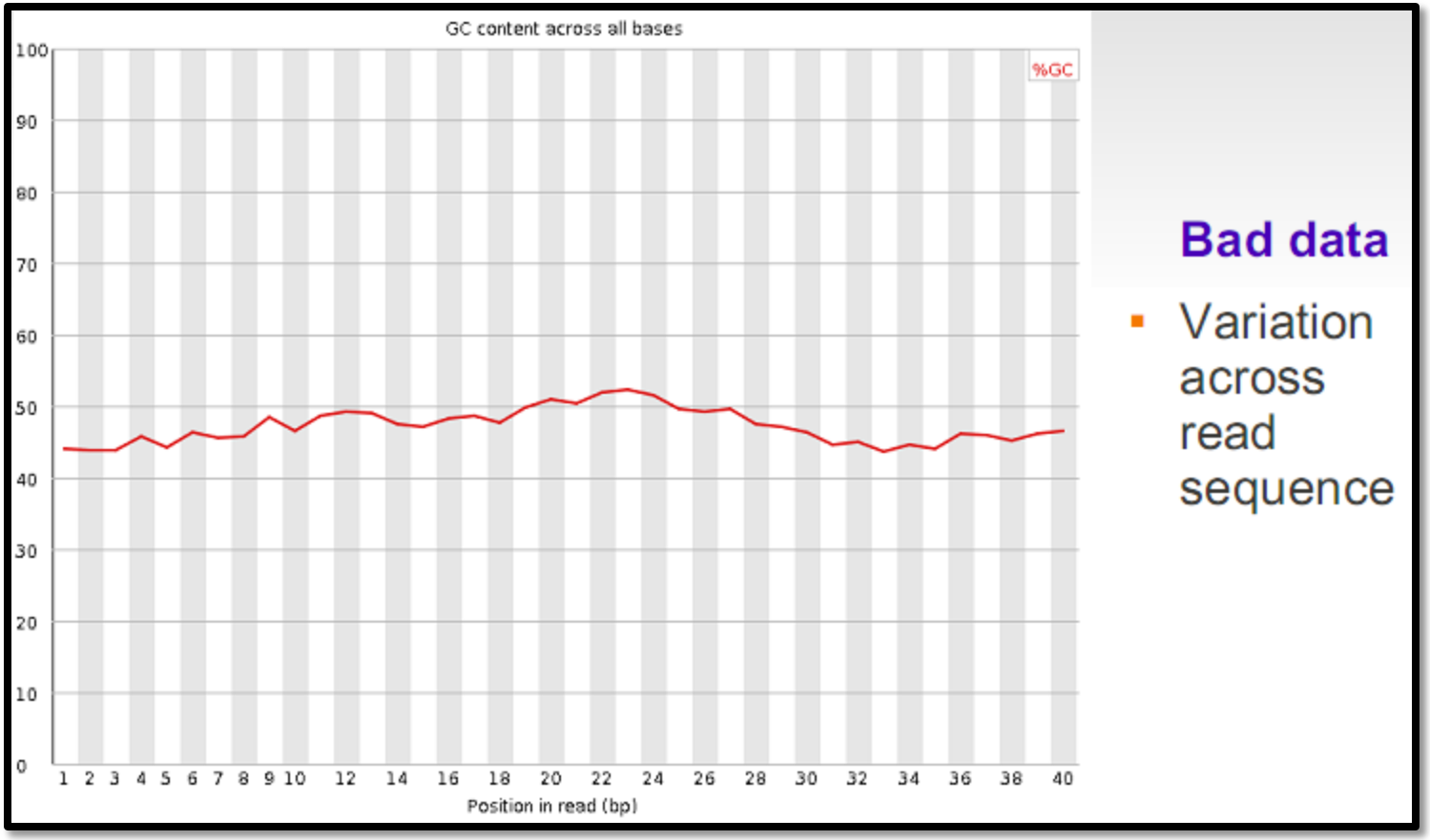
6. Per base GC distribution



Good data

- No variation across read sequence

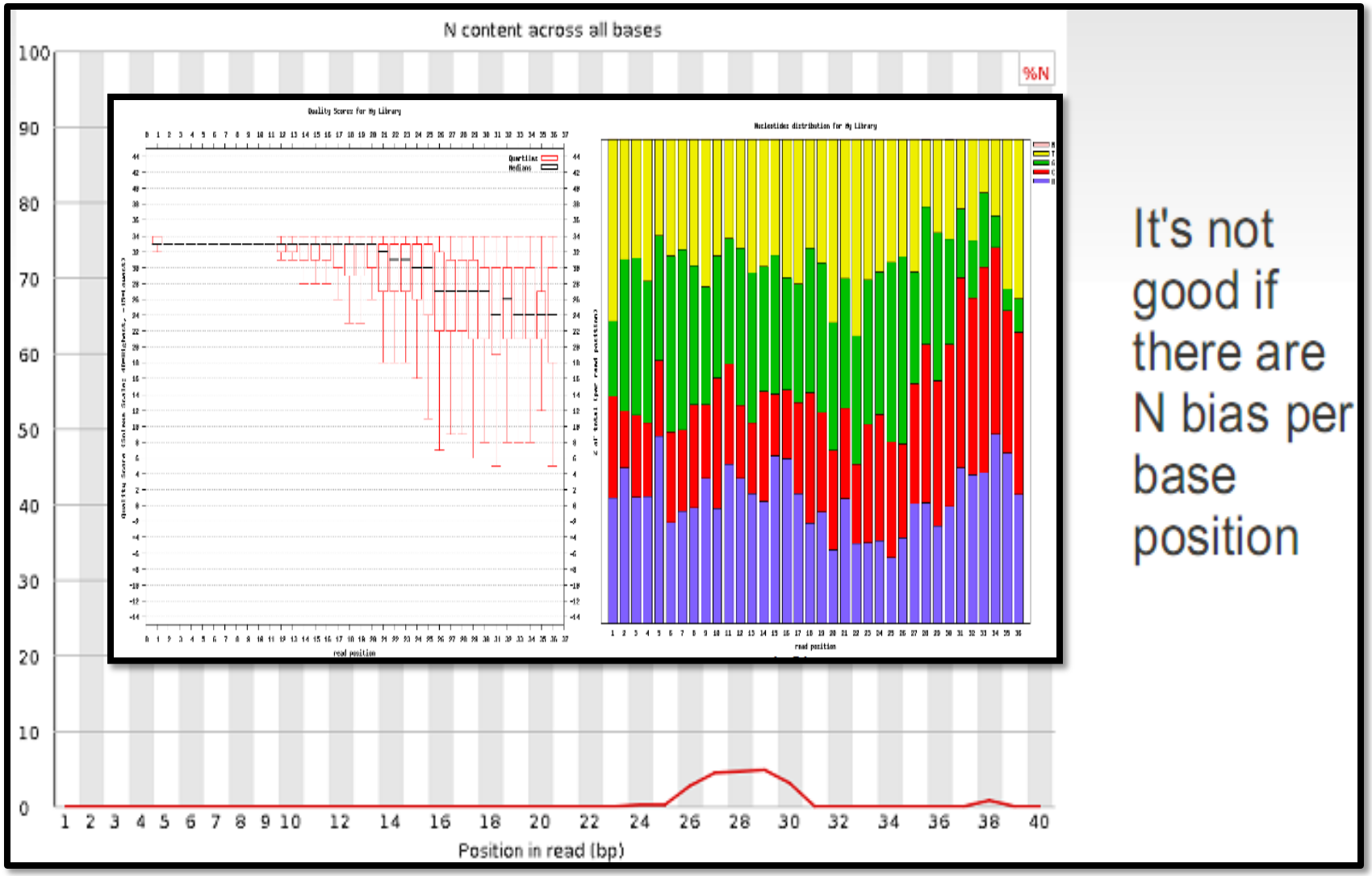
6. Per base GC distribution



Bad data

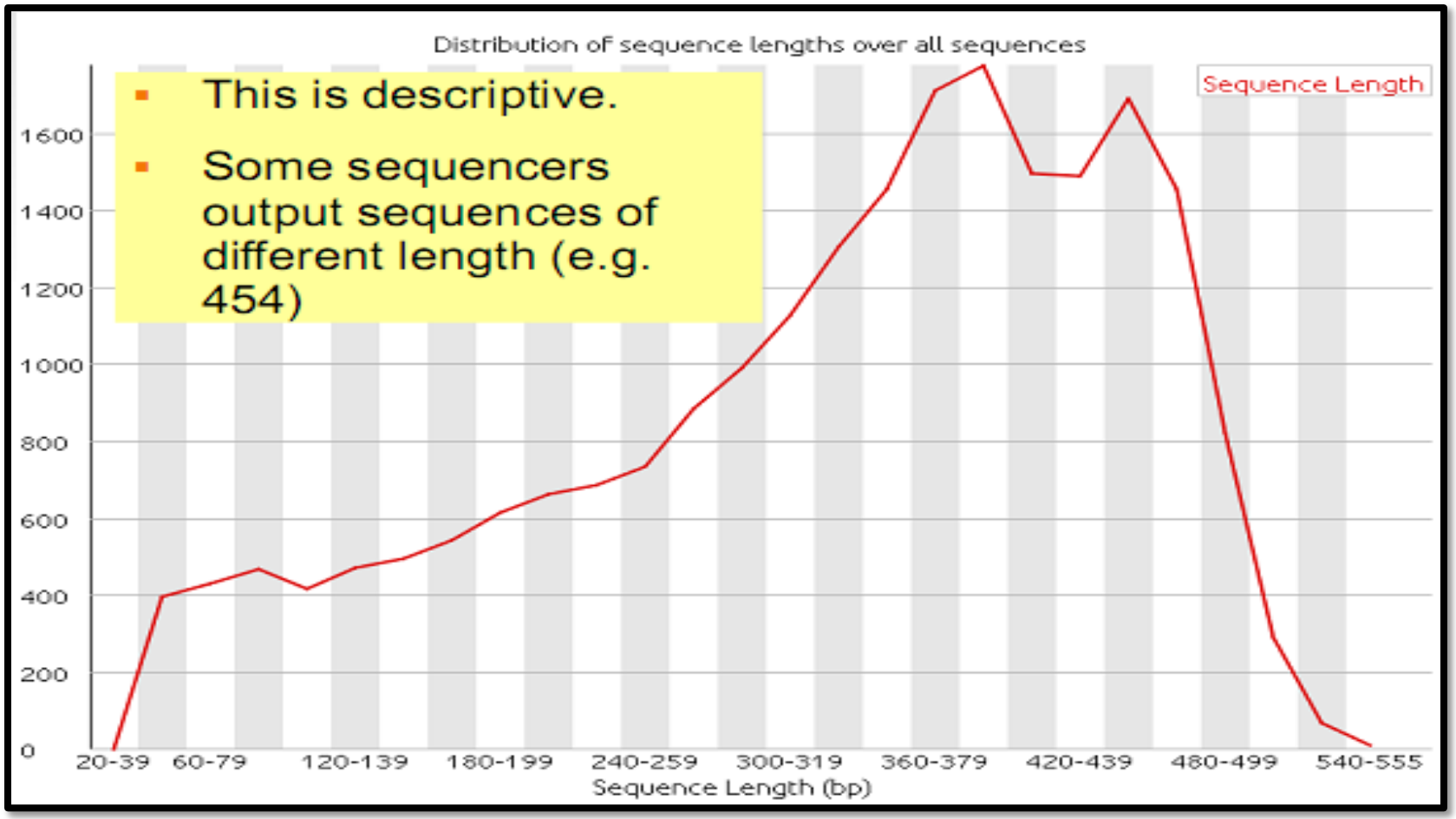
- Variation across read sequence

7. Per base N content



It's not good if there are N bias per base position

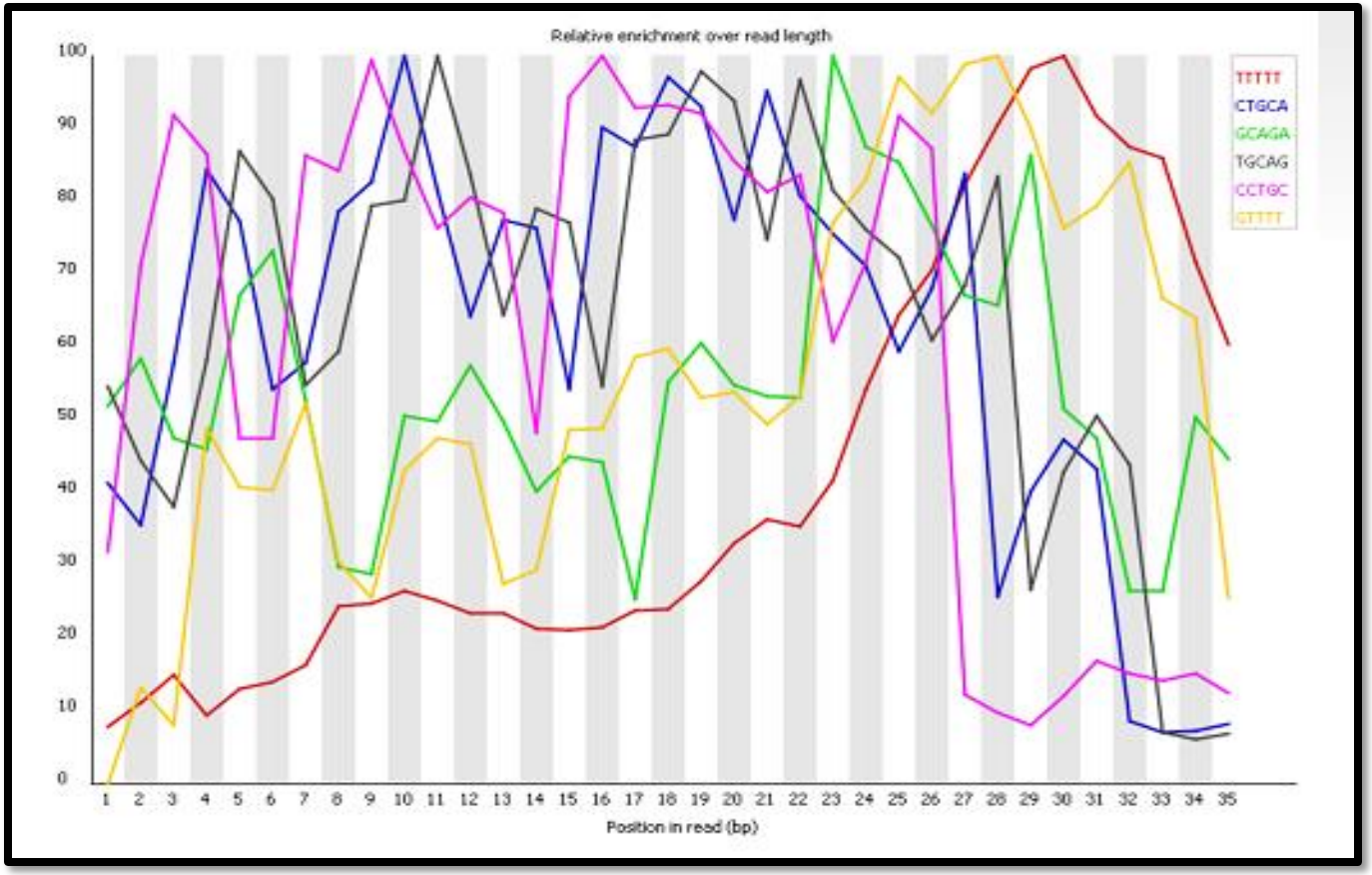
7. Length Distribution



- This is descriptive.
- Some sequencers output sequences of different length (e.g. 454)

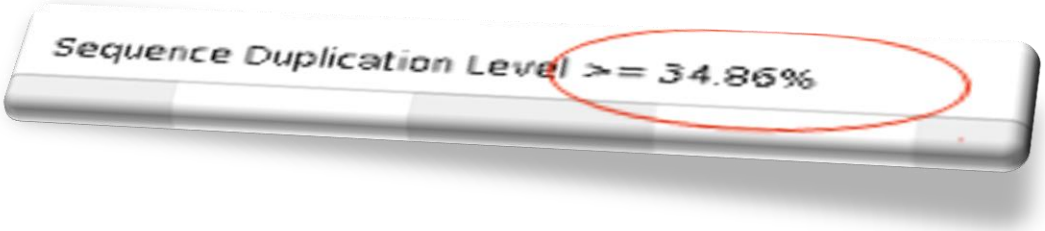
8. Kmer content

Any k-mer showing more than a 3 fold overall enrichment or a 5 fold enrichment at any given base position will be reported by this module.



9. Overrepresented/ duplicate sequences

Too many duplicate regions in the sequence will be due to sequencing problems



QC Report

- **Sequence Statistics**

Alignment statistics

```
Total Reads 15849154
Reads aligned   7746088
% Reads Aligned 48.8738
Total Genome Size   64022747
Genome Covered  28234853
%Coverage   44.1013
Avg Read Depth  1.50491
% Coverage at 1X    44.1013
% Coverage at 5X    10.7884
% Coverage at 10X   1.76412
% Coverage at 15X   0.297722
% Coverage at 20X   0.122413
% Coverage at 30X   0.0557255
% Coverage at 40X   0.0372789
```

- **Quality Statistics**

Preprocessing raw data

Removing technical artifacts

Adjusting biases

etc

Quality control of raw data

Proceed? Or rerun?

This QC can guide you to which preprocessing steps you need to apply **for sure**. The extra time and money needed to correct the biases can sometimes justify a rerun of the experiment.

This QC shows which preprocessing steps have already been made by the sequencing provider.

Preprocessing

Removing unwanted parts of the raw data so it helps as much as possible with reaching our goal: defining differentially expressed genes.

1) removing **technical** contamination
- Low quality read parts
- Technical sequences: adaptors
- PhiX internal control sequences

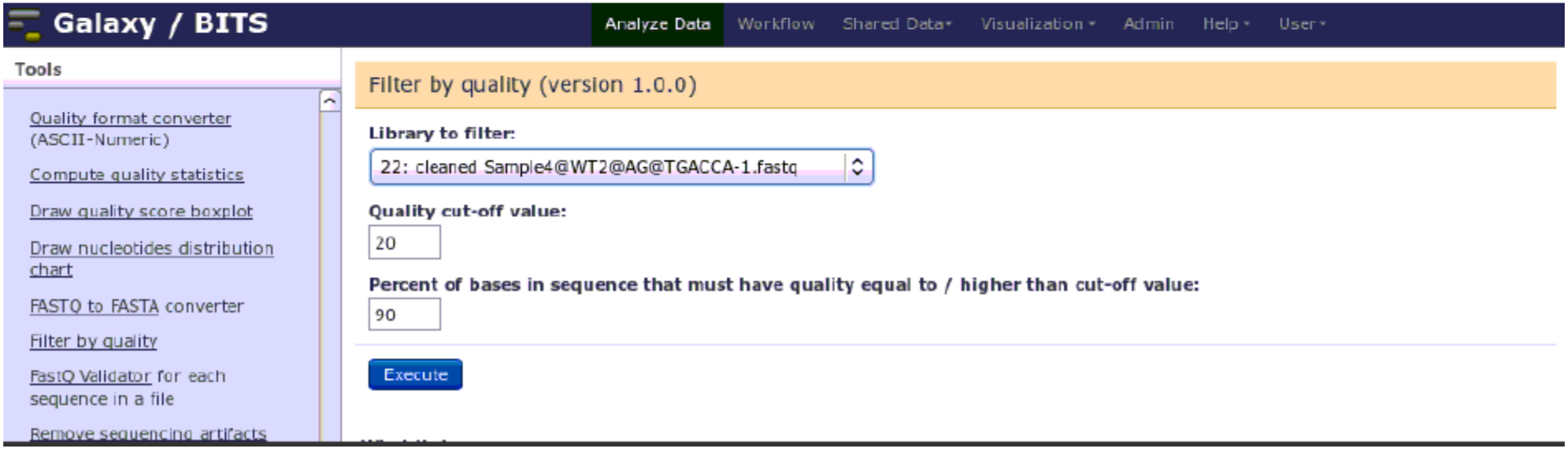
2) removing **biological** contamination
- polyA-tails
- rRNA sequences
- mtDNA sequences

After this, we run FastQC again.

Technical contamination

Our goal is to define DE expression, for this we need to assign reads with a high confidence to the correct genomic location.

Removal of **low quality** read parts: they have a higher chance to contain errors, and cause noise in our read counts.



Removing low quality reads

Galaxy / BITS

Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

Tools

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- FastQ Validator for each sequence in a file
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences
- Analyse SAM/BAM with bamqc to asses mapping quality metrics.
- FastqMcf sequence quality filtering and clipping
- Calculate the FASTQ sequence lengths calculates the length of each fastq entry
- Generate base quality report using quarc

Filter by quality (version 1.0.0)

Library to filter:

22: cleaned Sample4@WT2@AG@TGACCA-1.fastq

Quality cut-off value:

20

Percent of bases in sequence that must have quality equal to / higher than cut-off value:

90

Execute

What it does

This tool filters reads based on quality scores.

Using **percent = 100** requires all cycles of all reads to be at least the quality cut-off value.

Using **percent = 50** requires the median quality of the cycles (in each read) to be at least the quality cut-off value.

Quality score distribution (of all cycles) is calculated for each read. If it is lower than the quality cut-off value - the read is discarded.

Example:

```
@CSHL_4_FC042AG00II:1:2:214:584
GACAATAAAC
+CSHL_4_FC042AG00II:1:2:214:584
30 30 30 30 30 30 30 30 20 10
```

Using **percent = 50** and **cut-off = 30** - This read will not be discarded (the median quality is higher than 30).

Using **percent = 90** and **cut-off = 30** - This read will be discarded (90% of the cycles do no have quality equal to / higher than 30).

Using **percent = 100** and **cut-off = 20** - This read will be discarded (not all cycles have quality equal to / higher than 20).

This tool is based on FASTX-toolkit by Assaf Gordon.

Trimming reads

Tools

- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- FASTQ interlacer on paired end reads
- FASTQ de-interlacer on paired end reads
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- FastQ Validator for each sequence in a file
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement

FASTQ Quality Trimmer (version 1.0.0)

FASTQ File:

22: cleaned Sample4@WT2@AG@TGACCA-1.fastq

Keep reads with zero length:

☐

Trim ends:

5' and 3'

Window size:

1

Step Size:

1

Maximum number of bases to exclude from the window during aggregation:

0

Aggregate action for window:

min score

Trim until aggregate score is:

>=

Quality Score:

0.0

Execute

This tool allows you to trim the ends of reads based upon the aggregate value of quality scores found within a sliding window; a sliding window of size 1 is equivalent to 'simple' trimming of the ends.

The user specifies the aggregating action (min, max, sum, mean) to perform on the quality score values found within the sliding window to be used with the user defined comparison operation and comparison value.

The user can provide a maximum count of bases that can be excluded from the aggregation within the window. When set, this tool will first check the aggregation of the entire window, then after removing 1 value, then after removing 2 values, up to the number declared. Setting this value to be equal to or greater than the window size will cause no trimming to occur.

⚠ Trimming a color space read will cause any adapter base to be lost.

Citation

If you use this tool, please cite Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A; Galaxy Team. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5.

Technical contamination

Our goal is to define DE expression, for this we need to assign reads with a high confidence to the correct genomic location.

Removal of adaptor sequences (and other technical sequences, such as multiplex) as they cannot be mapped to the reference genome.

FastqMcf (version 1.0)

A fasta formatted adapter list:

4: non-coding sequences Candida SC5314

Reads: single or Left-hand of Paired End Reads:

22: cleaned Sample4@WT2@AG@TGACCA-1.fastq

Right-hand mates for Paired End Reads:

Selection is Optional

Trimming Options:

Use Defaults

-P phred-scale (64):

Remove adaptors & primers & etc

FastqMcf (version 1.0)

A fasta formatted adapter list:

4: non-coding sequences Candida SC5314

List of technical sequences

Reads: single or Left-hand of Paired End Reads:

22: cleaned Sample4@WT2@AG@TGACCA-1.fastq

Right-hand mates for Paired End Reads:

Selection is Optional

Trimming Options:

Use Defaults

Advised to use defaults

-P phred-scale (64):

Default is to determine automatically

-n Don't clip, just output what would be done:

Execute

What it does

fastq-mcf attempts to:

Detect and remove sequencing adapters and primers Detect limited skewing at the ends of reads and clip Detect poor quality at the ends of reads and clip Detect N's, and remove from ends Remove reads with CASAVA 'Y' flag (purity filtering) Discard sequences that are too short after all of the above Keep multiple mate-reads in sync while doing all of the above

Fastq-mcf-output

Data Viewer

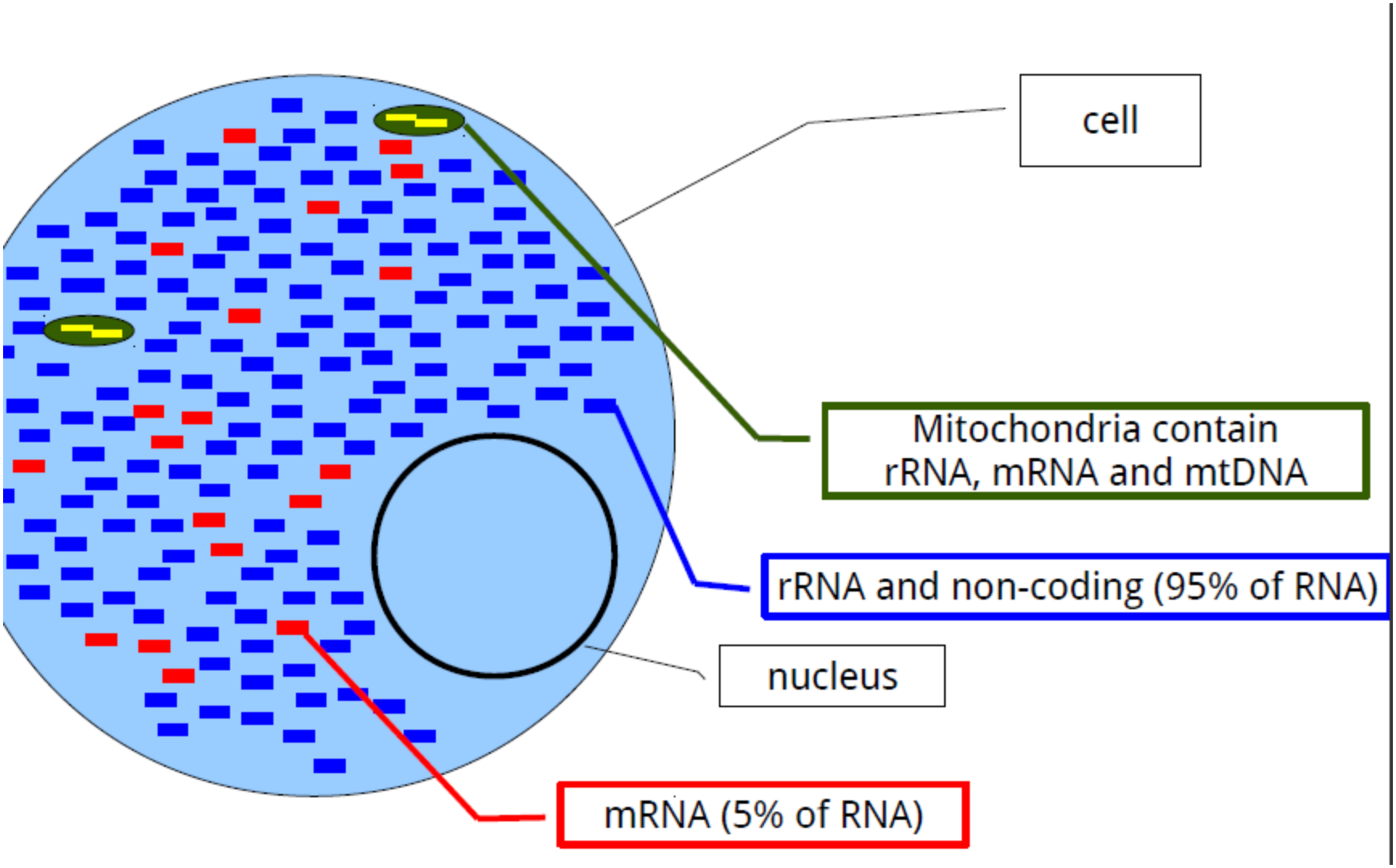
```
Scale used: 2.2
Phred: 33
Threshold used: 751 out of 300000
Adapter tag9 (GATCAG): counted 1527 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 6
Adapter tag8 (ACTTGA): counted 5488 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 5
Adapter tag7 (CAGATC): counted 2133 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 6
Adapter tag6 (GCCAAT): counted 7546 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 4
Adapter tag5 (ACAGTG): counted 3578 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 5
Adapter tag4 (TGACCA): counted 8205 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 4
Adapter tag1 (ATCACG): counted 1268 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 6
Adapter tag15 (ATGTCA): counted 3427 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 5
Adapter tag14 (AGTTCC): counted 1421 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 6
Adapter tag13 (AGTCAA): counted 7804 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 4
Adapter tag12 (CTTGTA): counted 7139 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 4
Adapter tag10 (TAGCTT): counted 9255 at the 'end' of '/mnt/galaxydb/files/016/dataset_16656.dat', clip set to 4
Files: 1
Total reads: 7239748
Too short after clip: 564286
Clipped 'end' reads: Count: 933479, Mean: 15.81, Sd: 8.72
Trimmed 811764 reads by an average of 5.26 bases on quality < 20
```

Technical contamination

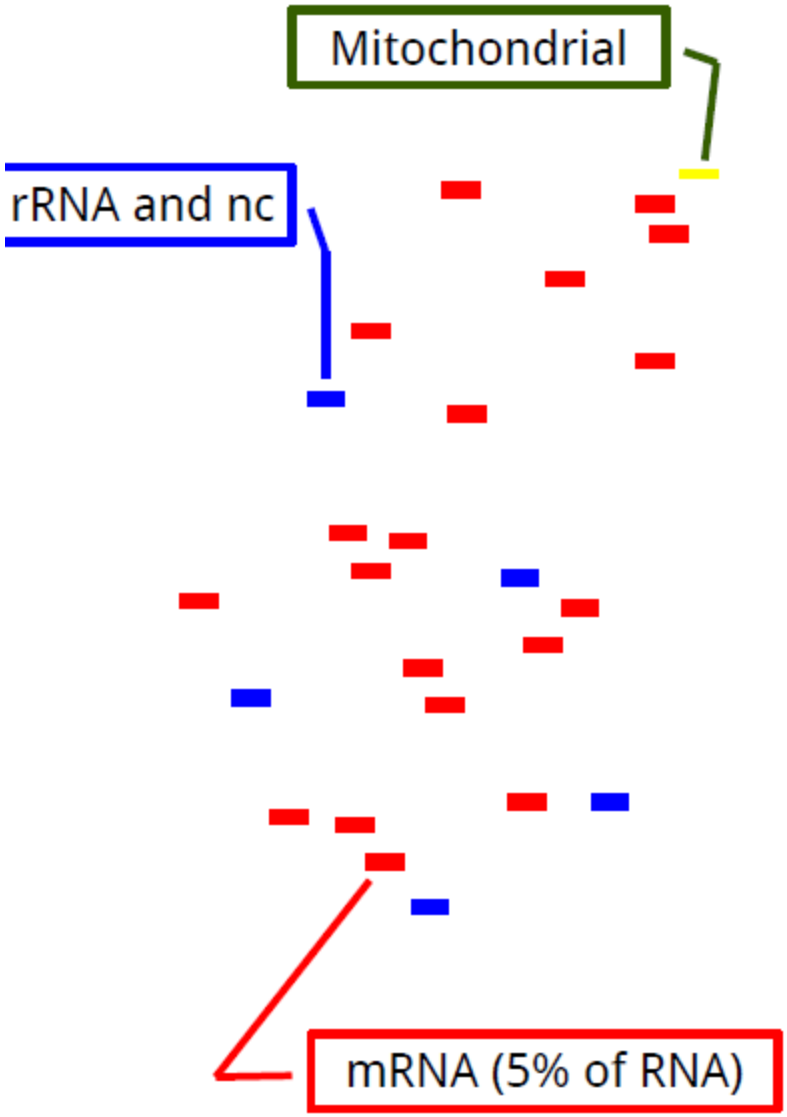
- **Never remove duplicate reads**! Highly expressed genes can have genuine duplicate reads, which are not due to the PCR amplification step in the protocol.

- **PhiX sequences**: the DNA of Phi X bacteriophage is spiked in to monitor and optimize sequencing on Illumina machines. Your sequencing provider should filter out those sequences before delivery. You can filter them out by aligning your reads to the PhiX genome.

Biological contamination



Biological contamination



mRNAs are captured with oligo-dT coated beads.

Occasionally, **non-protein coding sequences** are also captured (especially since mtRNA and rRNA can be relatively rich in AT).

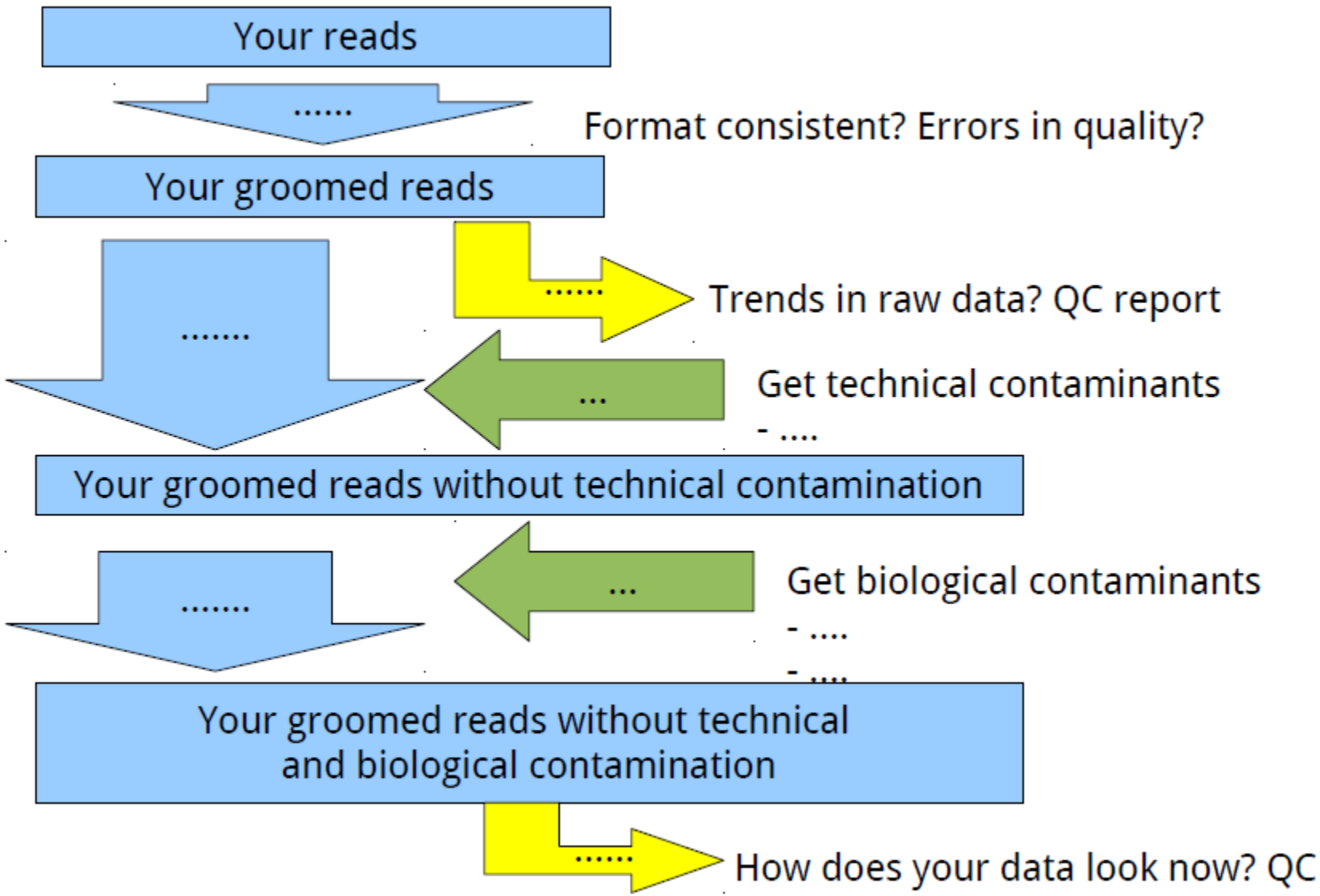
We can remove them via **homology searching** (BLAST) with known non-protein coding sequences.

Biological contamination

AAAAAAAAAAAAA

mRNAs are post-transcriptionally modified: e.g. the addition of a **poly-A tail**. If our goal is to map the reads to a reference genome sequence, the polyA tails should be removed. This can be viewed as some source of 'biological contamination' in our sequences (...).

Summary preprocessing



Acknowledgements

- Part of these slides has been prepared borrowing material from other presentations.
- I am particularly thankful to
  - Javier Santoyo
  - Joachim Jacob (VIB)
  - Karan Veer Singh (NBAGR)