



From QC to normalization of RNA-seq data

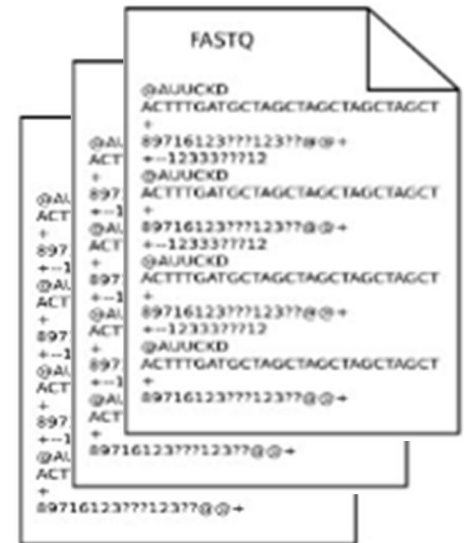
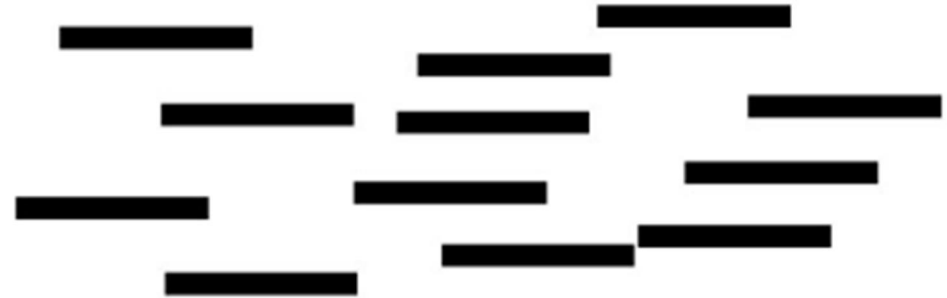
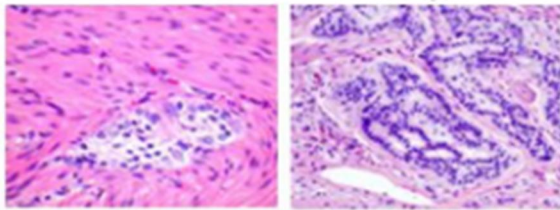
Ahmed DEBIT
5/22/2017

BIO-GIGA Training Session
GIGA-R (Liège – Belgium)



Overview of Sequencing

Input: Sample of interest



Output: .fastq files
.fq.gz



FASTQ file

- Standardized output format
- Contains millions of records
- Each record is represented by four lines

Illumina ASCII_BASE = 33

Base = T

Score = D (ASCII code = 68) = 35

```
Name      @ERR127302.1 HWI-EAS350_0441:1:1:1055:4898#0/1
Sequence  GGCTCATCTTGATACTGGGTGGCGACCGTCCCTGGCCCCTTCTTGACACCCA
+
Quality score 4=B@D99BDDDDDDDD:DD?B<<=?>6B#####
```

An orange arrow points from the text "Base = T" to the 'T' in the sequence line. A red box highlights the 'T' in the sequence line and the 'D' in the quality score line.



Quality Assessment & Data Filtering


QUALITY ASSESSMENT



Aim of QC ...

- Assess sequence qualities
- Collect statistics about NGS runs and sequence compositions

```
@ERR127302.1 HWI-EAS350_0441:1:1:1055:4898#0/1  
GGCTCATCTTGAAGTGGGTGGCGACCGTCCCTGGCCCCTTCTTGACACCCA  
+
```

 4=B@D99BDDDDDDDD:DD?B<<=?>6B#####



FASTQC

QC = Are the data correct enough for next step?

- Main FASTQC results:
 1. Basic statistics
 2. Per base sequence quality
 3. Per sequence quality scores
 4. Per base sequence content
 5. Per sequence GC content
 6. Sequence duplication levels
 7. Overrepresented sequences





1. Basic statistics

Measure	Value
Filename	NGS13-A521_2T_GCCAAT_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	39915280
Sequences flagged as poor quality	0
Sequence length	101
%GC	57

Raw data

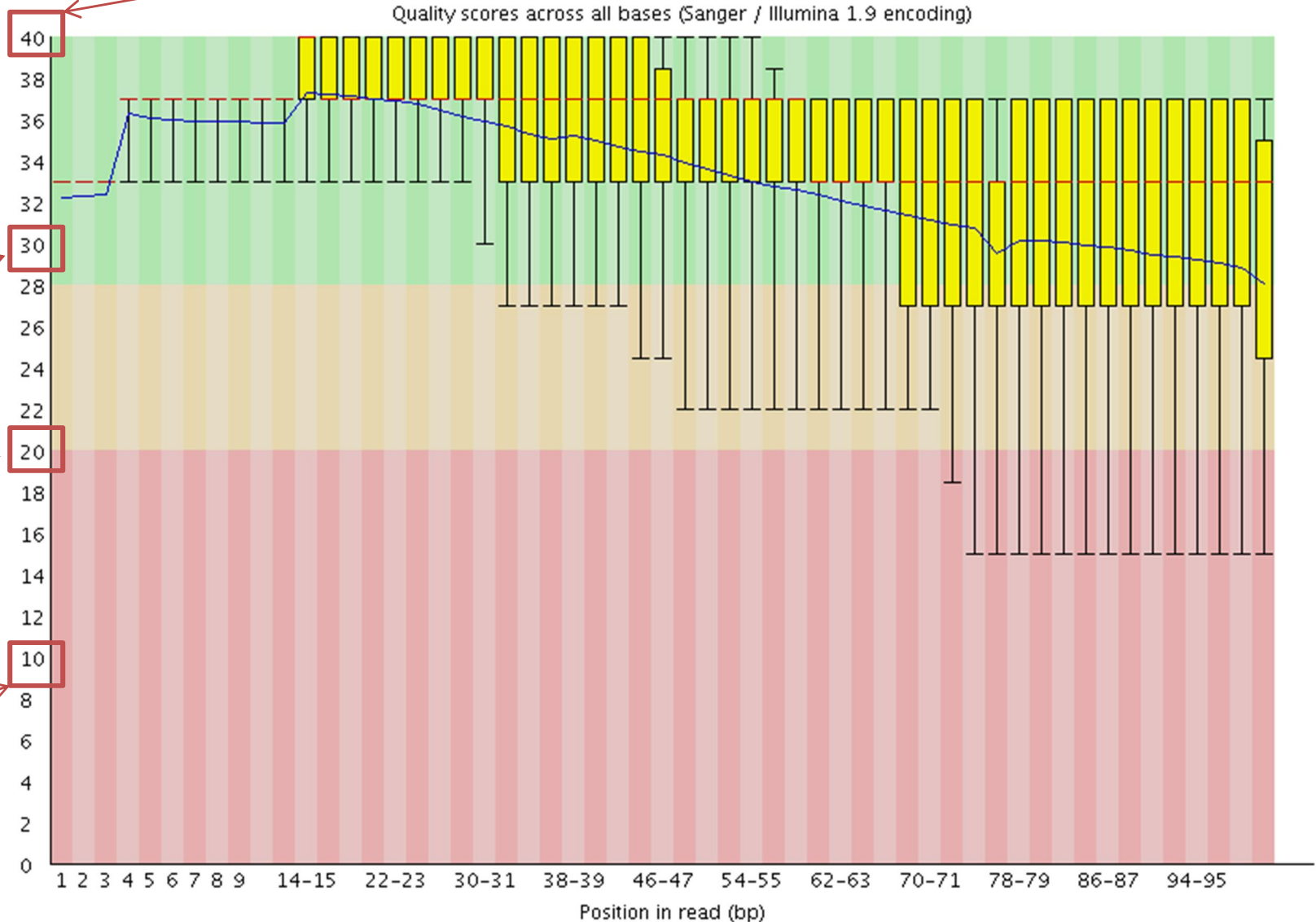


2. Per base sequence quality



$$Q = -10 \log_{10} p$$

1 error in every
10K bases = 99,99%
accuracy



1 error in every
1K bases = 99,90%
accuracy

1 error in every
100 bases = 99%
accuracy

1 error in every
10 bases = 90%
accuracy

Raw data

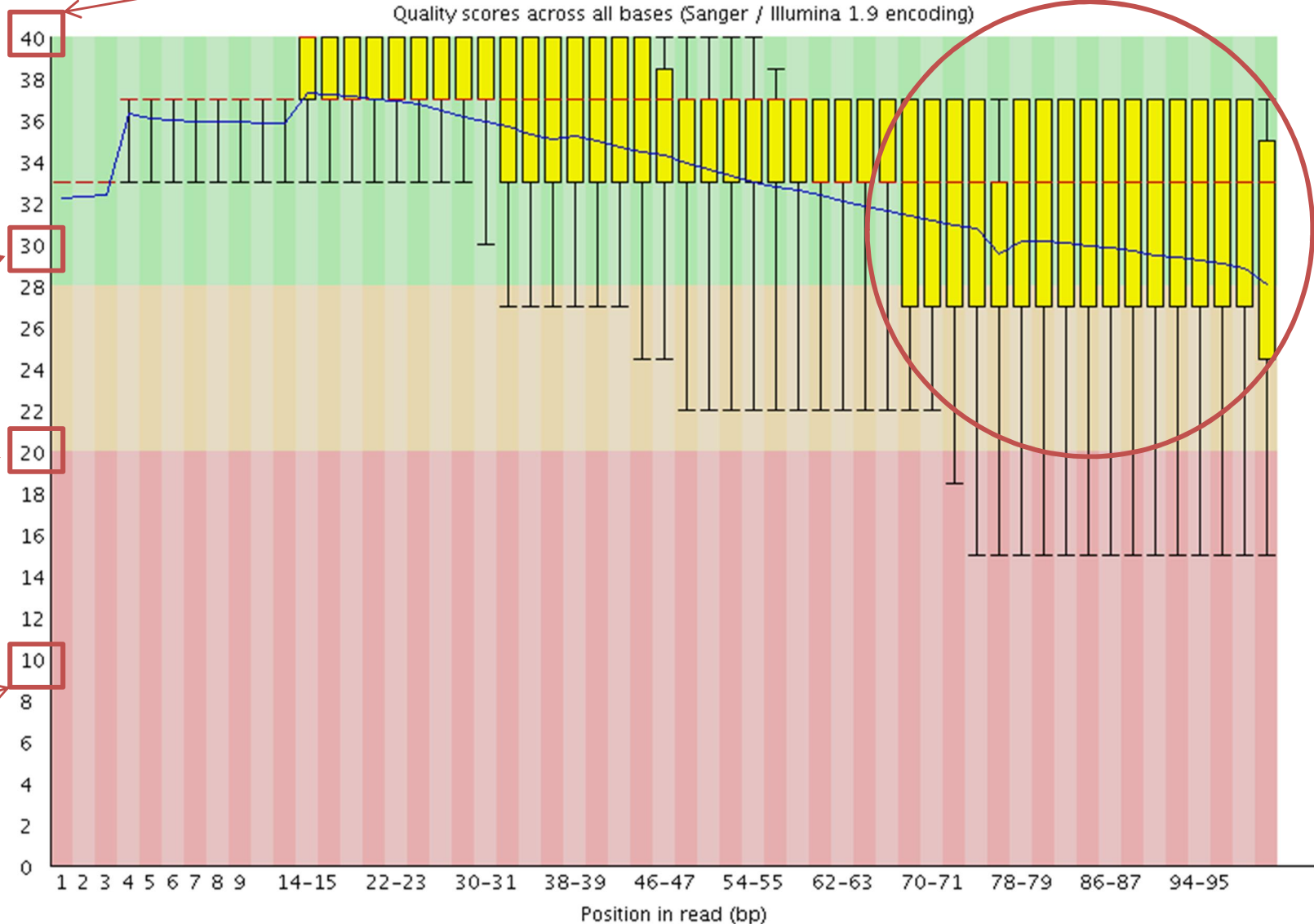


2. Per base sequence quality



$$Q = -10 \log_{10} p$$

1 error in every
10K bases = 99,99%
accuracy



1 error in every
1K bases = 99,90%
accuracy

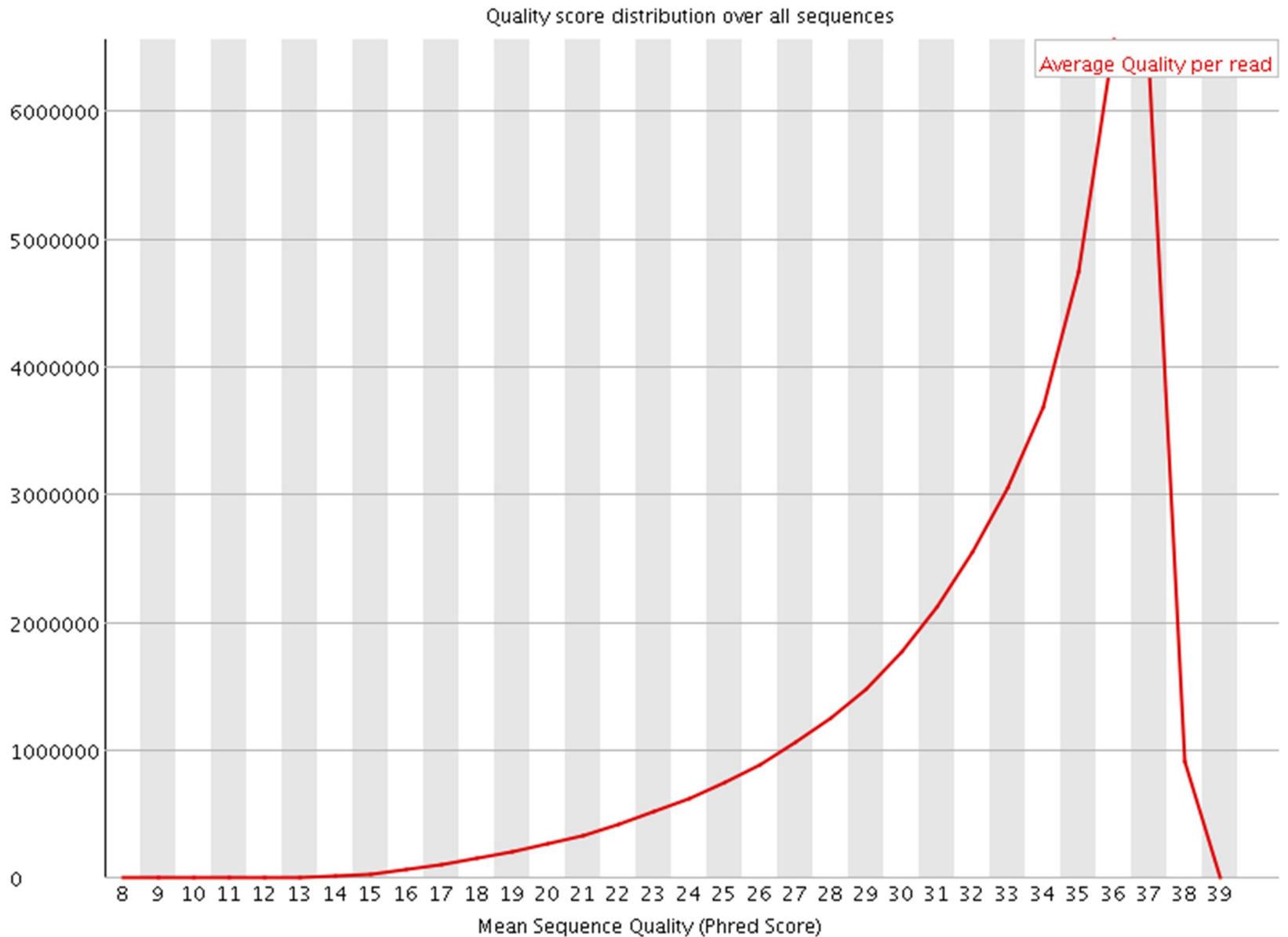
1 error in every
100 bases = 99%
accuracy

1 error in every
10 bases = 90%
accuracy

Raw data



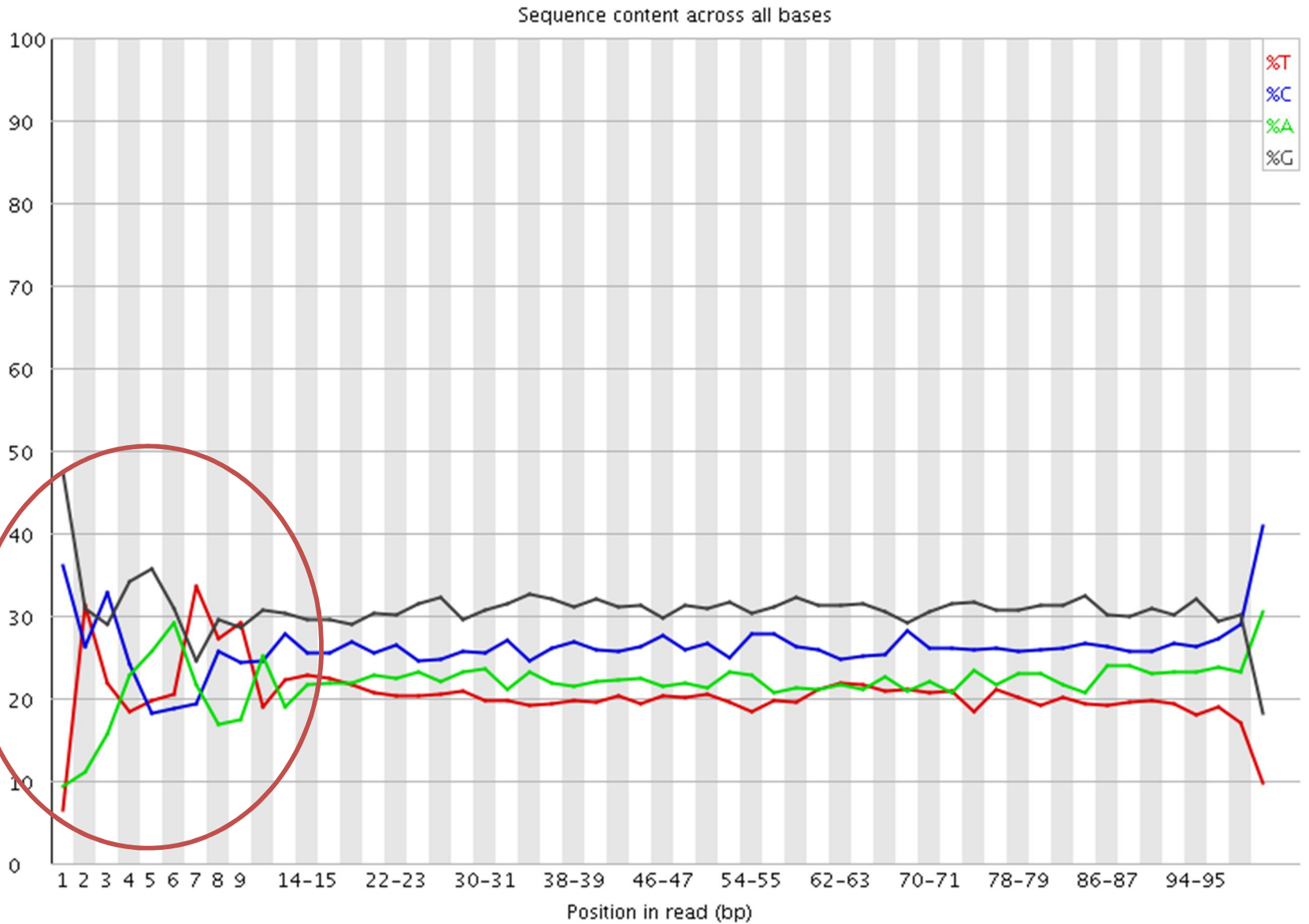
3. Per sequence quality scores



Raw data



4. Per base sequence content

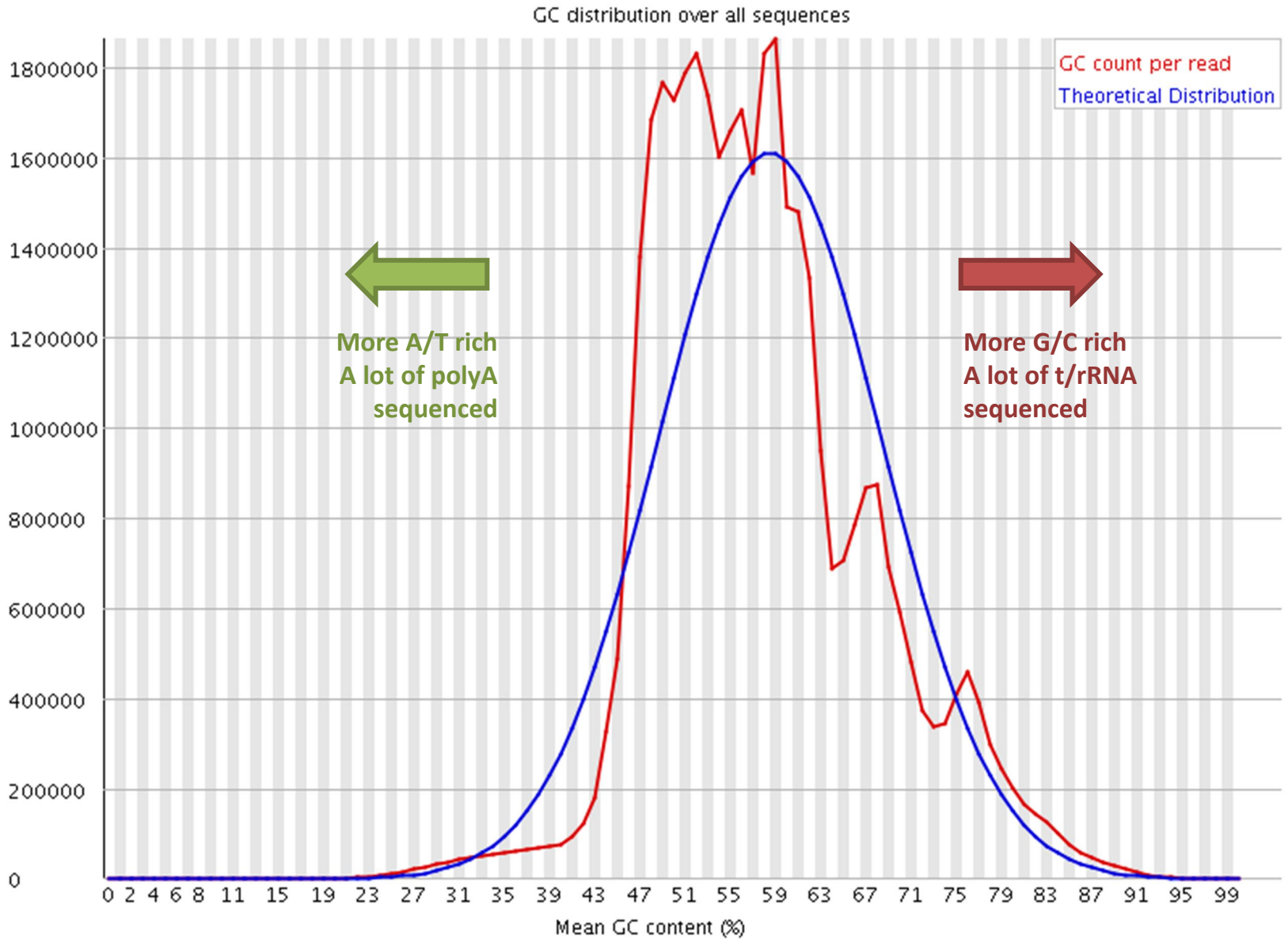


characteristic of
Illumina RNA
-seq data

Raw data



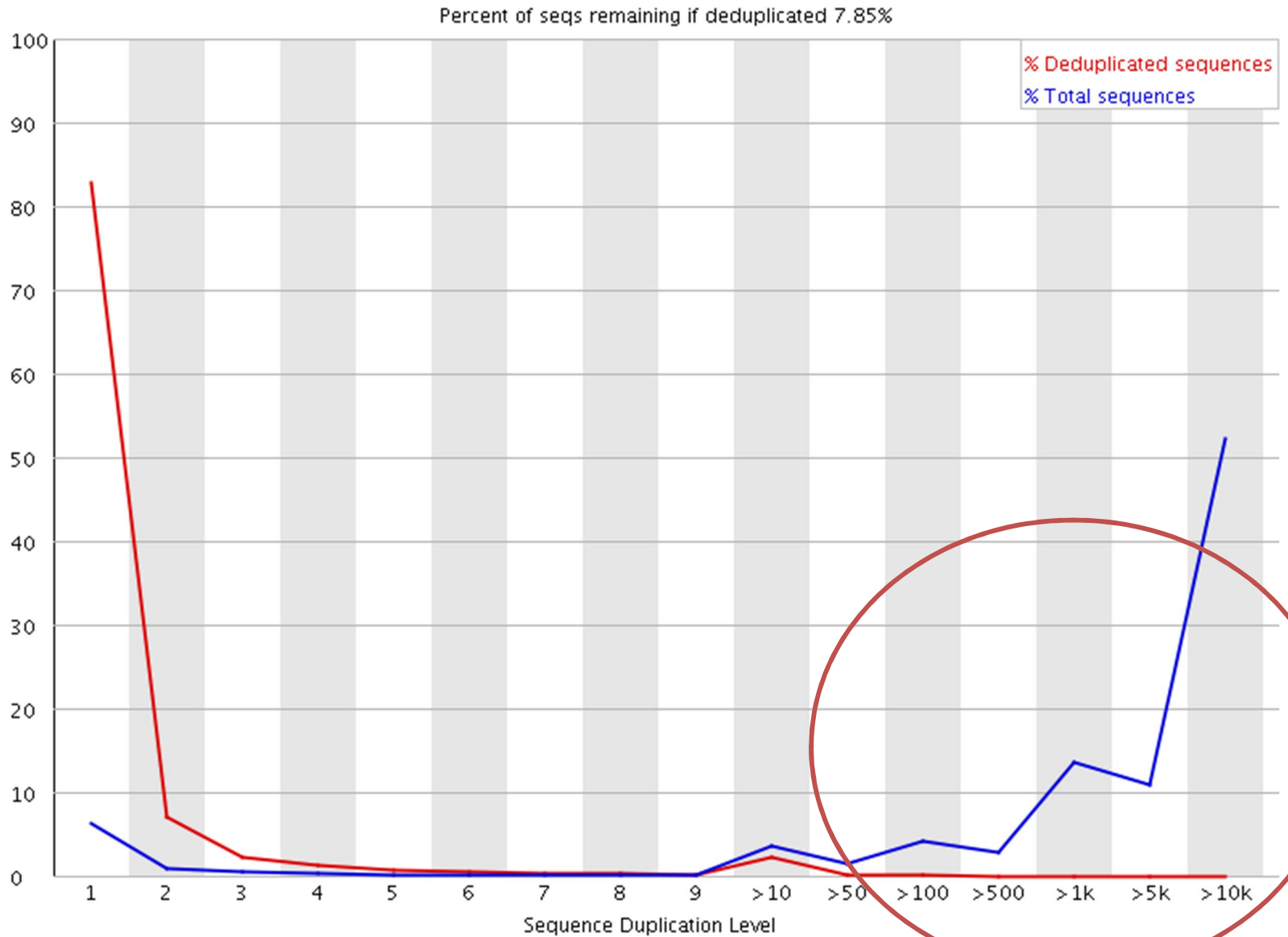
5. Per sequence GC content (Raw data)



Raw data



6. Sequence duplication levels



Raw data



7. Overrepresented sequences



Sequence	Count	Percentage	Possible Source
CTTCGATGTCGGCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGCGTT	229462	0.5748725801247041	No Hit
GTTGGATTGTTACCCACTAATAGGGAACGTGAGCTGGGTTTAGACCGTC	161617	0.40490007836597913	No Hit
CCAGTAAGTGCGGGTCATAAGCTTGCGTTGATTAAGTCCCTGCCCTTTGT	160593	0.40233464477763903	No Hit
GGGAGTTTGACTGGGGCGGTACACCTGTCAAACGGTAACGCAGGTGTCCT	158328	0.39666012614717977	No Hit
CAGAAAAGTTACCACAGGGATAACTGGCTTGTGGCGGCCAAGCGTTCATA	131459	0.3293450528218767	No Hit
GGCGGAGATGGGCGCCGCGAGGCGTCCAGTGCGGTAACGCGACCGATCCC	129503	0.32444467381914893	No Hit
CCTAAGGCGAGCTCAGGGAGGACAGAAACCTCCCGTGAGCAGAAGGGCA	119008	0.2981514848449015	No Hit
GGATTGTTACCCACTAATAGGGAACGTGAGCTGGGTTTAGACCGTCGTG	118427	0.2966959019202671	No Hit
GGCGTACGGAAGACCCGCTCCCCGGCGCCGCTCGTGGGGGGCCCAAGTCC	118090	0.2958516137178544	No Hit
GCCGAAGTGGAGAAGGCTTCCATGTGAACAGCAGTTGAACATGGGTCAAGT	115883	0.29032240284923466	No Hit
TGATGATGTGTTGTTGCCATGGTAATCCTGCTCAGTACGAGAGGAACCGC	109247	0.27369719064979625	No Hit
GTTTTAAGCAGGAGGTGTCAGAAAAGTTACCACAGGGATAACTGGCTTGT	108340	0.27142487789137393	No Hit
GTCGGCTCTTCCTATCATTGTGAAGCAGAATTCACCAAGCGTTGGATTGT	97615	0.24455546848224538	No Hit
GCTGGGTTTAGACCGTCGTGAGACAGGTTAGTTTTACCCTACTGATGATG	96711	0.2422906716425389	No Hit
GTCCGGGGCTGCACGCGCGCTACACTGACTGGCTCAGCGTGTGCCTACCC	95826	0.24007347562136605	No Hit
GTTCAAAGCAGGCCCGAGCCGCCTGGATACCGCAGCTAGGAATAATGGAA	95332	0.23883585433949103	No Hit
GTCTGTGATGCCCTTAGATGTCCGGGGCTGCACGCGCGCTACACTGACTG	94870	0.23767840285725167	No Hit

Raw data

rRNA or other contaminants

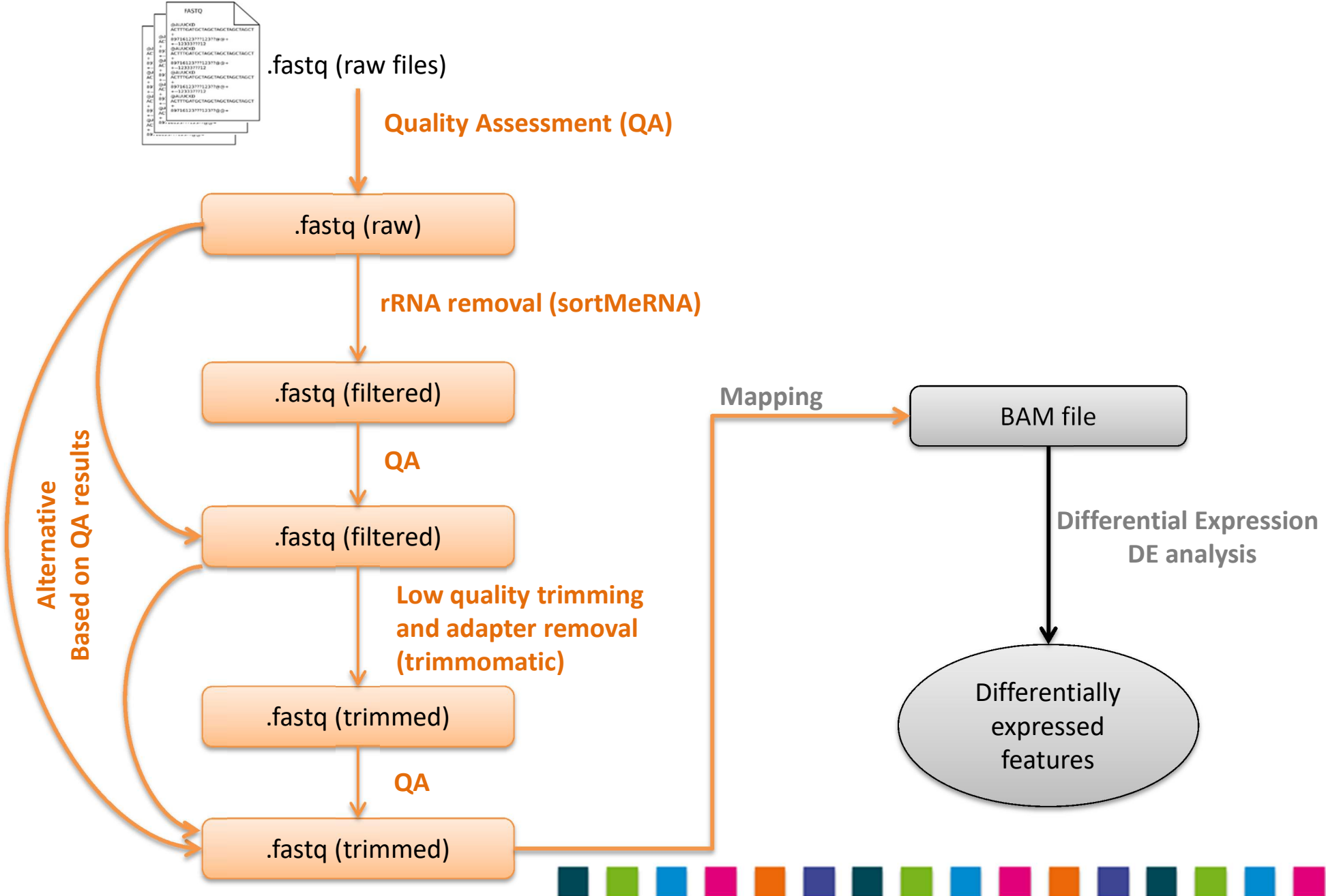


After first pass QC ...

We should remove rRNA



RNA-seq preprocessing and Analysis workflow



Quality Assessment & Data Filtering

DATA FILTERING



... Filtering

- Remove:
 - Sequencing adaptors (trimmomatic, Cutadapt)
 - Low quality reads (fastx toolkit, PRINSEQ)
 - rRNA and others RNA contaminants (SortMeRNA)



rRNA out-filtering

- Acceptable rate between 0.1% and 3% (rRNA depletion)
- $> 3\%$ may effect the usable number of reads.
- Tool used hereafter: SortMeRNA



SortMeRNA pipeline



Create rRNA databases (ref. DB)

Indexing of all the databases

R1.fastq

R2.fastq

Merging of the paired-end files

sortMeRNA

--other

--aligned

rRNA-free file

rRNA aligned file

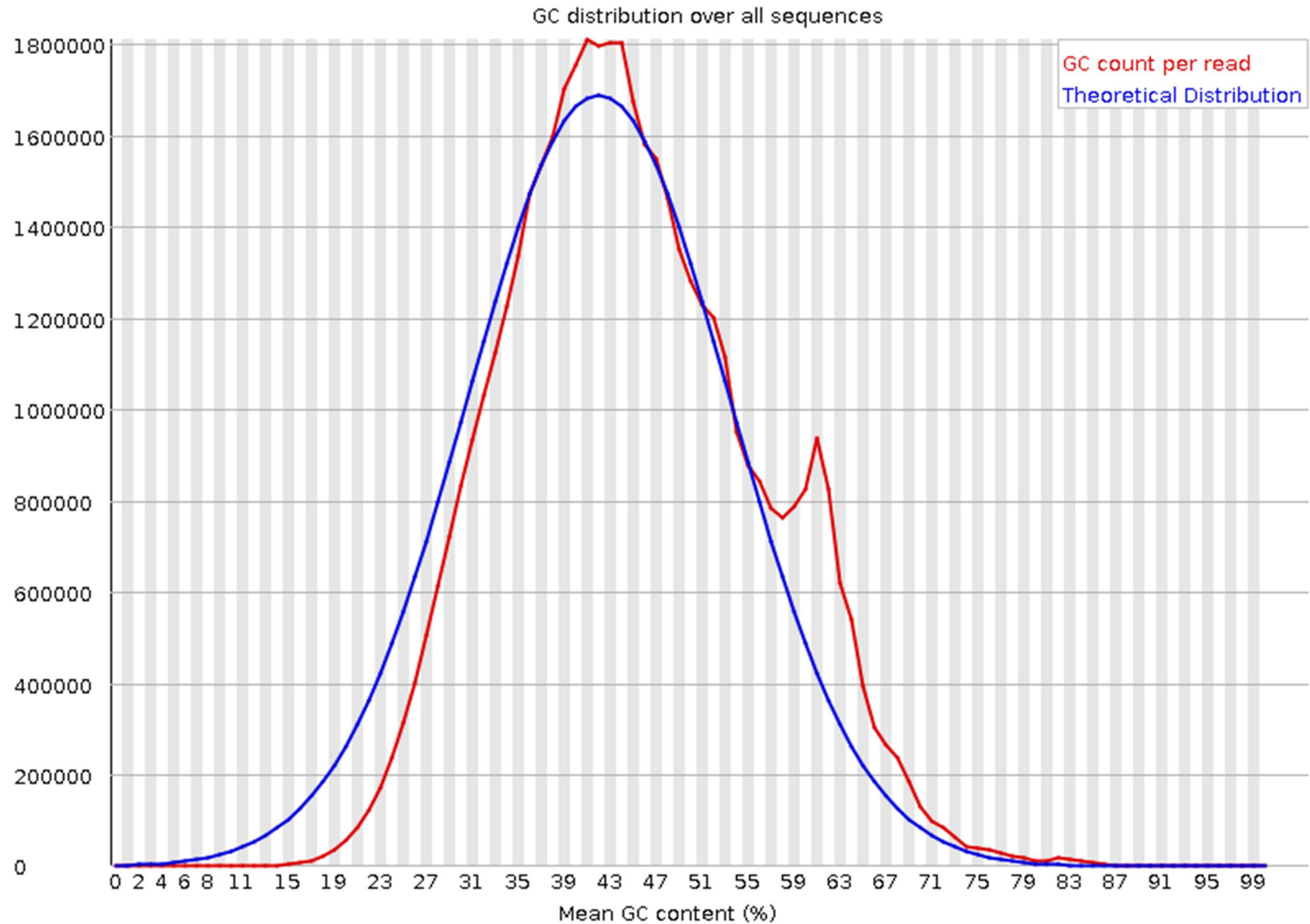
Unmerging

R1.fastq

R2.fastq



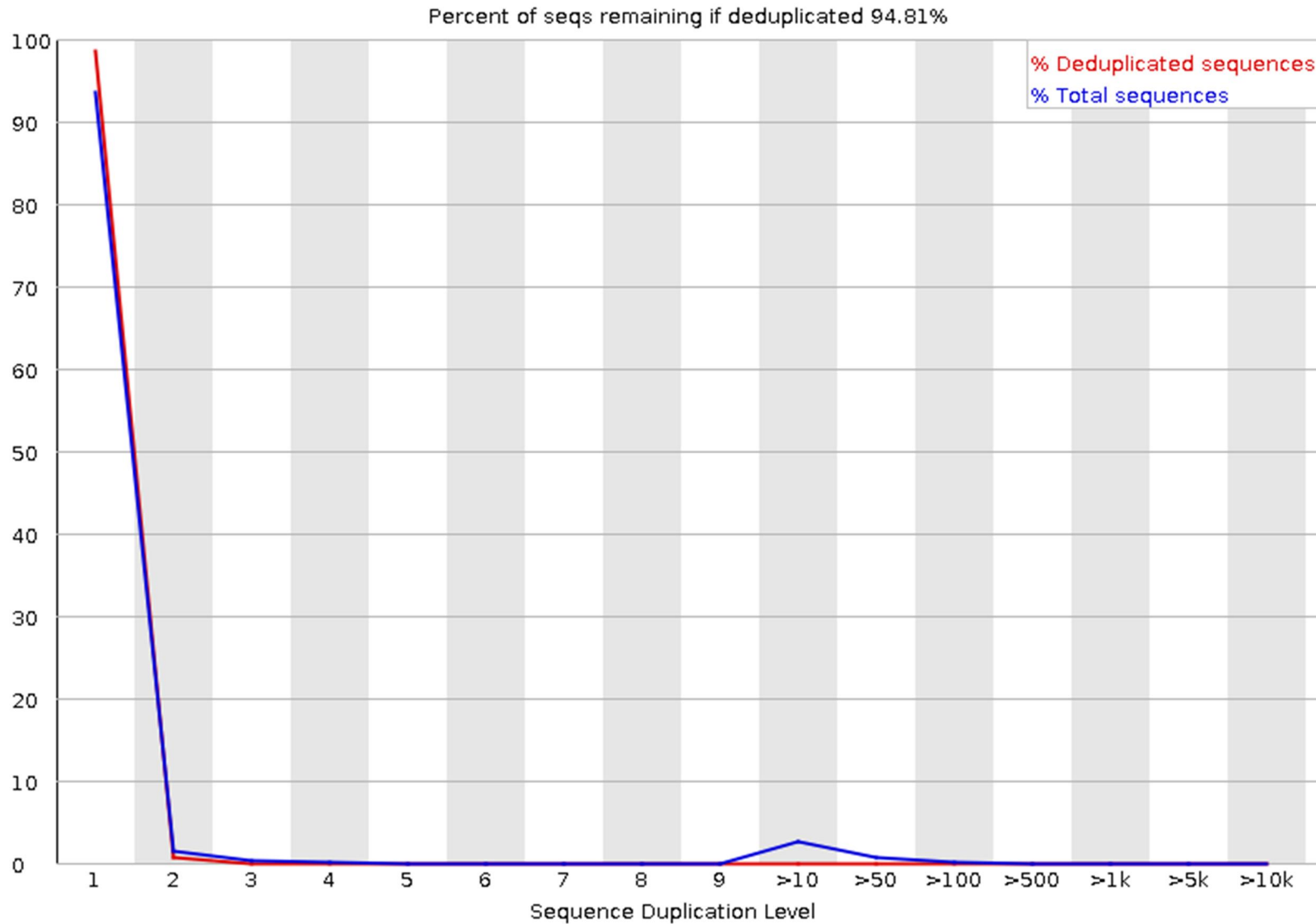
After rRNA filtering



rRNA-free data



After rRNA filtering



rRNA-free data



Quality trimming and adapter removal

- Sequencing process: bases in the later cycles receive a lower average quality than the earliest cycles.



Quality trimming and adapter removal

- Sequencing process: bases in the later cycles receive a lower average quality than the earliest cycles.
- Trim low quality bases from the 3' until the quality reaches a selected Phred score threshold.



Quality trimming and adapter removal

- Sequencing process: bases in the later cycles receive a lower average quality than the earliest cycles.
- Trim low quality bases from the 3' until the quality reaches a selected Phred score threshold.
- Presence of partial adapter sequences within sequenced reads: Adapter removal



Quality trimming and adapter removal

- Sequencing process: bases in the later cycles receive a lower average quality than the earliest cycles.
- Trim low quality bases from the 3' until the quality reaches a selected Phred score threshold.
- Presence of partial adapter sequences within sequenced reads: Adapter removal
- Tool: Trimmomatic / cutadapt



Example of adapter contamination

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	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	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)
CGGTTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

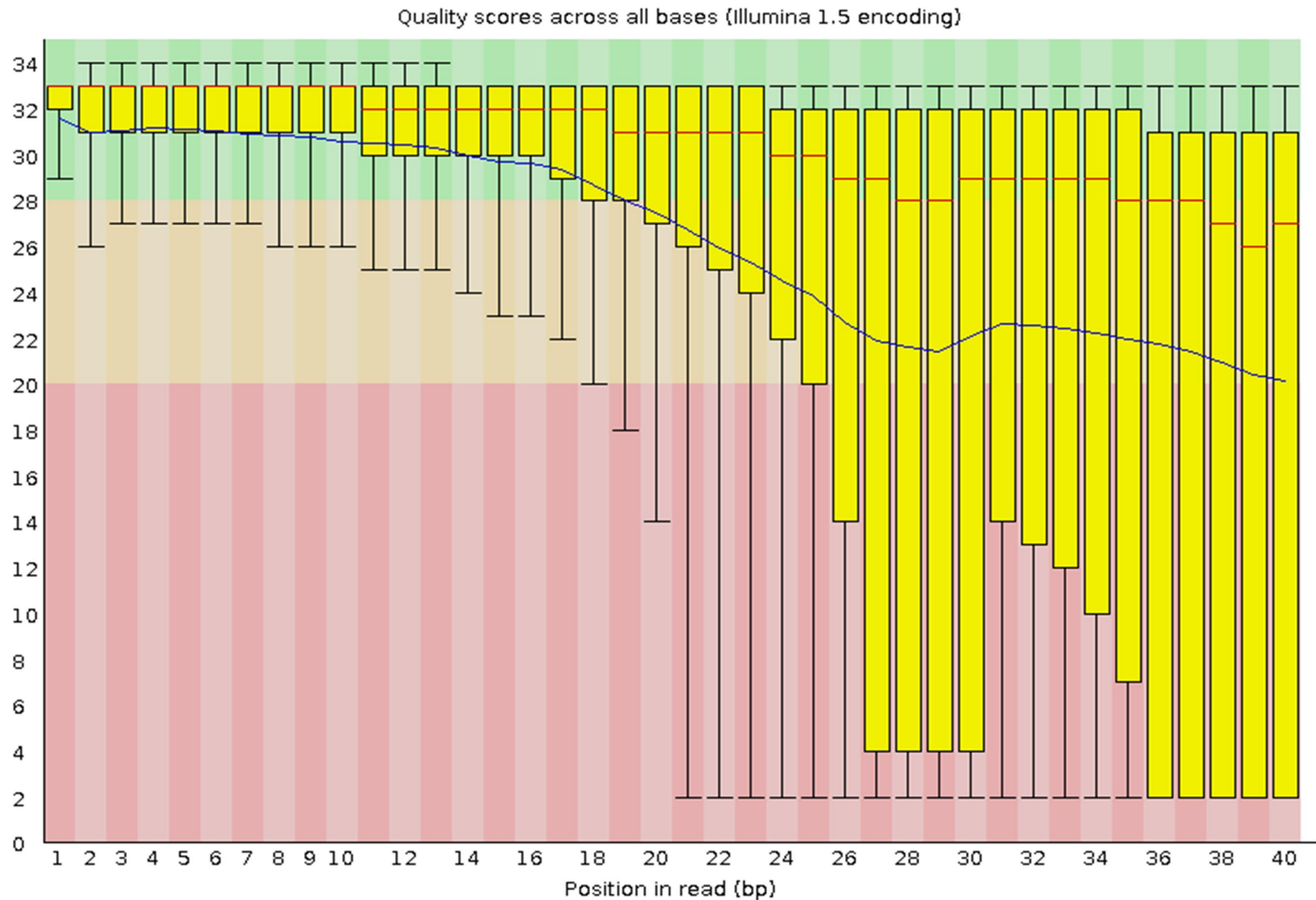


Example of adapter contamination

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGATCGGAAG	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)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGGAAG	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)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)



Example of low/bad quality



General remarks on QC using FASTQC

- A “**bad**” result from FASTQC doesn’t always mean the data are not useful or valuable



Alternate tools for QC and filtering

- FASTQC [*Andrews S et al. 2010*]: **QC Standard tool**
 - AfterQC [*Shifu Chen et al. 2017*]
 - RSeQC [*Liguo Wang et al. 2012*]
 - RNA-SeQC [*David DeLuca et al. 2012*]
 - Picard [<http://picard.sourceforge.net>]
- } Specific to RNA-seq data



Normalization for differential expression analysis

NORMALIZATION

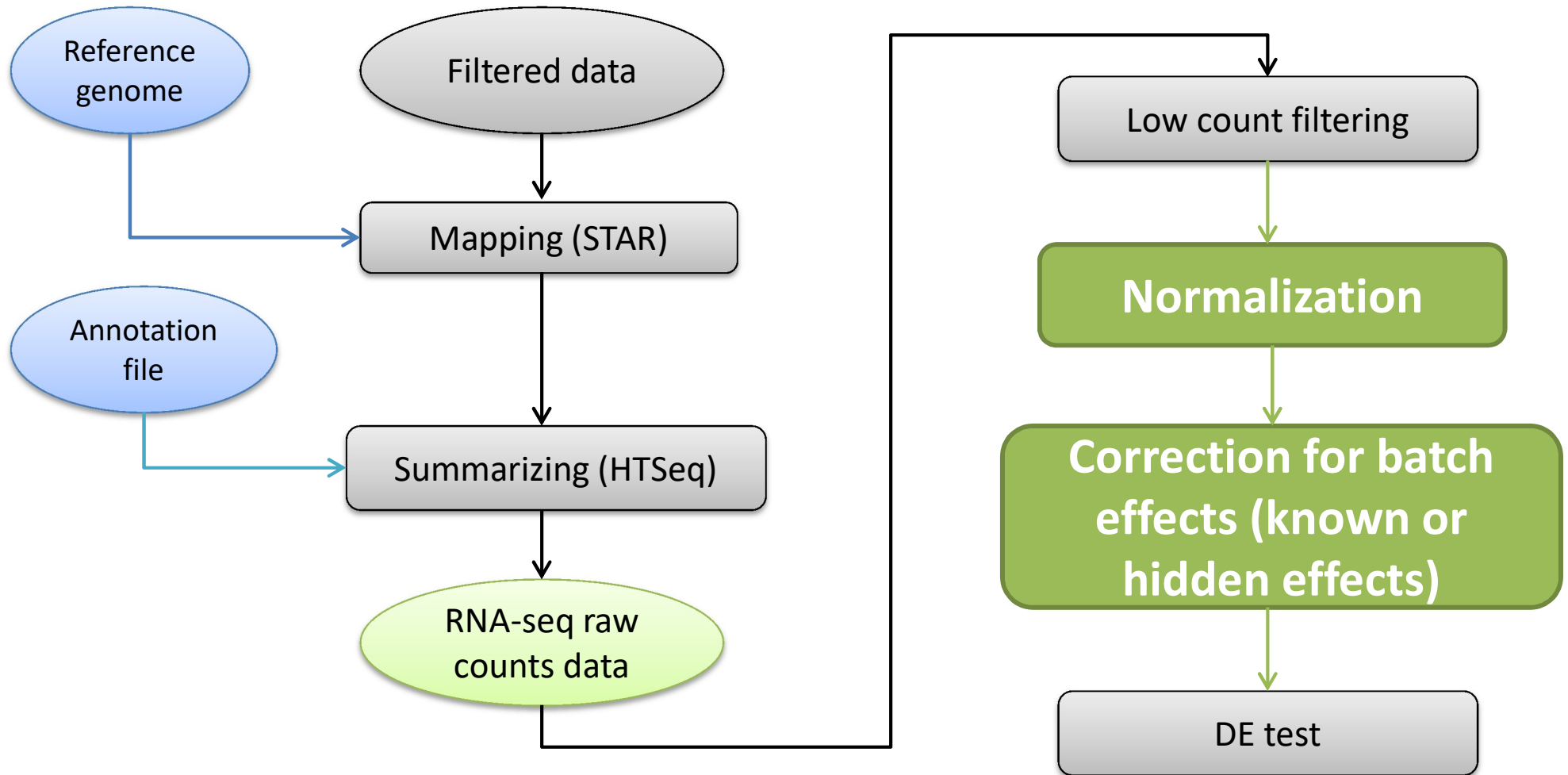


What next ? Differential Expression (DE) analysis

- Differential expression analysis means taking the normalized read counts data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups.
- Aim: identify genes that are differentially expressed between two conditions/groups




Typical DE pipeline



RNA-seq raw count matrix

Condition 1

Condition 2



Genes	Sample#1	Sample#2	Sample#3	Sample#4	Sample#5	Sample#6
Gene#1	33	18	12	77	33	40
Gene#2	2	1	0	2	3	4
Gene#3	1233	233	2200	120	2900	3300
Gene#4	544	88	110	23	129	455



Normalization

“ Normalization is a data analysis technique that adjusts global properties of measurements for individual samples so that they can be appropriately compared ”

[Jeffrey T. Leek et al. 2010]



Aims of normalization

- Normalization (including correction for batch effects) has a great impact on Differential Expression (DE) results (Bullard et al. 2010)
- Accurate estimation of gene expression levels
- Reliable DE analysis
- Reduce FP DE genes



A good normalization method including correction for batch effects (even known or hidden) must be carefully selected and applied



Biases

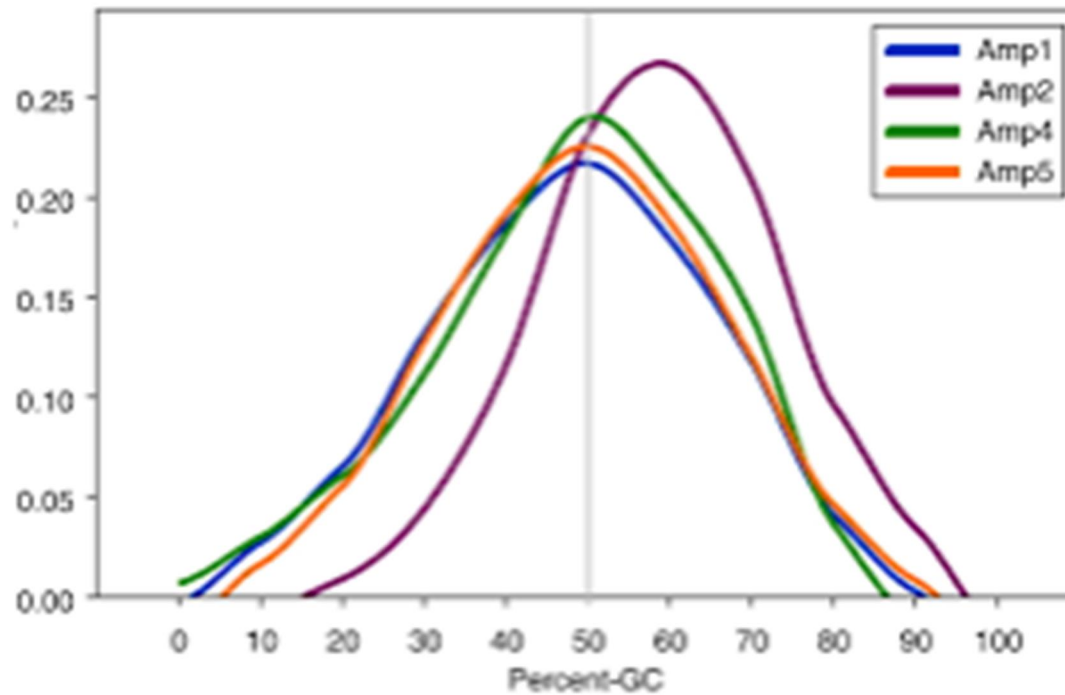
- Within sample biases: gene length, nucleotide composition (GC content), ...
- Between sample biases: library size (aka. sequencing depth), known and potential unknown batch effects



Gene length bias



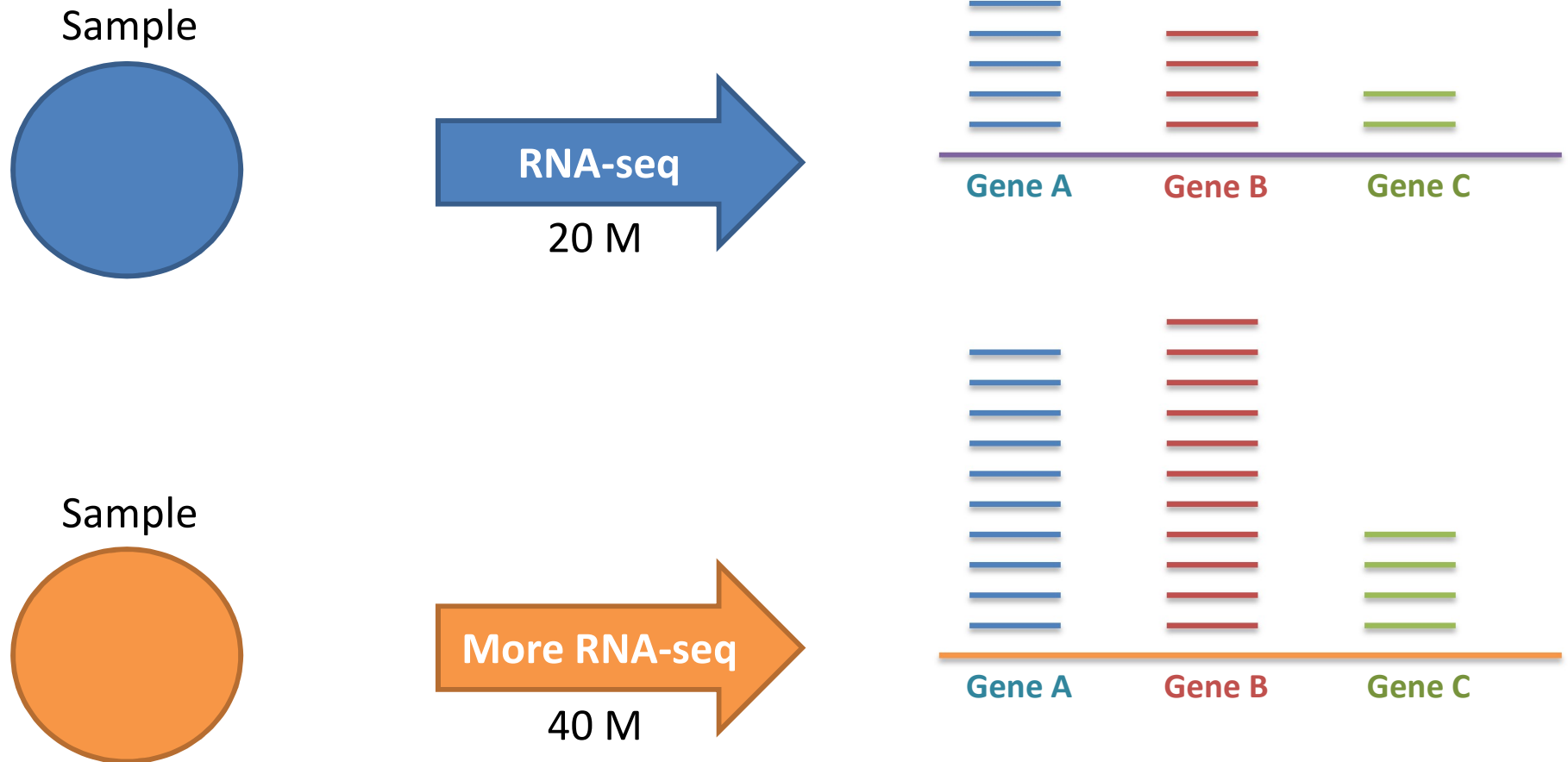
GC content bias



[Tarazona S et al. 2012]



Library size (aka. Sequencing depth) bias



Normalization methods

- RPKM [Mortazavi et al. 2008]
 - UQ, TC
 - CQN [Hansen et al. 2012]
 - TMM (edgeR) [Robinson MD et al. 2010]
 - DESeq2 [Anders S et al. 2010]
- distribution adjustment of read counts
- Library size

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis [Dillies M.A. et al. 2012]



Reads Per Kilobase per Million mapped reads (RPKM)

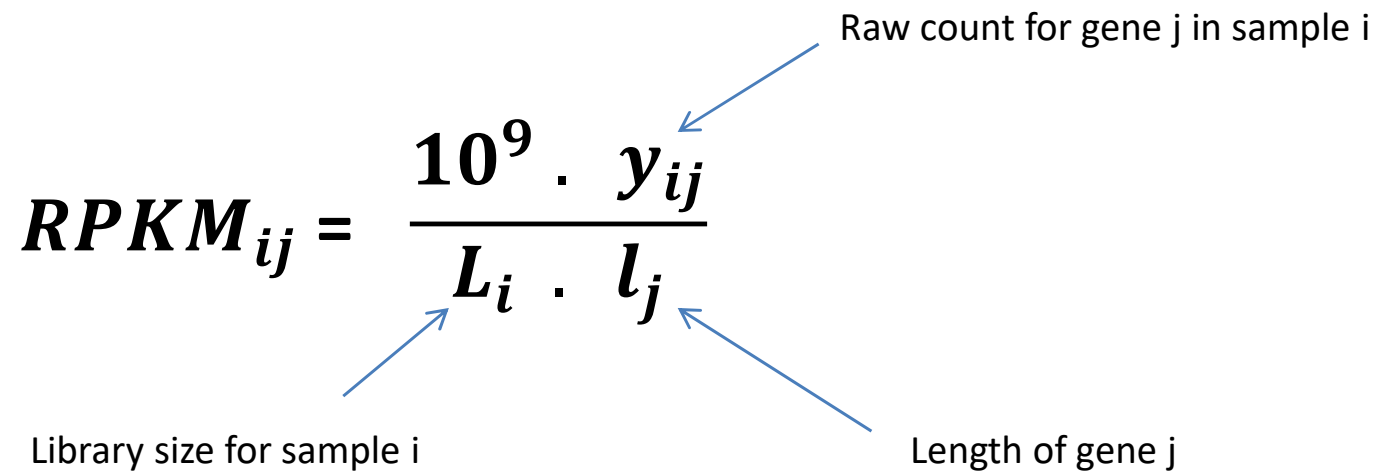
- Removing library sizes and gene length effects

$$RPKM_{ij} = \frac{10^9 \cdot y_{ij}}{L_i \cdot l_j}$$

Raw count for gene j in sample i

Library size for sample i

Length of gene j

The diagram shows the RPKM formula with three labels and arrows. An arrow points from 'Raw count for gene j in sample i' to the variable y_{ij} in the numerator. Another arrow points from 'Library size for sample i' to the variable L_i in the denominator. A third arrow points from 'Length of gene j' to the variable l_j in the denominator.



Total Count (TC) and Upper Quartile (UQ) normalizations

- Total Count

$$TC_{ij} = \frac{y_{ij} \cdot \sum_{i=1}^n L_i}{L_i \cdot n}$$

Total number of samples



Total Count (TC) and Upper Quartile (UQ) normalizations

- Total Count

$$TC_{ij} = \frac{y_{ij} \cdot \sum_{k=1}^n L_k}{L_i \cdot n}$$

Total number of samples

- Upper Quartile

$$UQ_{ij} = \frac{y_{ij} \cdot \sum_{k=1}^n UQ_k}{UQ_i \cdot n}$$



DESeq

- **Assumption:** Most genes are equivalently expressed (EE) across samples.



DESeq

- **Assumption:** Most genes are equivalently expressed (EE) across samples.
- An estimated size factor \hat{S}_j (scalable factor) is calculated for each sample j .
- Scale the counts to the corresponding size factor for each sample



DESeq – Calculating of size factors

1. Relative expression of gene i in sample j :

$$e_{ij} = \frac{k_{ij}}{\left(\prod_{v=1}^m k_{iv}\right)^{1/m}}$$

Read count for gene i in sample j

Geometric mean of gene i across all the samples



DESeq – Calculating of size factors

1. Relative expression of gene i in sample j :

$$e_{ij} = \frac{k_{ij}}{\left(\prod_{v=1}^m k_{iv}\right)^{1/m}}$$

Read count for gene i in sample j

Geometric mean of gene i across all the samples

2. Estimated size factor for sample j :

$$\hat{s}_j = \text{median} \{ e_{ij} : \text{gene } i \text{ is EE across the samples} \}$$



Trimmed Mean of M-values (TMM)

- The same principle as DESeq
- Assume most of genes are not differentially expressed across samples.
- edgeR package



Trimmed Mean of M-values (TMM)

- The same principle as DESeq
- Assume most of genes are not differentially expressed across samples.
- edgeR package
- TMM normalization factors across several samples can be calculated by selecting one sample as a reference and calculating the TMM factor for each non-reference sample



TMM normalization factors

- Steps:

1. Calculate the M-value for each gene g

$$M_g = \log_2 \frac{y_{gk} / N_k}{y_{gk'} / N_{k'}} \quad k' \text{ reference sample}$$



TMM normalization factors

- Steps:

1. Calculate the M-value for each gene g

$$M_g = \log_2 \frac{y_{gk} / N_k}{y_{gk'} / N_{k'}} \quad k' \text{ reference sample}$$

2. Calculate the absolute expression level for each gene (A value)

$$A_g = \frac{1}{2} \log_2 \left(y_{gk} / N_k \cdot y_{gk'} / N_{k'} \right)$$



TMM normalization factors

- Steps:

3. Trimming of M-values and A-values

M_1 M_2 M_3 M_4 ... M_G

A_1 A_2 A_3 A_4 ... A_G



TMM normalization factors

- Steps:

3. Trimming of M-values and A-values

~~M_1~~ ~~M_2~~ M_3 M_4 ... ~~M_G~~

~~A_1~~ ~~A_2~~ A_3 A_4 ... ~~A_G~~



TMM normalization factors

- Steps:

3. Trimming of M-values and A-values

~~M_1~~ ~~M_2~~ M_3 M_4 ... ~~M_G~~

~~A_1~~ ~~A_2~~ A_3 A_4 ... ~~A_G~~

4. Calculate the weighted mean of the remaining M-values



TMM normalization factors

- Steps:

5. We use the set of the genes with a valid M-value and A-value to calculate the TMM normalization factor for each sample using a reference sample



EDASeq to correct for GC content bias

GC-Content Normalization for RNA-Seq Data

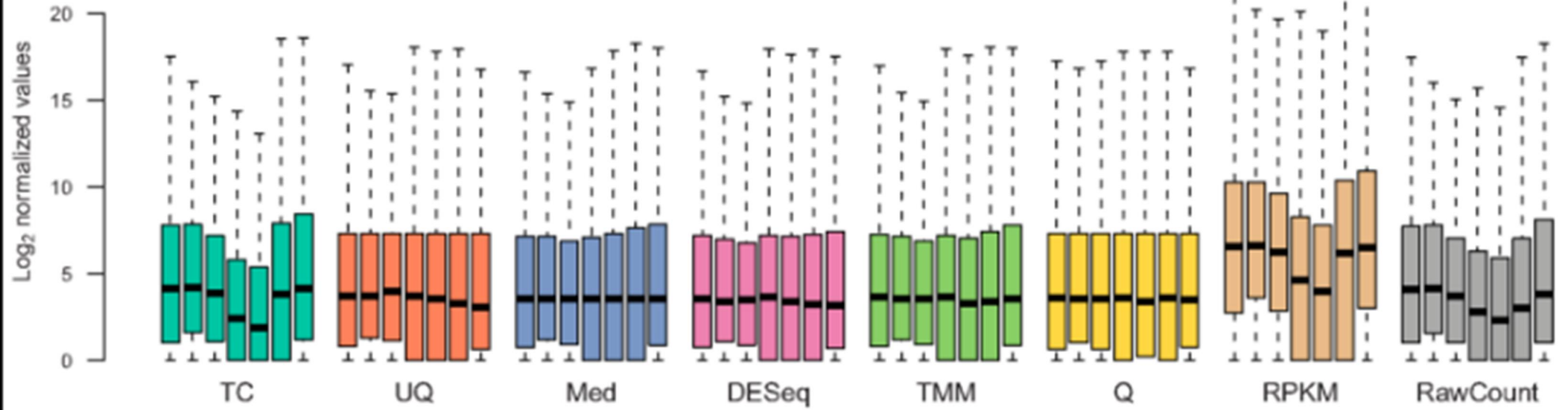
Davide Risso¹, Katja Schwartz², Gavin Sherlock² and Sandrine Dudoit^{3*}

Abstract

Background: Transcriptome sequencing (RNA-Seq) has become the assay of choice for high-throughput studies of gene expression. However, as is the case with microarrays, major technology-related artifacts and biases affect the resulting expression measures. Normalization is therefore essential to ensure accurate inference of expression levels and subsequent analyses thereof.



Comparison



[Dillies et al. 2012]



Normalization is not enough !

- Normalization does not remove batch effects, which affect specific subsets of genes and may affect different genes in different ways [Davide Risso et al. 2015]
- Most of the normalization methods proposed in the literature don't correct for **unknown** batch effects: RPKM, TMM, UQ, DESeq, ...



Batch effects

- Different sequencing centers
- Chemical reagent lots,
- Personnel
- Date of the experiment, and,
- Many other unknown technical variation



What if the batch effect is unknown ?

PCA plot shows no clustering of the samples according to the factor of interest



What if the batch effect is unknown ?

PCA plot shows no clustering of the samples according to the factor of interest → a hidden effect that hampers the data to be clustered



What if the batch effect is unknown ?

PCA plot shows no clustering of the samples according to the factor of interest → a hidden effect that hampers the data to be clustered → Hidden noise(s) to be determined and removed



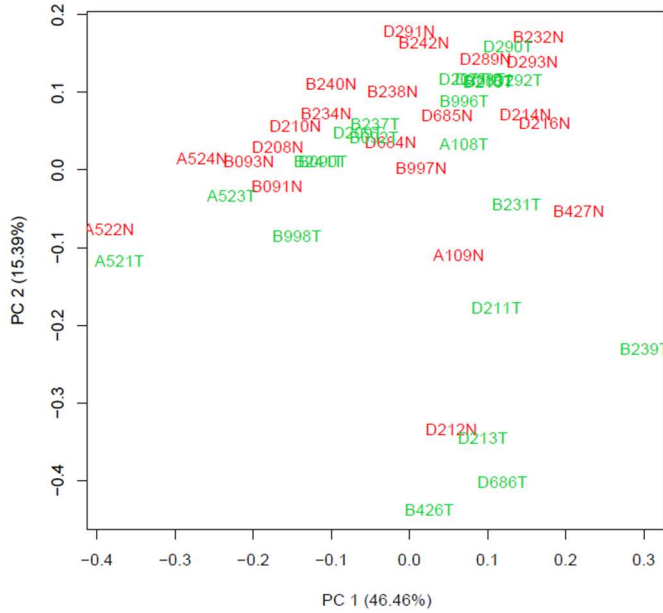
Methods to correct for unknown batch effects

- SVA [Jeffrey T. Leek, 2014]
- Combat [Johnson WE et al. 2007]
- RUVg [Davide Risso et al. 2014]
- ARSyN [Maria j. Nueda et al. 2012]

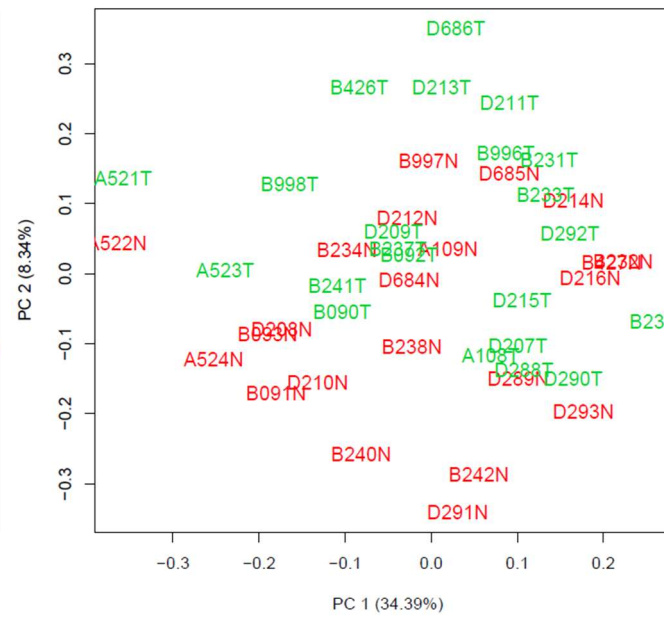


Example of removing batch effects

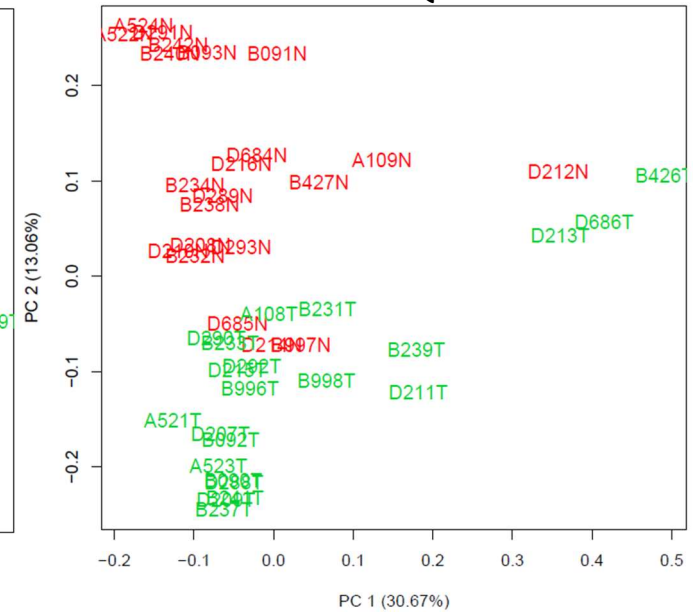
Raw data



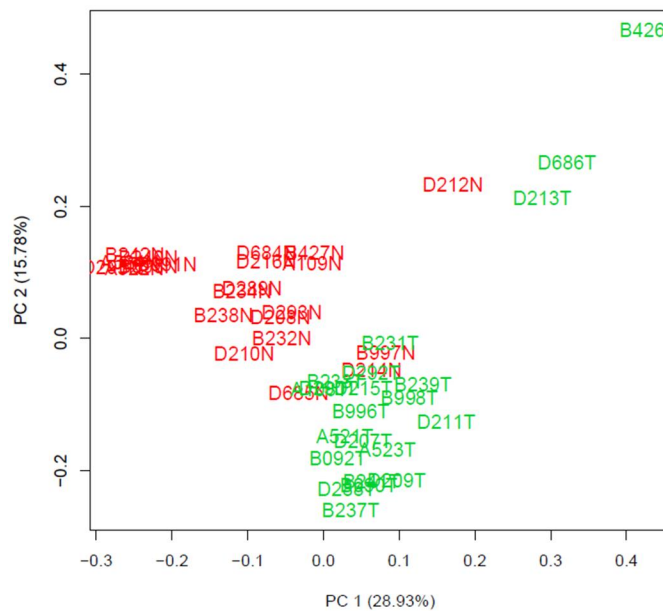
RPKM



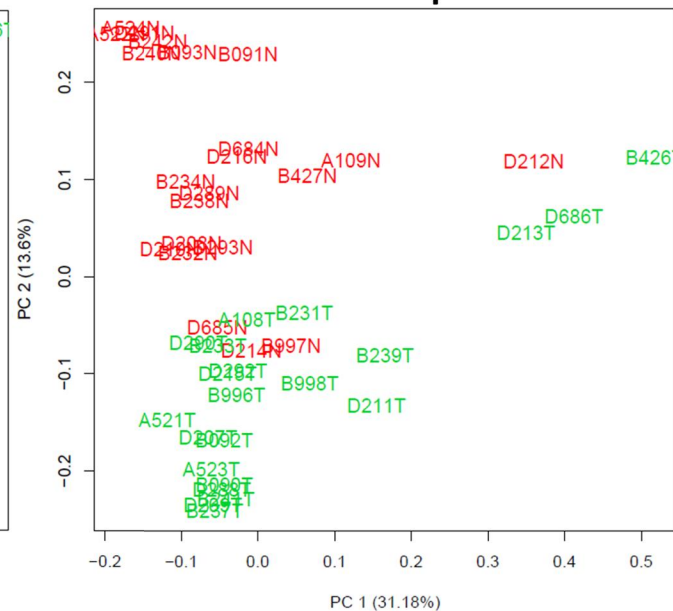
UQ



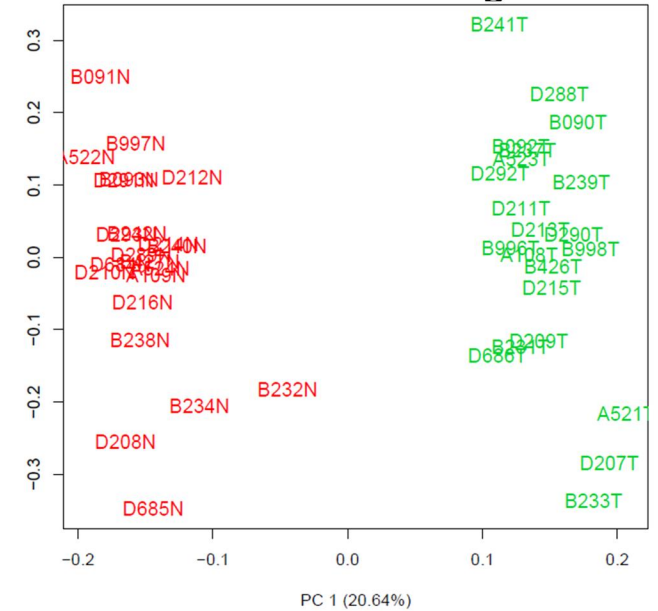
TMM



DESeq



After removing UV



What next



- Differential expression analysis
- Alternative splicing analysis



Thank you

Questions ?

