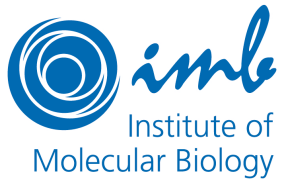
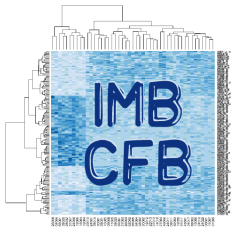
Design and Analysis of NGS Experiments

Nastasja Kreim & Anke Busch

Bioinformatics Core Facility
Institute of Molecular Biology, Mainz





July 6th, 2020

imb
Institute of
Molecular Biology

Experimental Design

Why experimental design?

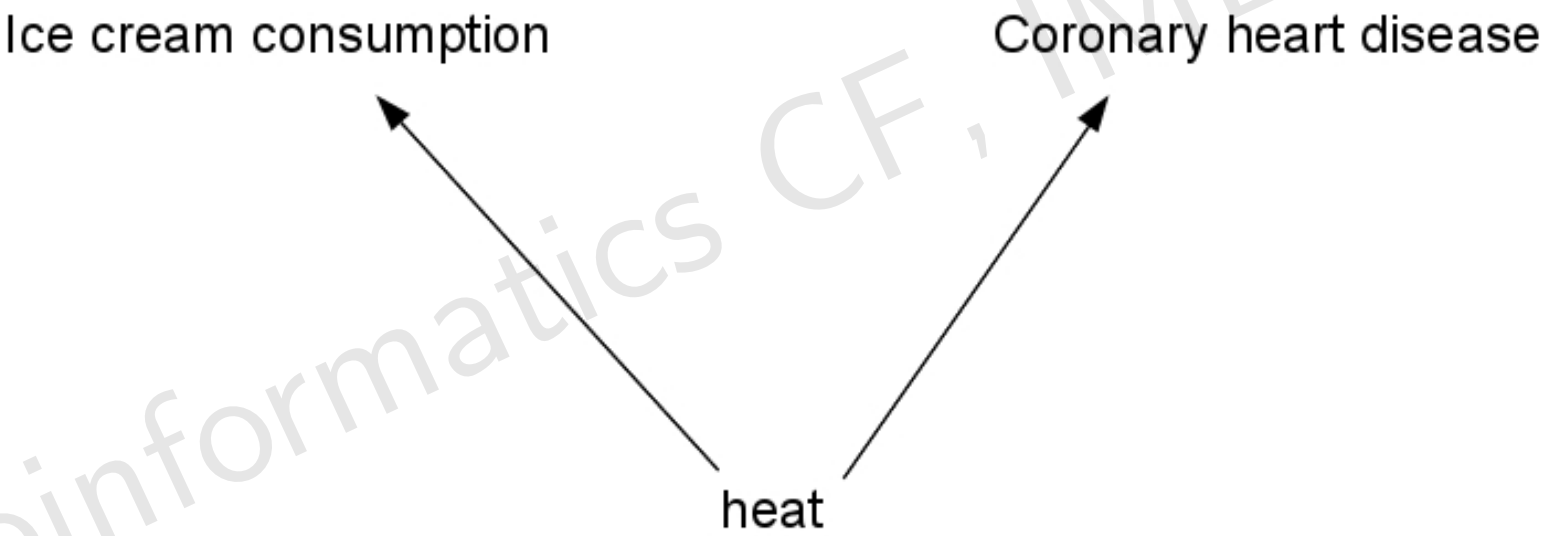
- to enable unbiased comparison between subjects, conditions, treatment groups
- to account for random variation
- to establish a relationship between cause and effect
- to disentangle biological variability from technical variability
- to enable the generalisation of findings

Hypothesis-driven research

- Is drug A better than drug B?
- Is there a genetic interaction between gene X and gene Y?
- Are transfected cells behaving differently than control cells?

Correlation does not mean Causation

Association does not mean Causation



Correlation does not mean Causation

Variability

Modes of Variability

- biological variability between subjects /samples
- variability between conditions /groups
- technical variability (e.g. sample extraction, library preparation)

$\Rightarrow$ we want to determine the variability between groups

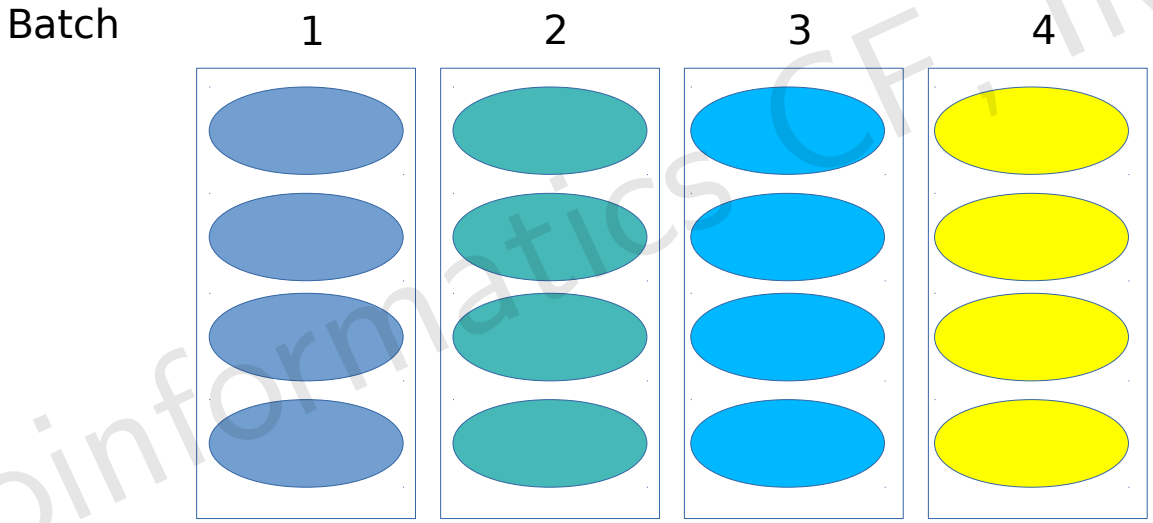
Basic rules of experimental design

- Blocking for known confounding factors
- Randomisation for unknown confounding factors
- Replication to estimate the variability within a group/condition

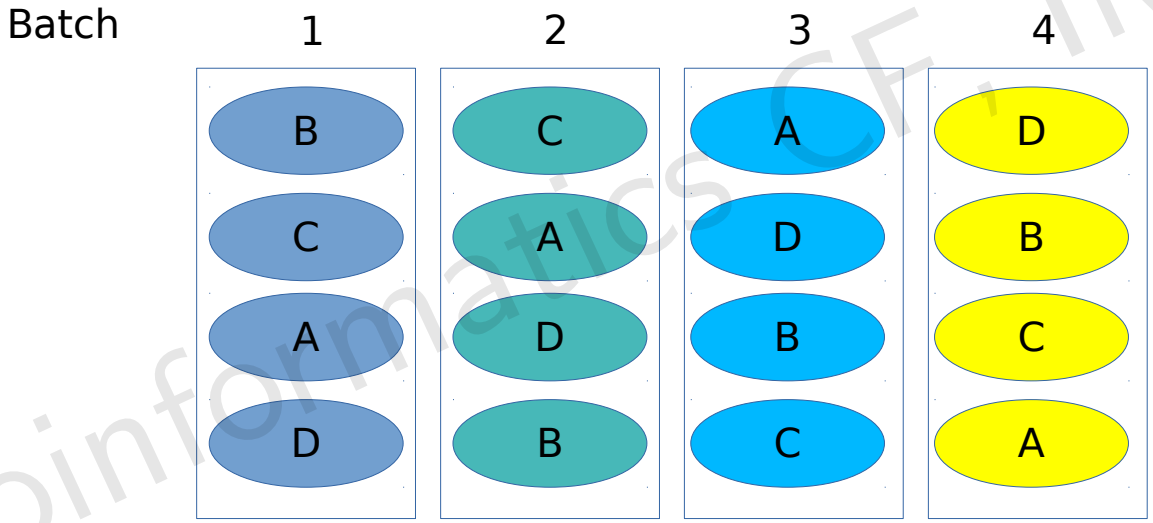
Blocking

Creation of homogeneous sample sets (for known confounding factors) with a varying factor of interest.

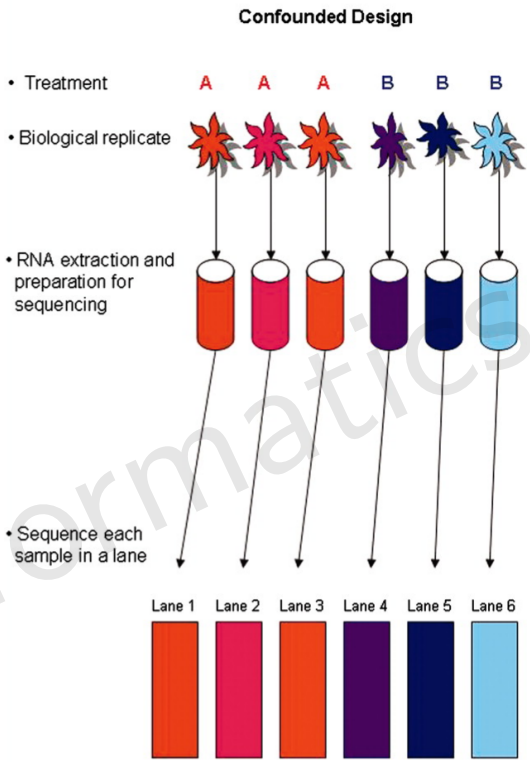
Randomised Block Design



Randomised Block Design

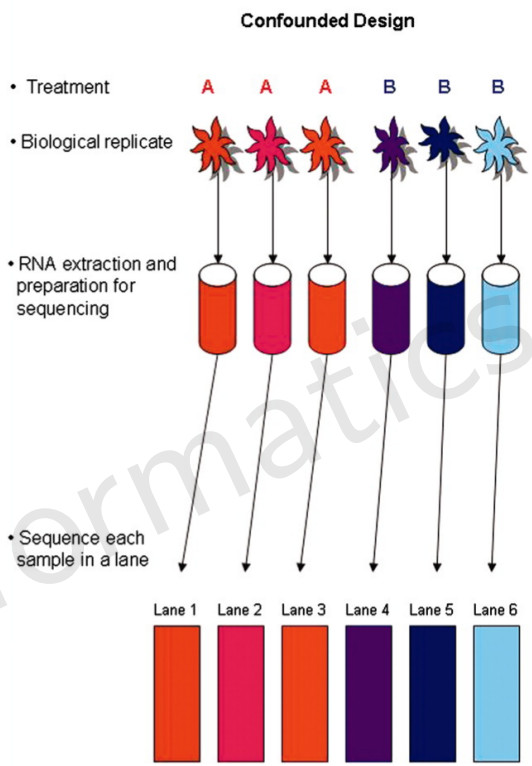


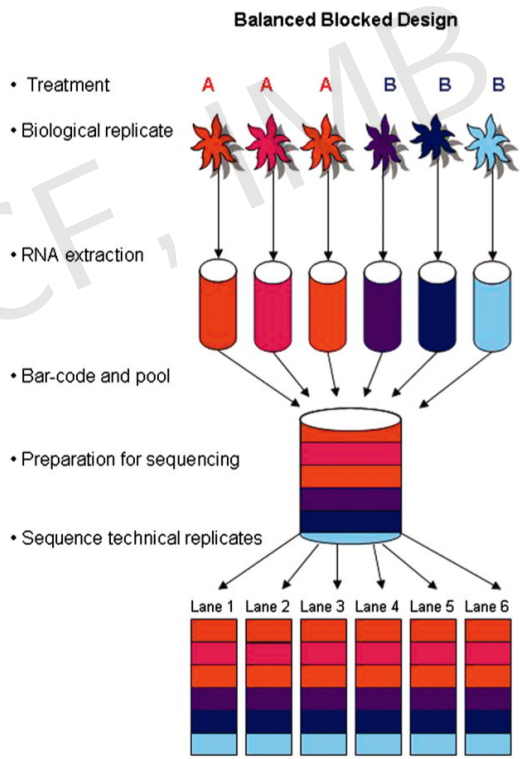
Example: flowcell design for testing differential expression



Auer, Paul L., and R. W. Doerge. "Statistical design and analysis of RNA sequencing data." Genetics 185.2 (2010): 405-416.

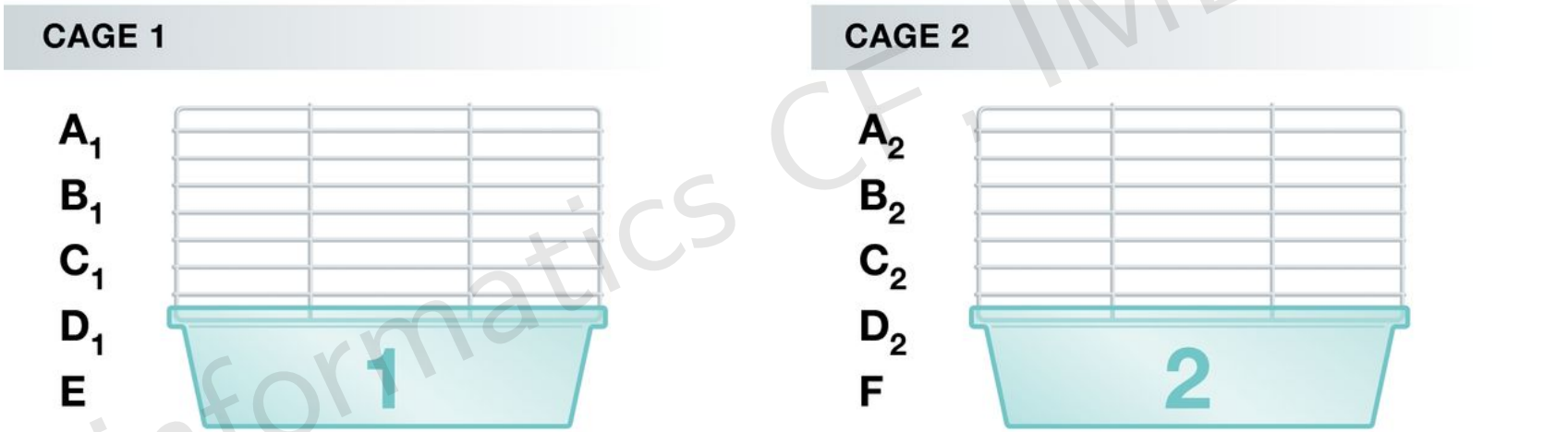
Example: flowcell design for testing differential expression





Auer, Paul L., and R. W. Doerge. "Statistical design and analysis of RNA sequencing data." Genetics 185.2 (2010): 405-416.

Example: Cages



$$\text{CAGE EFFECT} \quad \frac{1}{4} \times \left[(A_1 - A_2) + (B_1 - B_2) + (C_1 - C_2) + (D_1 - D_2) \right]$$

Institute of Molecular Biology

Klaus, Bernd. "Statistical relevancerelevant statistics, part I." The EMBO journal 34.22 (2015): 2727-2730.

Replication

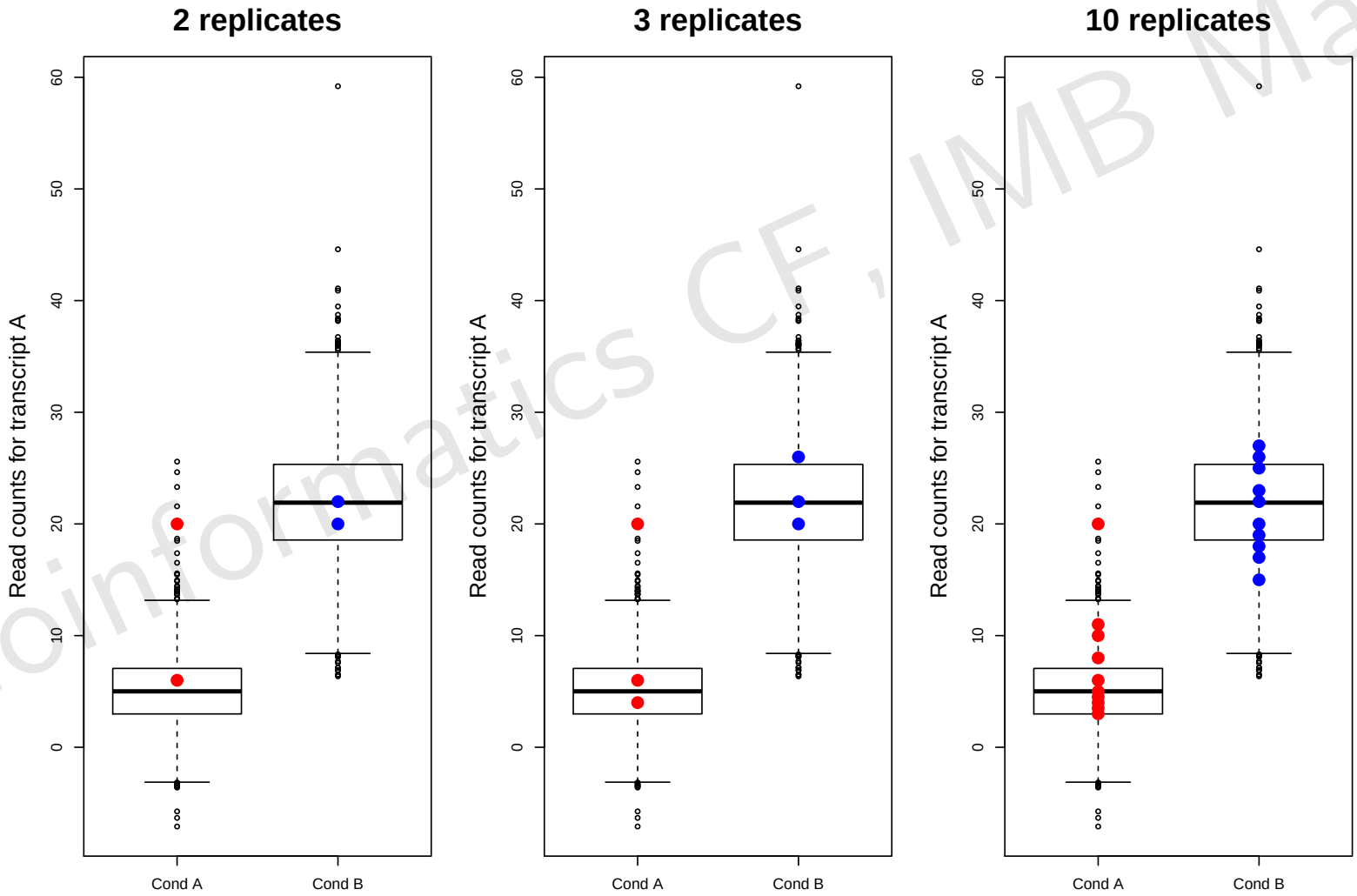
- to estimate the effect size
- to estimate how precise the effect size estimates are (SD)

Replication

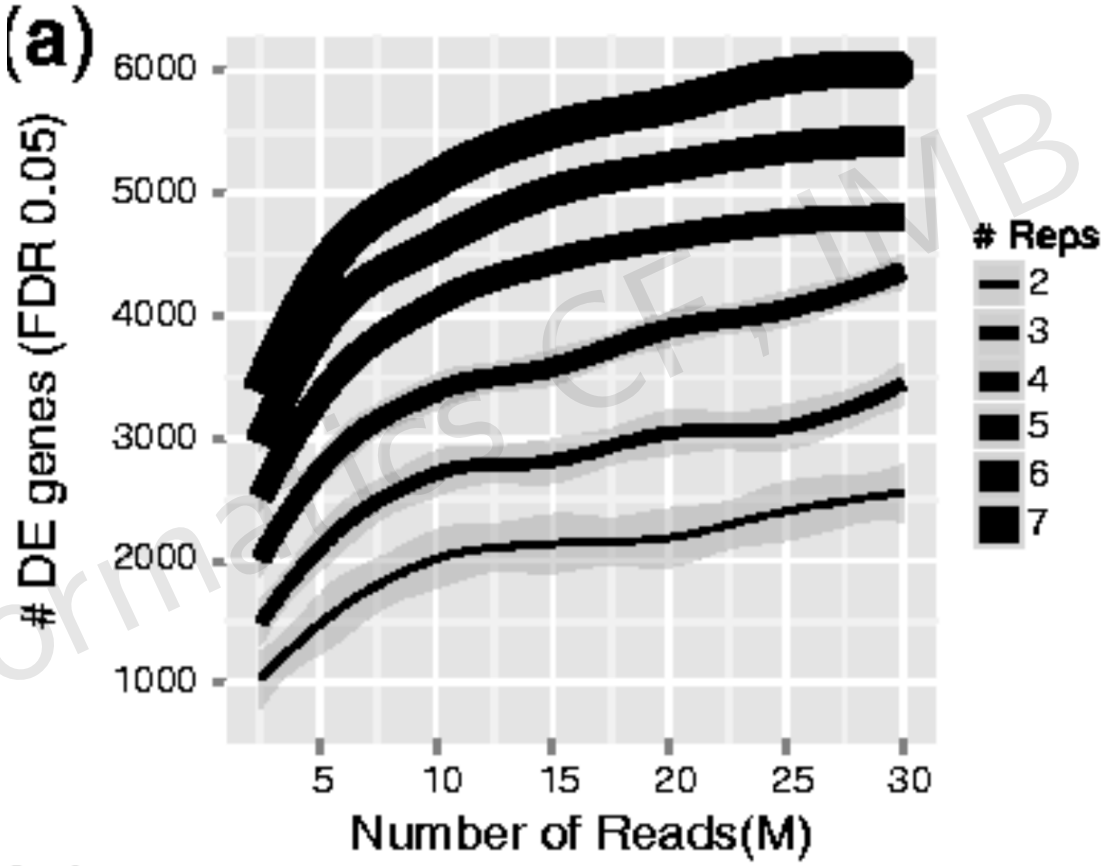
- to estimate the effect size
- to estimate how precise the effect size estimates are (SD)

$\Rightarrow$ to generalise findings

Replication

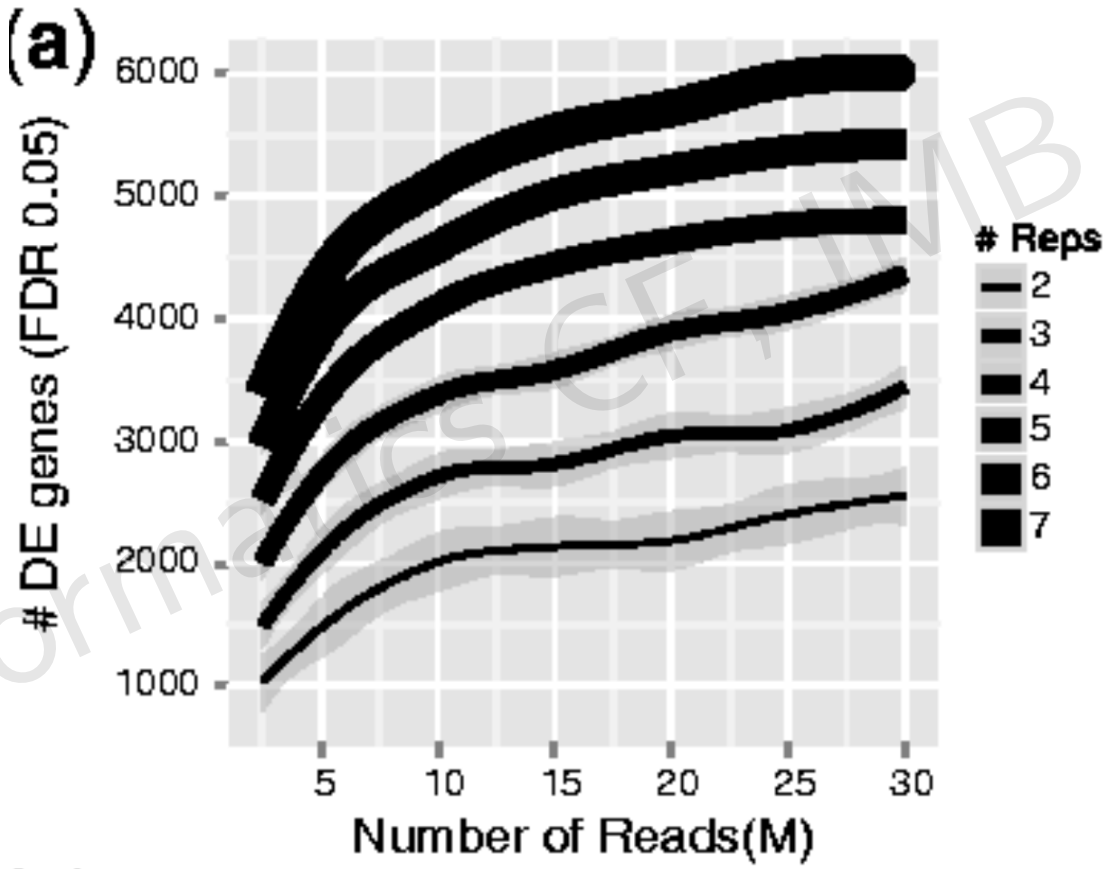


Replication



Liu, Yuwen, Jie Zhou, and Kevin P. White. "RNA-seq differential expression studies: more sequence or more replication?." Bioinformatics 30.3 (2014): 301-304.

Replication



$\Rightarrow$ **Replicates over depth!**

Liu, Yuwen, Jie Zhou, and Kevin P. White. "RNA-seq differential expression studies: more sequence or more replication?." Bioinformatics 30.3 (2014): 301-304.

Replicates

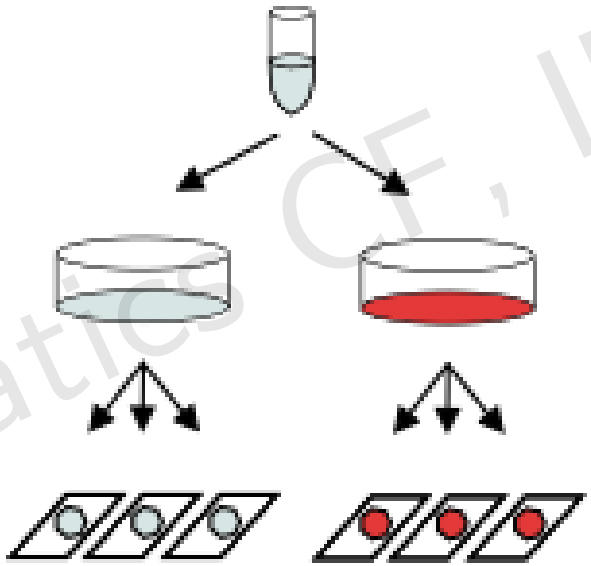
Technical Replicates

Technical replicates are replicates which have the same biological sample as origin and are processed and measured multiple times.

Biological Replicates

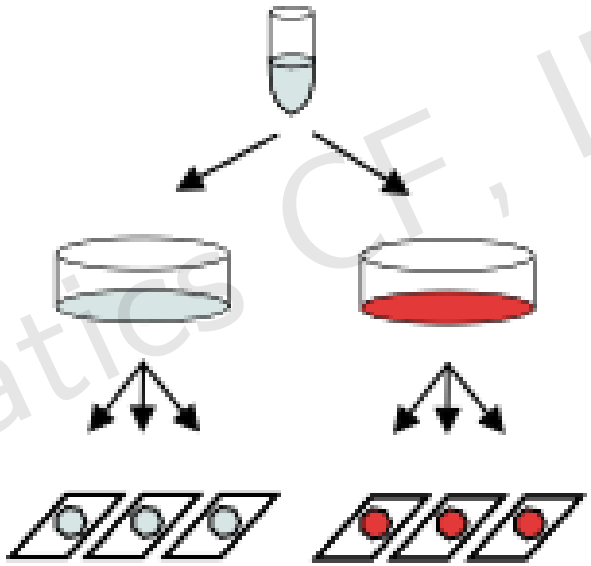
- in vivo: samples from different individuals
- in vitro: are there biological replicates?

In vitro: Cell culture



http://labstats.net/articles/cell_culture_n.html

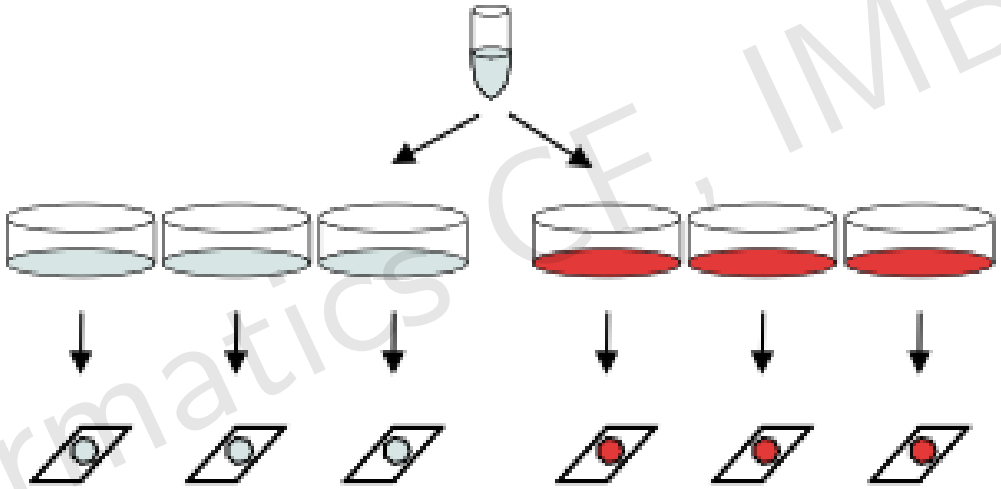
In vitro: Cell culture



$\Rightarrow$ number of biological replicates is 1 !

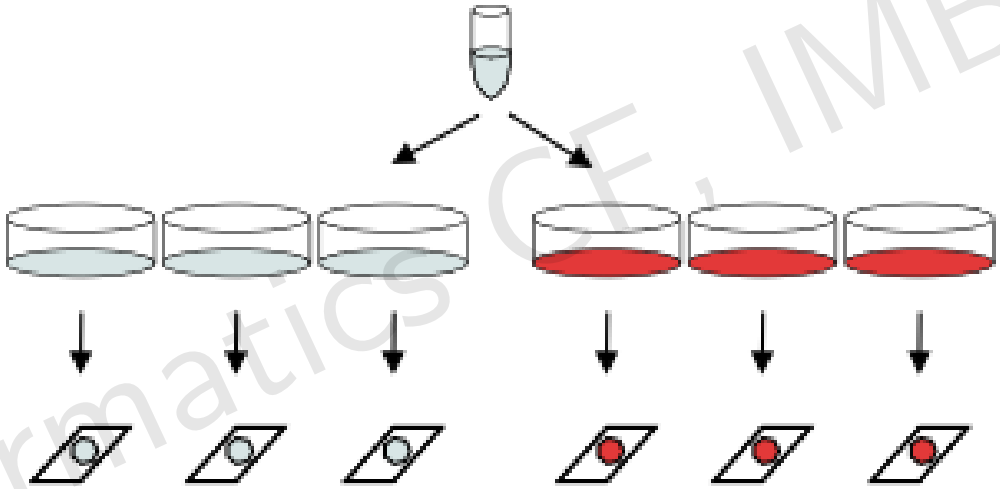
http://labstats.net/articles/cell_culture_n.html

In vitro: Cell culture



http://labstats.net/articles/cell_culture_n.html

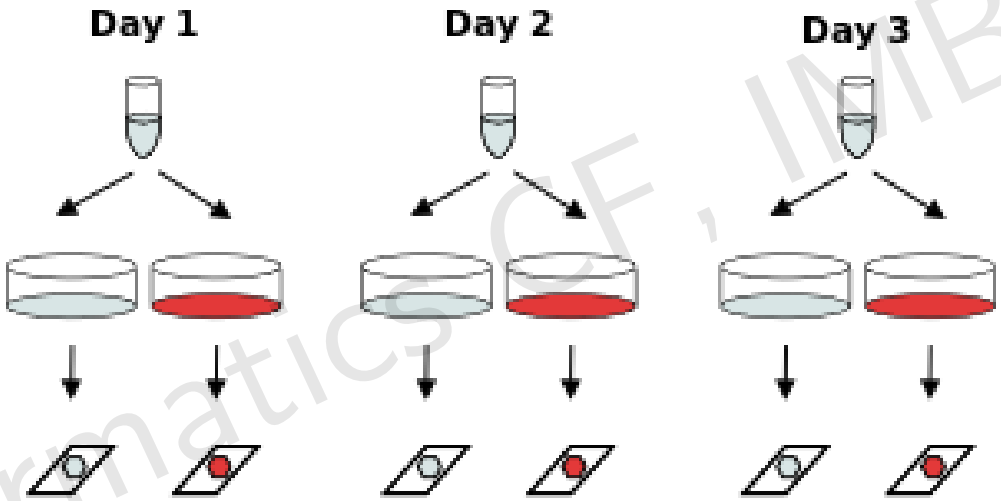
In vitro: Cell culture



$\Rightarrow$ a little bit better. More variance than before through split up higher in the hierarchy.

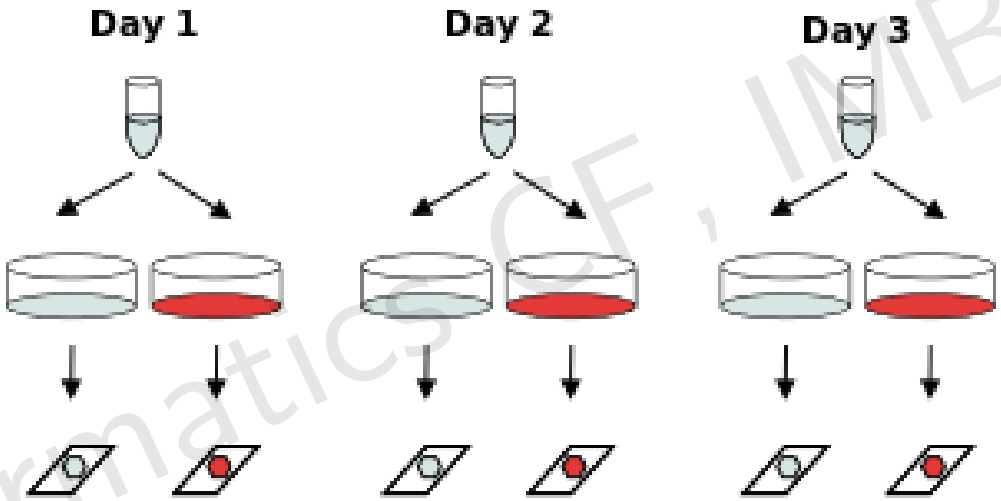
http://labstats.net/articles/cell_culture_n.html

In vitro: Cell culture



http://labstats.net/articles/cell_culture_n.html

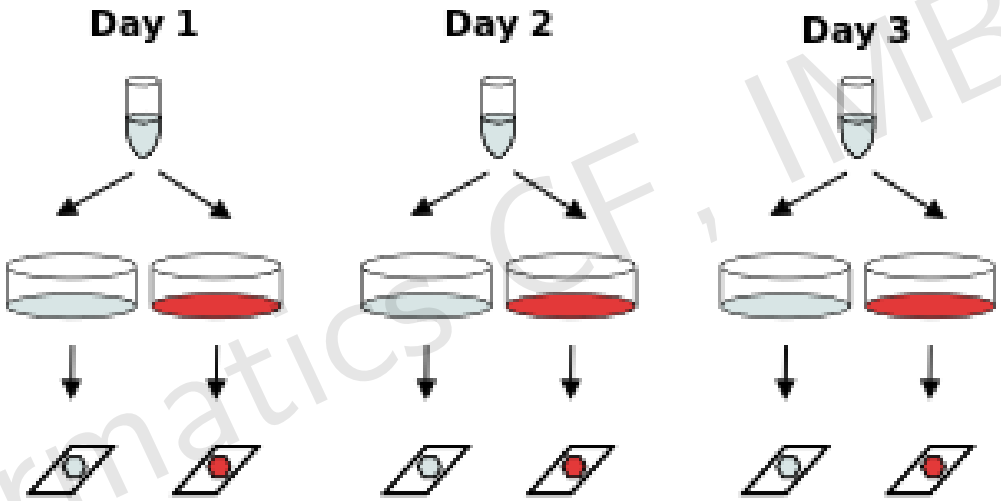
In vitro: Cell culture



Not ideal either maybe the best solution depending on the circumstances.

http://labstats.net/articles/cell_culture_n.html

In vitro: Cell culture



Not ideal either maybe the best solution depending on the circumstances.
$\Rightarrow$ ideal would be to have cell cultures from different individuals of the same cell type.

http://labstats.net/articles/cell_culture_n.html

Control groups

- confirm validity of the experiment
- reference an experimental manipulation is compared to
- positive and negative controls should be included
- control groups should be as similar as possible to your experimental groups

Summary I

- Define your hypothesis and formulate your expectations
- Use an appropriate control
- Replicate
- Block for known confounding factors
- Randomise for unknown confounding factors

After the experiment...

Data analysis

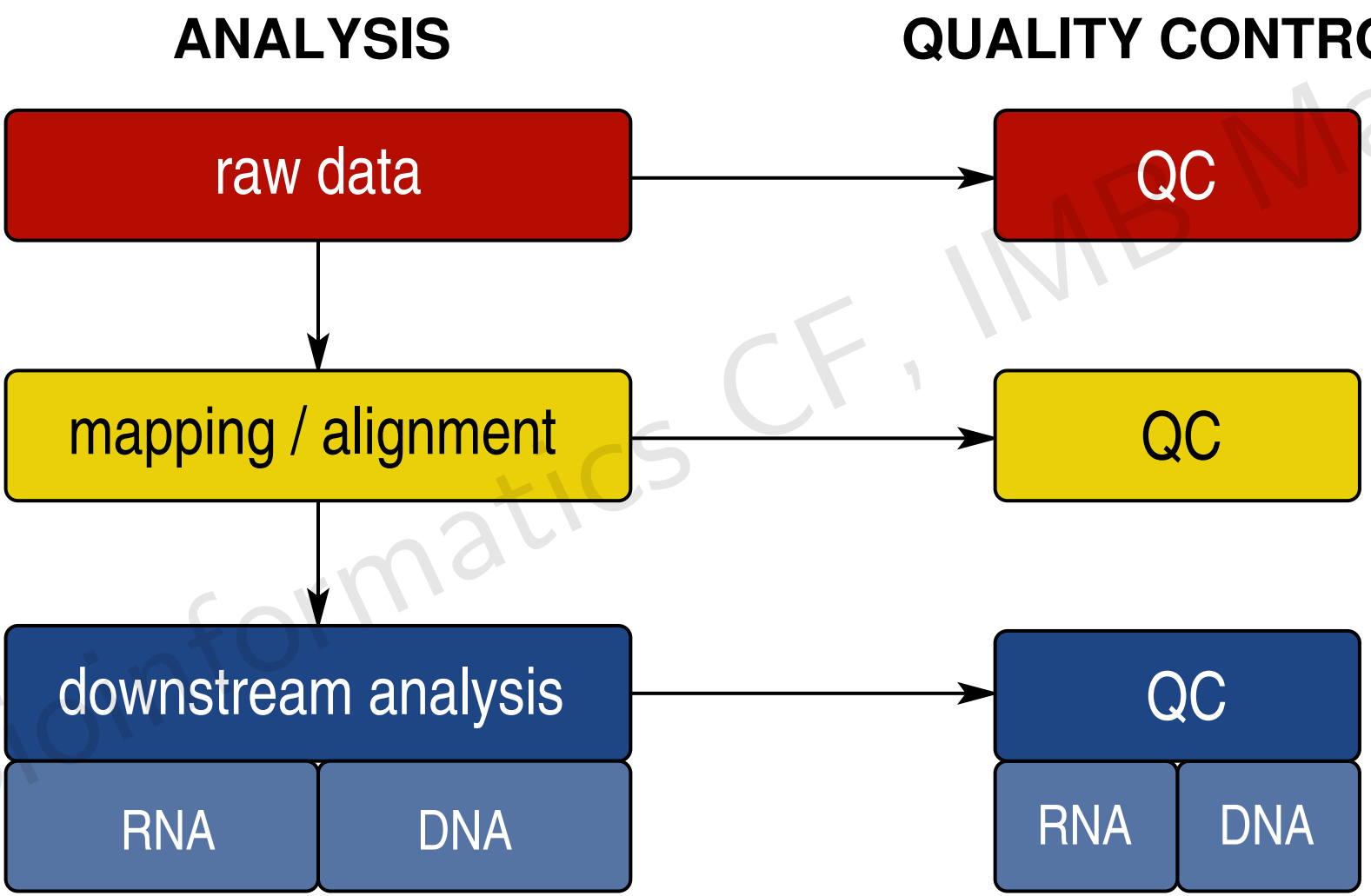
Reminder: RNA-seq and ChIP-seq

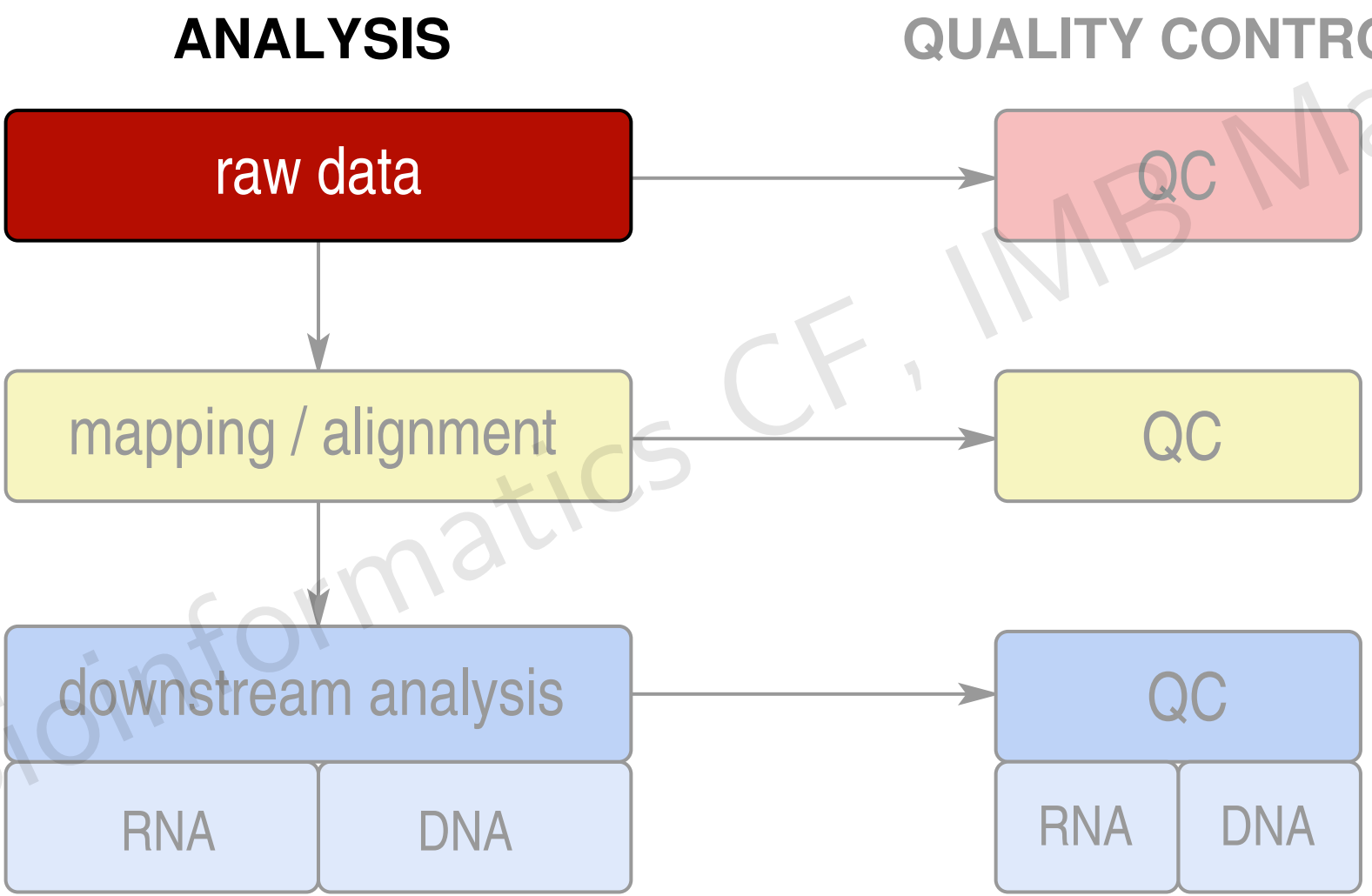
RNA-seq

- sequence expressed (m)RNA
- **Goal:** find differences in expression / splicing

ChIP-seq

- sequence DNA bound to DNA-binding proteins
- **Goal:** find binding sites of DNA-binding protein





Short reads

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG
+
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI
@HWI-ST558:257:C4AJHACXX:1:1101:2458:2678 1:N:0:ATCACG
AGGTAAACAGACCATTGGATGGGAGATAGCAAGAACAATAGACTCCCTCAG
+
:@?4ADDDFFDBDGFEBGF<<;A@F8<A4?FGEBBFFG<BB?@FG@D>?FB
@HWI-ST558:257:C4AJHACXX:1:1101:3208:2718 1:N:0:ATCACG
AGGAGGAGGAAGGTGATATCACTGCACAATTTTTCATCTGTTATGATCAAT
+
@CCFFFDFDDHHDAACCBCHHIIGHIIIIIIIIIGHHEGGFEFHGGEHHGH
@HWI-ST558:257:C4AJHACXX:1:1101:3358:2699 1:N:0:ATCACG
AGTGTGCCATAGAGCATGCTTGCTATTCCTACAACCCATCCTCTTCAAGCC
+
===DBBDFDHBDFHIGIGIHIIEGEHGHGH@F@FHII;GGGHGGIGGIGII
@HWI-ST558:257:C4AJHACXX:1:1101:3627:2685 1:N:0:ATCACG
TGGACATATTTTGCATATGTTATCAACATTCATTCTCAGCCCCTTAATGCA
+
BCCDFFFFHHHGHJJIJJJHHIJJJJJJJJJJIJIGGIJJJJJJJIJJIJ
```

Short reads

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG
+
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI
@HWI-ST558:257:C4AJHACXX:1:1101:2458:2678 1:N:0:ATCACG
AGGTAAACAGACCATTGGATGGGAGATAGCAAGAACAATAGACTCCCTCAG
+
:@?4ADDDFFDBDGFEBGF<<;A@F8<A4?FGEBBFFG<BB?@FG@D>?FB
@HWI-ST558:257:C4AJHACXX:1:1101:3208:2718 1:N:0:ATCACG
AGGAGGAGGAAGGTGATATCACTGCACAATTTTTCATCTGTTATGATCAAT
+
@CCFFFDFDDHHDAACCBCHHIIGHIIIIIIIIIGHHEGGFEFHGGEHHGH
@HWI-ST558:257:C4AJHACXX:1:1101:3358:2699 1:N:0:ATCACG
AGTGTGCCATAGAGCATGCTTGCTATTCCTACAACCCATCCTCTTCAAGCC
+
===DBBDFDHBDFHIGIGIHIIEGEHGHGH@F@FHII;GGGHGGIGGIGII
@HWI-ST558:257:C4AJHACXX:1:1101:3627:2685 1:N:0:ATCACG
TGGACATATTTTGCATATGTTATCAACATTCATTCTCAGCCCCTTAATGCA
+
BCCDFFFFHHHGHJJIJJJHHIJJJJJJJJJJIJIGGIJJJJJJJIJJIJ
```

Short reads

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG
+
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI
@HWI-ST558:257:C4AJHACXX:1:1101:2458:2678 1:N:0:ATCACG
AGGTAAACAGACCATTGGATGGGAGATAGCAAGAACAATAGACTCCCTCAG
+
:@?4ADDDFFDBDGFEBGF<<;A@F8<A4?FGEBBFFG<BB?@FG@D>?FB
@HWI-ST558:257:C4AJHACXX:1:1101:3208:2718 1:N:0:ATCACG
AGGAGGAGGAAGGTGATATCACTGCACAATTTTTCATCTGTTATGATCAAT
+
@CCFFFDFDDHHDAACCBCHHIIGHIIIIIIIIIGHHEGGFEFHGGEHHGH
@HWI-ST558:257:C4AJHACXX:1:1101:3358:2699 1:N:0:ATCACG
AGTGTGCCATAGAGCATGCTTGCTATTCCTACAACCCATCCTCTTCAAGCC
+
===DBBDFDHBDFHIGIGIHIIEGEHGHGH@F@FHII;GGGHGGIGGIGII
@HWI-ST558:257:C4AJHACXX:1:1101:3627:2685 1:N:0:ATCACG
TGGACATATTTTGCATATGTTATCAACATTCATTCTCAGCCCCTTAATGCA
+
BCCDFFFFHHHGHJJIJJJHHIJJJJJJJJJJIJIGGIJJJJJJJIJJIJ
```

Short reads

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG
+
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI
@HWI-ST558:257:C4AJHACXX:1:1101:2458:2678 1:N:0:ATCACG
AGGTAAACAGACCATTGGATGGGAGATAGCAAGAACAATAGACTCCCTCAG
+
:@?4ADDDFFDBDGFEBGF<<;A@F8<A4?FGEBBFFG<BB?@FG@D>?FB
@HWI-ST558:257:C4AJHACXX:1:1101:3208:2718 1:N:0:ATCACG
AGGAGGAGGAAGGTGATATCACTGCACAATTTTTCATCTGTTATGATCAAT
+
@CCFFFDFDDHHDAACCBCHHIIGHIIIIIIIIIGHHEGGFEFHGGEHHGH
@HWI-ST558:257:C4AJHACXX:1:1101:3358:2699 1:N:0:ATCACG
AGTGTGCCATAGAGCATGCTTGCTATTCCTACAACCCATCCTCTTCAAGCC
+
===DBBDFDHBDFHIGIGIHIIEGEHGHGH@F@FHII;GGGHGGIGGIGII
@HWI-ST558:257:C4AJHACXX:1:1101:3627:2685 1:N:0:ATCACG
TGGACATATTTTGCATATGTTATCAACATTCATTCTCAGCCCCTTAATGCA
+
BCCDFFFFHHHGHJJIJJJHHIJJJJJJJJJJIJIGGIJJJJJJJIJJIJ
```

FASTQ format / base qualities

Single read

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG   header
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG      sequence
+                                                        (header2)
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI     qualities
```

FASTQ format / base qualities

Single read

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG   header
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG      sequence
+                                                        (header2)
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI     qualities
```

Quality translation

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
||||||                |     |    |        |
012345............20....26...31.......40.
```

FASTQ format / base qualities

Single read

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG   header
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG      sequence
+                                                        (header2)
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI     qualities
```

Quality translation

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
||||||                |    |    |        |
012345............20....26...31.......40.
```

Phred qual. score Q

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

$$\text{or } P = 10^{\frac{-Q}{10}}$$

FASTQ format / base qualities

Single read

```
@HWI-ST558:257:C4AJHACXX:1:1101:2005:2735 1:N:0:ATCACG   header
TAGCGGAACCAAGTGAGGAACTATGCCAGAGTCTATTACCATCTGTATCTG      sequence
+                                                        (header2)
@CCFFFDFHHHHHHIIJJGIJJJJHJJIGIIHGJIIGJJJJEGIJIGHIJI     qualities
```

Quality translation

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
||||||                  |     |     |     |
012345.............20....26...31.......40.
```

Phred qual. score Q

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

$$\text{or } P = 10^{\frac{-Q}{10}}$$

Phred score Q	prob. of incorrect base call P	base call accuracy
10	$0.1 = 1$ in 10	90%
20	$0.01 = 1$ in 100	99%
30	$0.001 = 1$ in 1000	99.9%
40	$0.0001 = 1$ in 10000	99.99%

http://en.wikipedia.org/wiki/Phred_quality_score

FASTQ format / base qualities

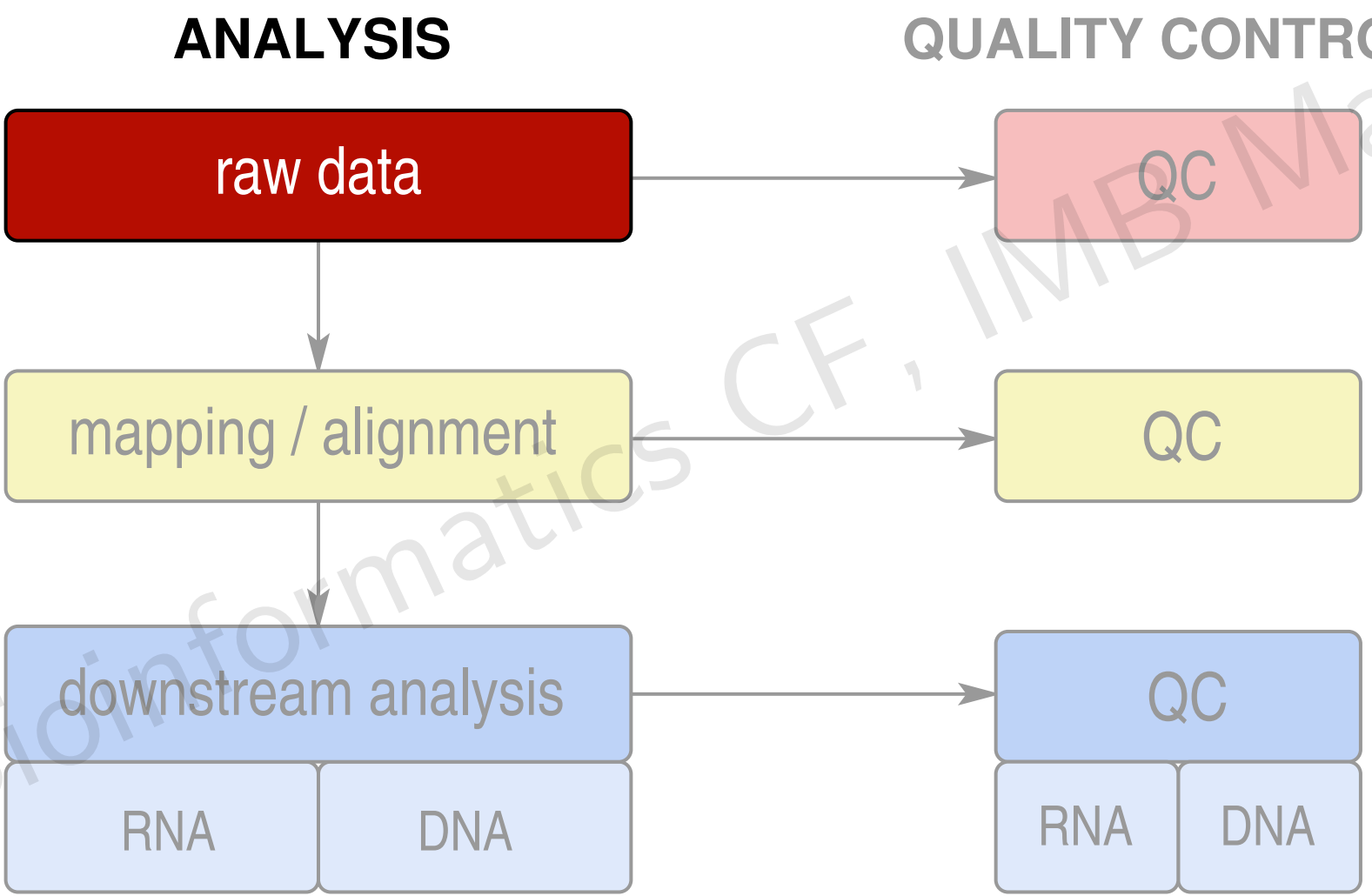
NextSeq Q-score binning

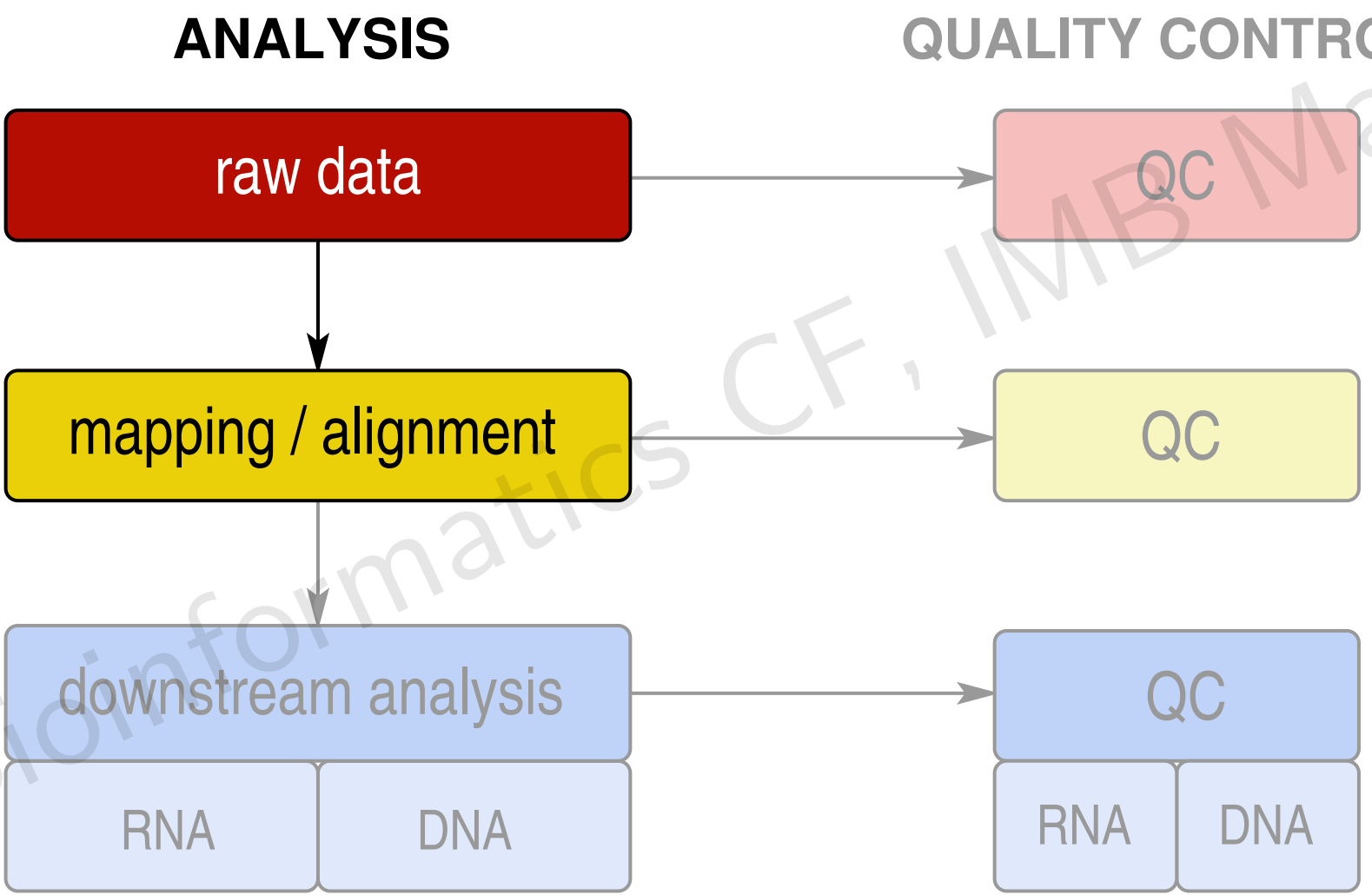
Q-score bins	new Q-score
$2-9$	6
$10-19$	15
$20-24$	22
$25-29$	27
$30-34$	33
$35-39$	37
≥ 40	40

FASTQ format / base qualities

NextSeq Q-score binning

Q-score bins	new Q-score	accuracy bins	assigned accuracy
$2-9$	6	$36.90-87.41\%$	74.88%
$10-19$	15	$90.00-98.74\%$	96.84%
$20-24$	22	$99.00-99.60\%$	99.37%
$25-29$	27	$99.68-99.87\%$	99.80%
$30-34$	33	$99.90-99.96\%$	99.95%
$35-39$	37	$99.97-99.99\%$	99.98%
≥ 40	40	$\geq 99.99\%$	99.99%





Read mapping

...ACTGATCCCATTTCATTCAAAAATCAAAATAAACTCGCCTAAATCACACAACCAAACCTAAAACT...

CTTTTCGCCA
ATTAAACGACC
AGGCA
CTTTCGACCA
ACAGGTATAC
TCAGACAAACCCT
AATTAA
TACCAGAAGGCC
GAAACGTTGAAGT
CAGACAAACCC
AAAATCAGA
GCATTGAACAGAAAG
ATTAACAACA
AAATTACACACA
ACCAGGTAAAGA
CATTACACAGAGACAAA
ACCAAAAAATTTA
ACAGTTAGAACACA
ACAGTTAGACACA
ACAGATAGACA
ACAGTTATAGCA
ACAGGATTTGTGAGAC
ACAGTAGACACAAC
ACACACAT
ACATAGACGGCAACAA
ACAG
ACAGACGATTTAACACA

Read mapping

Alignment of reads to genomic positions

- millions of short reads
- human genome: $\sim 3 * 10^9$ bp
- repetitive / low complexity regions
- sequencing: error-prone
- SNPs / InDels

$\Rightarrow$ complex task, computationally expensive

Read mapping

Alignment of reads to genomic positions

- millions of short reads
- human genome: $\sim 3 * 10^9$ bp
- repetitive / low complexity regions
- sequencing: error-prone
- SNPs / InDels

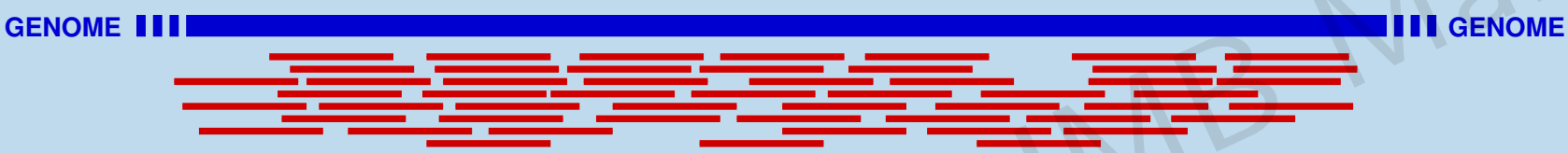
$\Rightarrow$ complex task, computationally expensive

Tools

- large number of **specialized NGS read mappers**
- **non-splice-aware:** `Bowtie`, `BWA`, **...** (differences in accuracy, speed, memory)
- **splice-aware:** `STAR`, `TopHat`, **...** (differences in accuracy, speed, memory)
- mapping **parameters** can have huge **impact** on mapping results

Types of mapping

Mapping to the genome (Bowtie, BWA)



Types of mapping

Mapping to the genome (Bowtie, BWA)

Splice-aware mapping (TopHat, STAR)

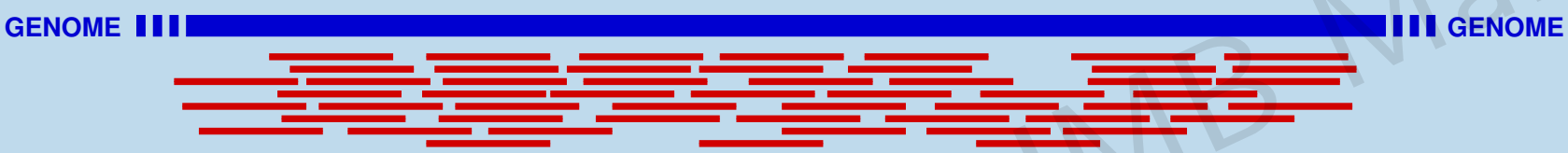
Types of mapping

Mapping to the genome (Bowtie, BWA)

Splice-aware mapping (TopHat, STAR)

Types of mapping

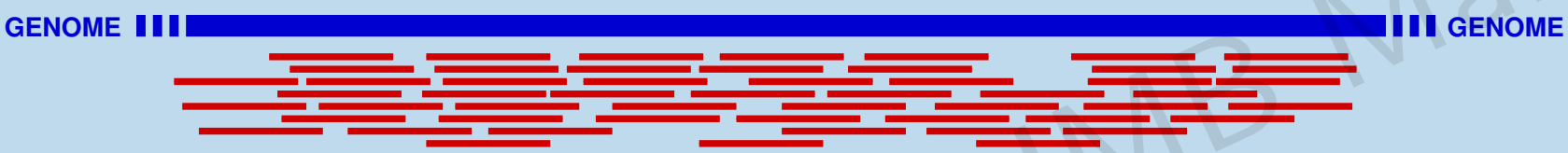
Mapping to the genome (Bowtie, BWA)



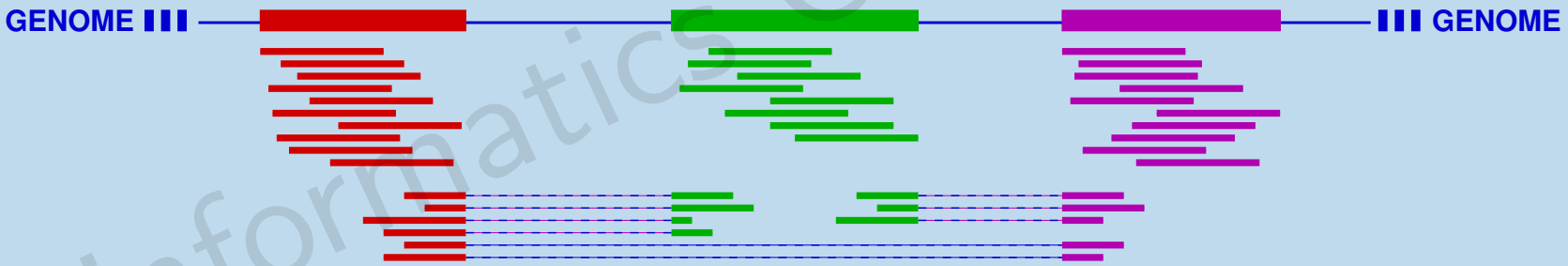
Splice-aware mapping (TopHat, STAR)

Types of mapping

Mapping to the genome (Bowtie, BWA)

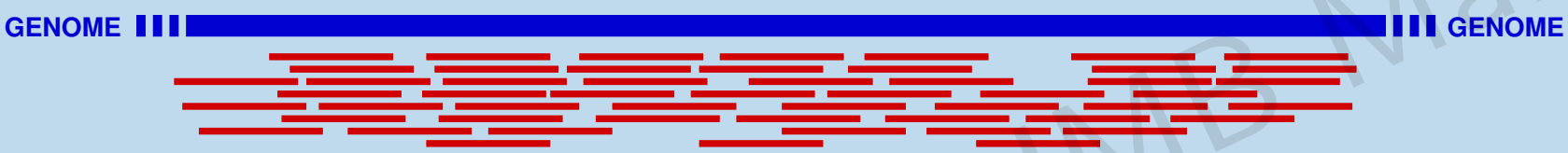


Splice-aware mapping (TopHat, STAR)

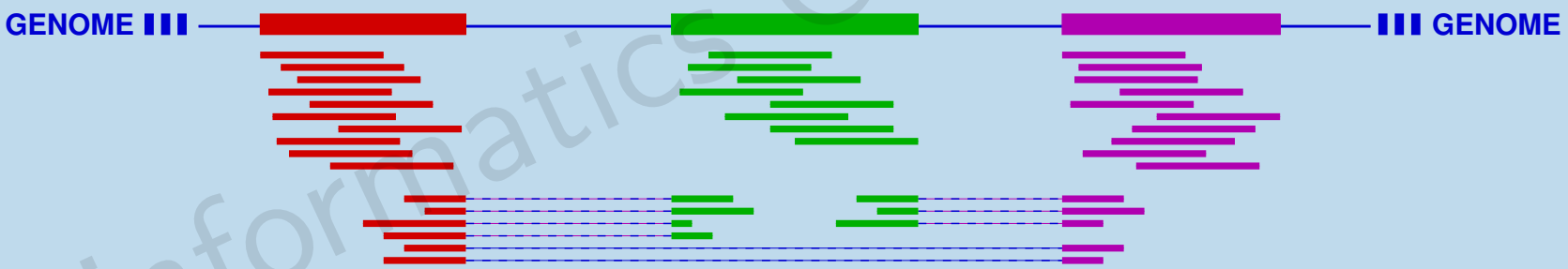


Types of mapping

Mapping to the genome (Bowtie, BWA)



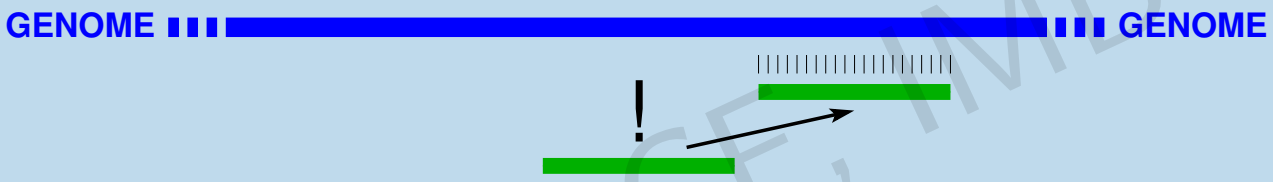
Splice-aware mapping (TopHat, STAR)



Mapping to transcripts

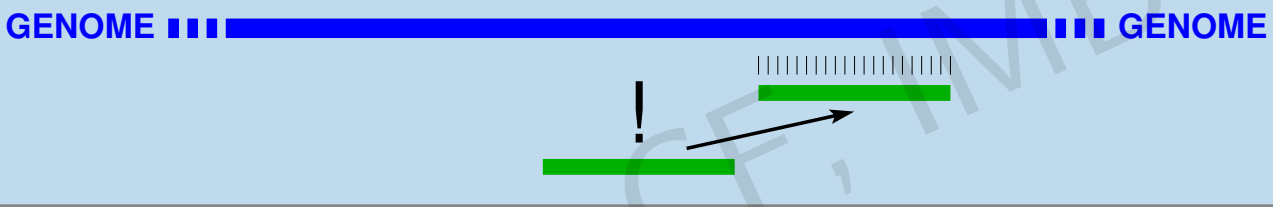
Types of mapping

Uniquely mapped reads

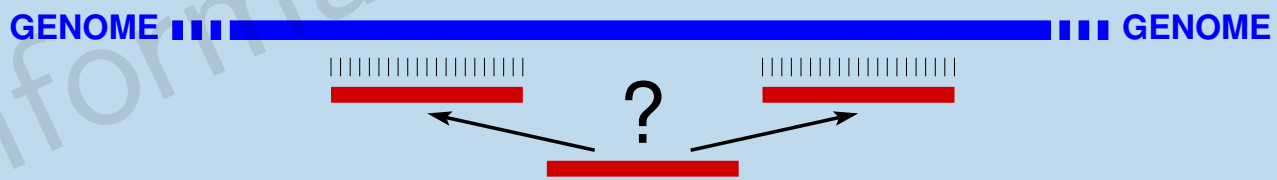


Types of mapping

Uniquely mapped reads

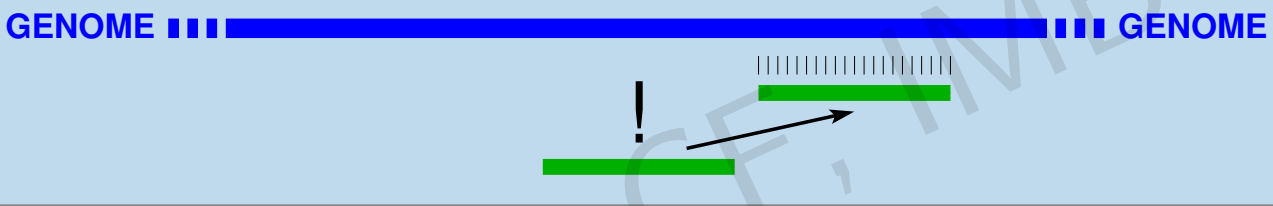


Multi-mapped reads (non-uniquely mapped)

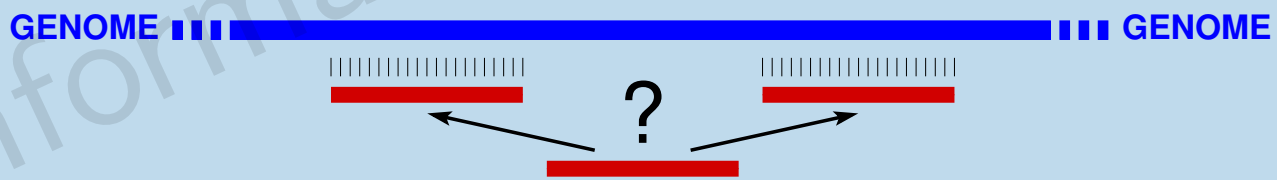


Types of mapping

Uniquely mapped reads

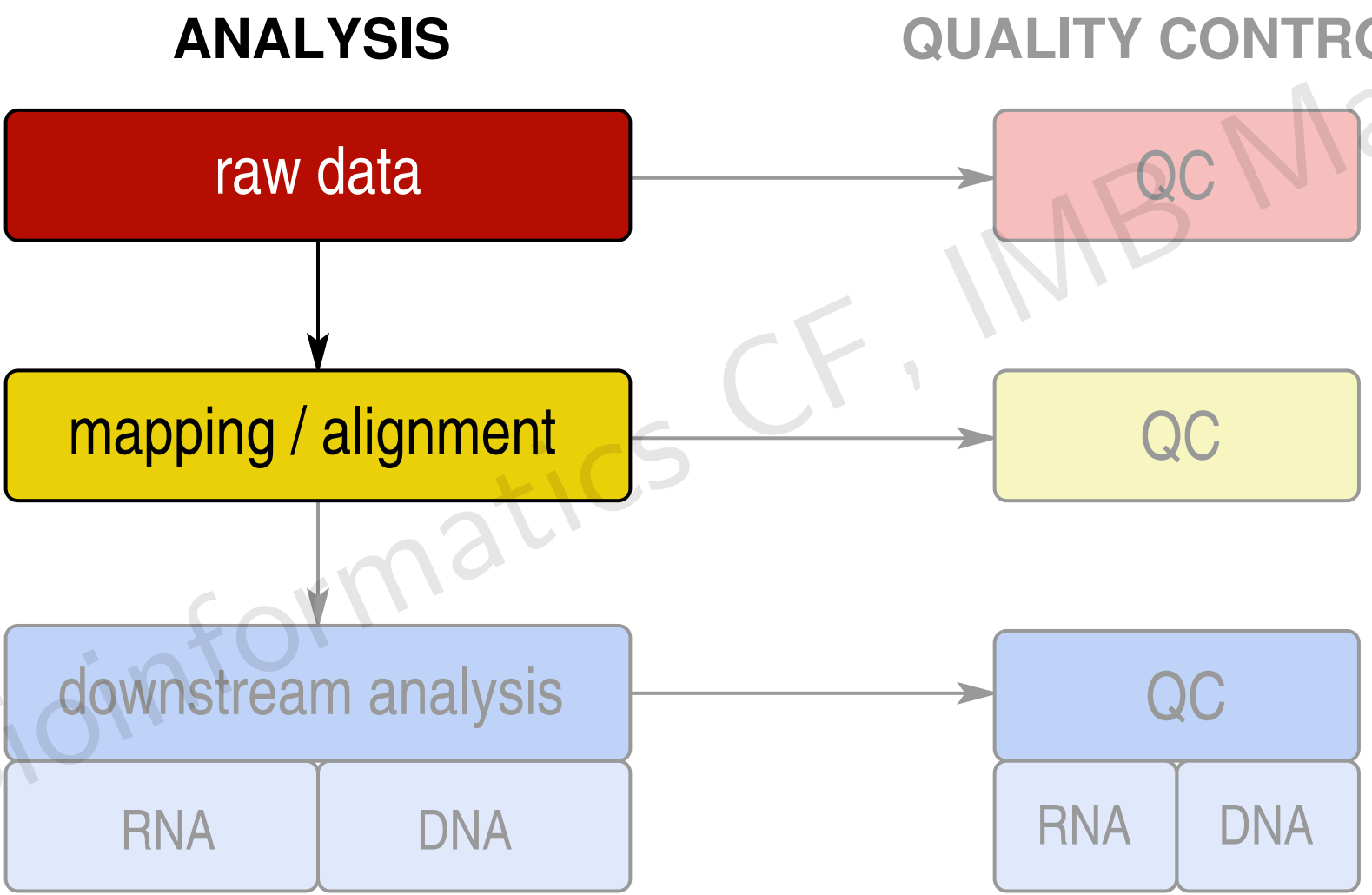


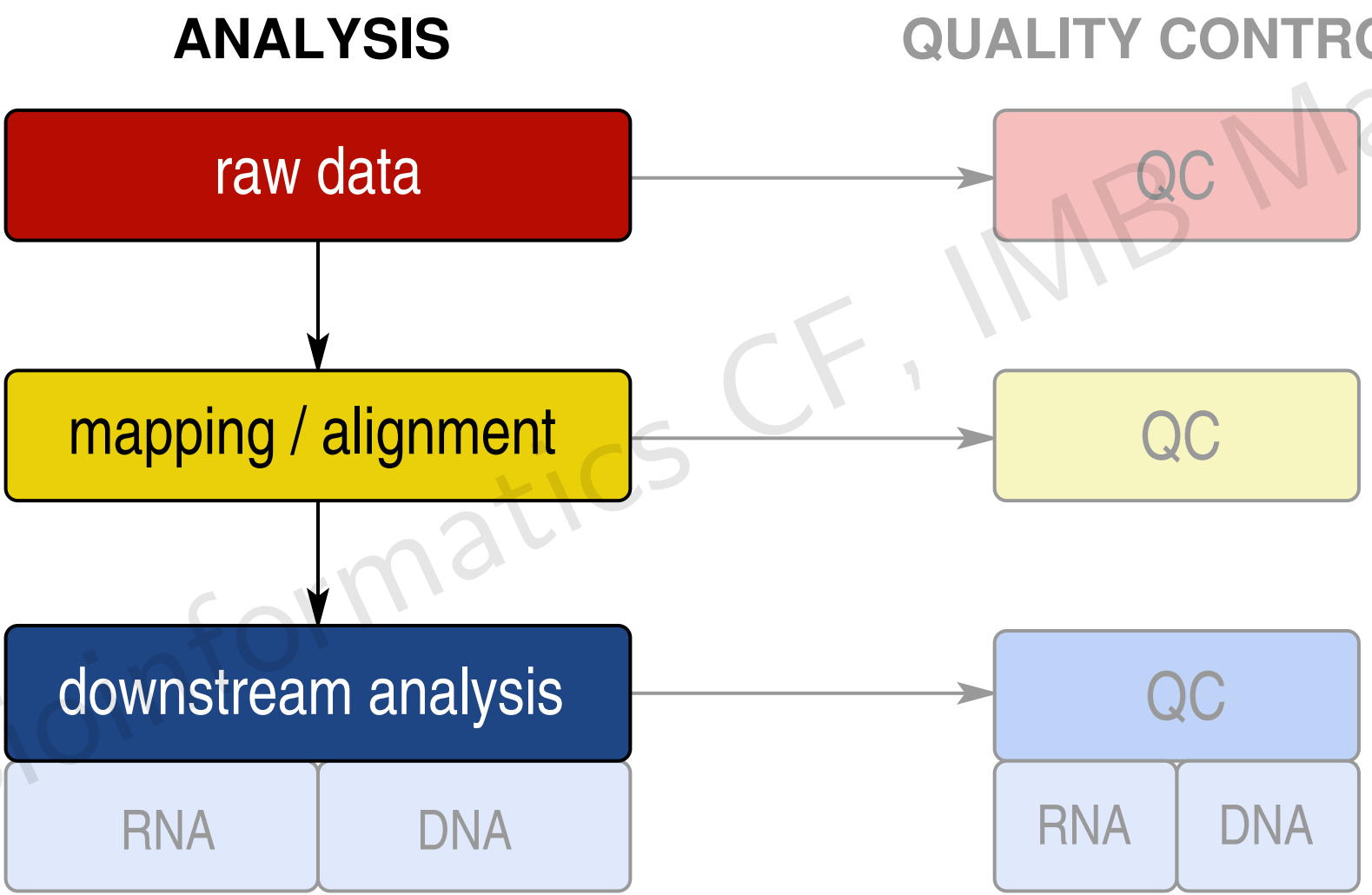
Multi-mapped reads (non-uniquely mapped)



Unmapped reads

Visualization: browser tracks



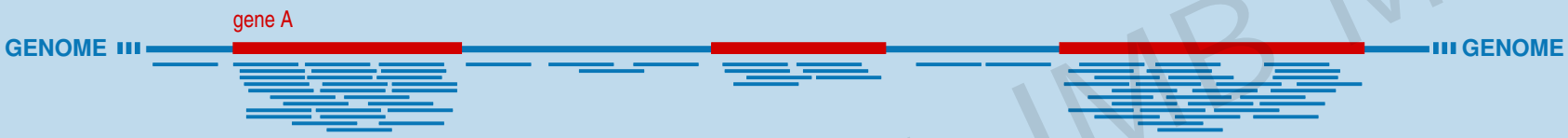


Quantification of reads

Quantify/count reads on **RNA**

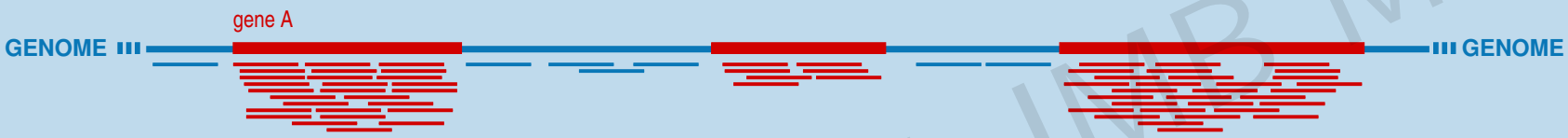
Quantification of reads

Quantify/count reads on **RNA**



Quantification of reads

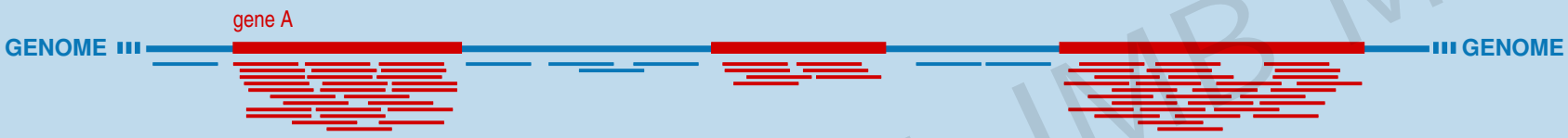
Quantify/count reads on **RNA**



- count reads in annotated regions

Quantification of reads

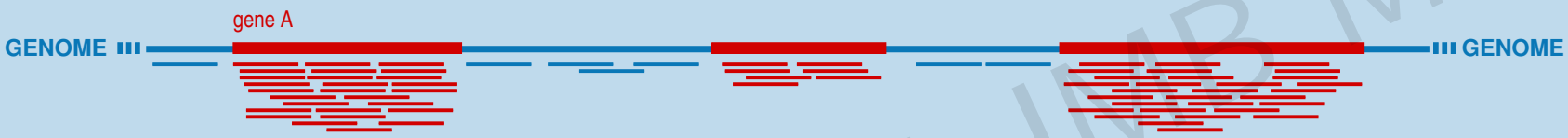
Quantify/count reads on **RNA**



- count reads in annotated regions
- #reads on gene A: 64

Quantification of reads

Quantify/count reads on **RNA**

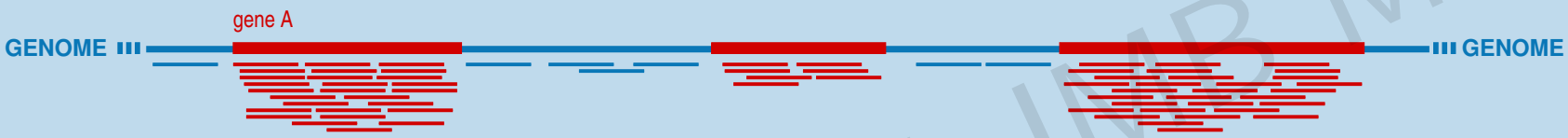


- count reads in annotated regions
- #reads on gene A: 64

Quantify/count reads on **DNA**

Quantification of reads

Quantify/count reads on **RNA**

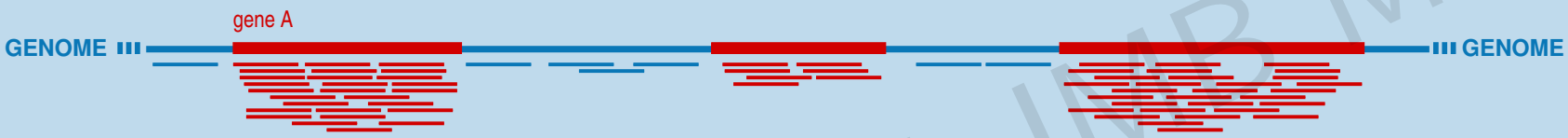


- count reads in annotated regions
- #reads on gene A: 64

Quantify/count reads on **DNA**

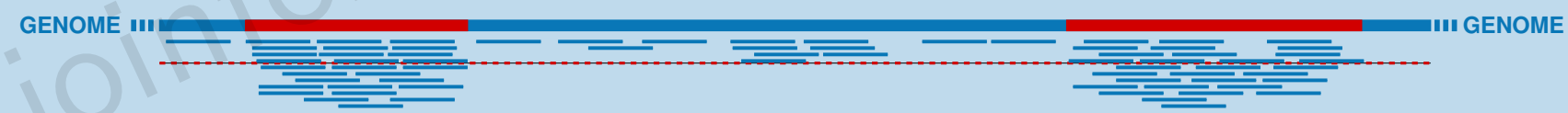
Quantification of reads

Quantify/count reads on **RNA**



- count reads in annotated regions
- #reads on gene A: 64

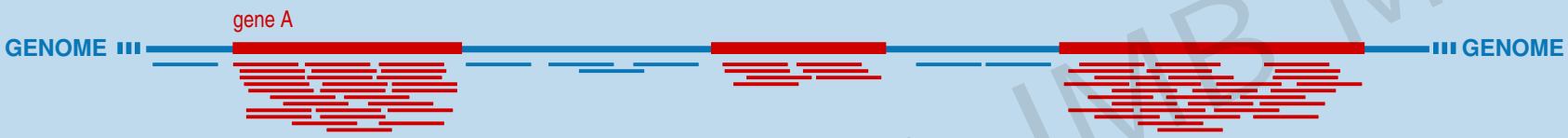
Quantify/count reads on **DNA**



- search for enriched areas: peak calling

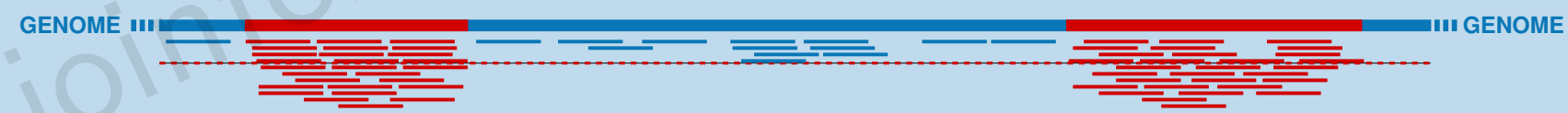
Quantification of reads

Quantify/count reads on **RNA**



- count reads in annotated regions
- #reads on gene A: 64

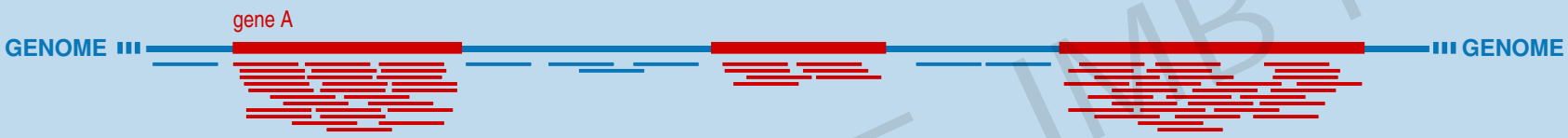
Quantify/count reads on **DNA**



- search for enriched areas: peak calling
- count reads in enriched areas

Quantification of reads

Quantify/count reads on **RNA**



- count reads in annotated regions
- #reads on gene A: 64

Quantify/count reads on **DNA**

- search for enriched areas: peak calling
- count reads in enriched areas

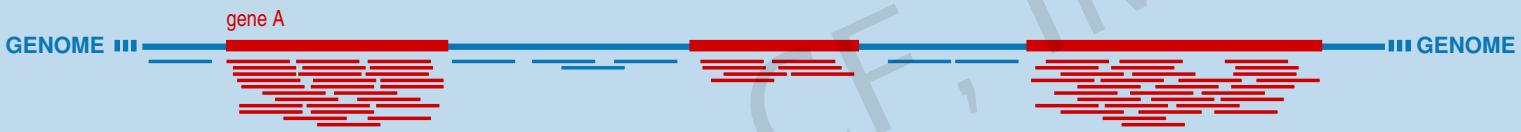
RNAseq: differential expression analysis

Steps

1. RNA quantification
2. normalization
3. data modeling, distribution estimation
4. visualization

RNA quantification

Counting reads



- only count reads on exons
- combine different isoforms
- result: **read count per gene**

```
gene A: 64
gene B: 0
gene C: 135
gene D: 21
gene E: 209
...
```

Normalization - why?

Why normalization? What to normalize for?

the number of counts is related to:

- mRNA expression level (proportional)

but also:

- the sequencing depth
- the transcript length

Thus, we need **within** and **between** sample normalization.

Normalization methods

RPKM

(Mortazavi et al., Nat. Methods, 2008)

- **R**eads **P**er **K**ilobase per **M**illion mapped reads
- read counts are divided by the transcript length (kb) and the total number of millions of mapped reads

$$\text{RPKM} = \frac{\text{read count}}{\frac{\text{gene length}}{10^3} * \frac{\text{total read count}}{10^6}}$$

Normalization methods

RPKM

(Mortazavi et al., Nat. Methods, 2008)

- **R**eads **P**er **K**ilobase per **M**illion mapped reads
- read counts are divided by the transcript length (kb) and the total number of millions of mapped reads

$$\text{RPKM} = \frac{\text{read count}}{\frac{\text{gene length}}{10^3} * \frac{\text{total read count}}{10^6}}$$

- problem: how to determine correct gene length in case of several isoforms? only some might be expressed
  $\rightarrow$ not useful for diff. expression analysis

Normalization methods

RPKM

(Mortazavi et al., Nat. Methods, 2008)

- **R**eads **P**er **K**ilobase per **M**illion mapped reads
- read counts are divided by the transcript length (kb) and the total number of millions of mapped reads

$$\text{RPKM} = \frac{\text{read count}}{\frac{\text{gene length}}{10^3} * \frac{\text{total read count}}{10^6}}$$

- problem: how to determine correct gene length in case of several isoforms? only some might be expressed
  $\rightarrow$ not useful for diff. expression analysis

RPM

- **R**eads **P**er **M**illion mapped reads
- read counts are divided by the total number of millions of mapped reads

$$\text{RPM} = \frac{\text{read count}}{\frac{\text{total read count}}{10^6}}$$

Normalization methods

RPKM

(Mortazavi et al., Nat. Methods, 2008)

- **R**eads **P**er **K**ilobase per **M**illion mapped reads
- read counts are divided by the transcript length (kb) and the total number of millions of mapped reads

$$\text{RPKM} = \frac{\text{read count}}{\frac{\text{gene length}}{10^3} * \frac{\text{total read count}}{10^6}}$$

- problem: how to determine correct gene length in case of several isoforms? only some might be expressed
  $\rightarrow$ not useful for diff. expression analysis

RPM

- **R**eads **P**er **M**illion mapped reads
- read counts are divided by the total number of millions of mapped reads

$$\text{RPM} = \frac{\text{read count}}{\frac{\text{total read count}}{10^6}}$$

- only suitable for between sample comparison (no within sample comp.)

Data modeling

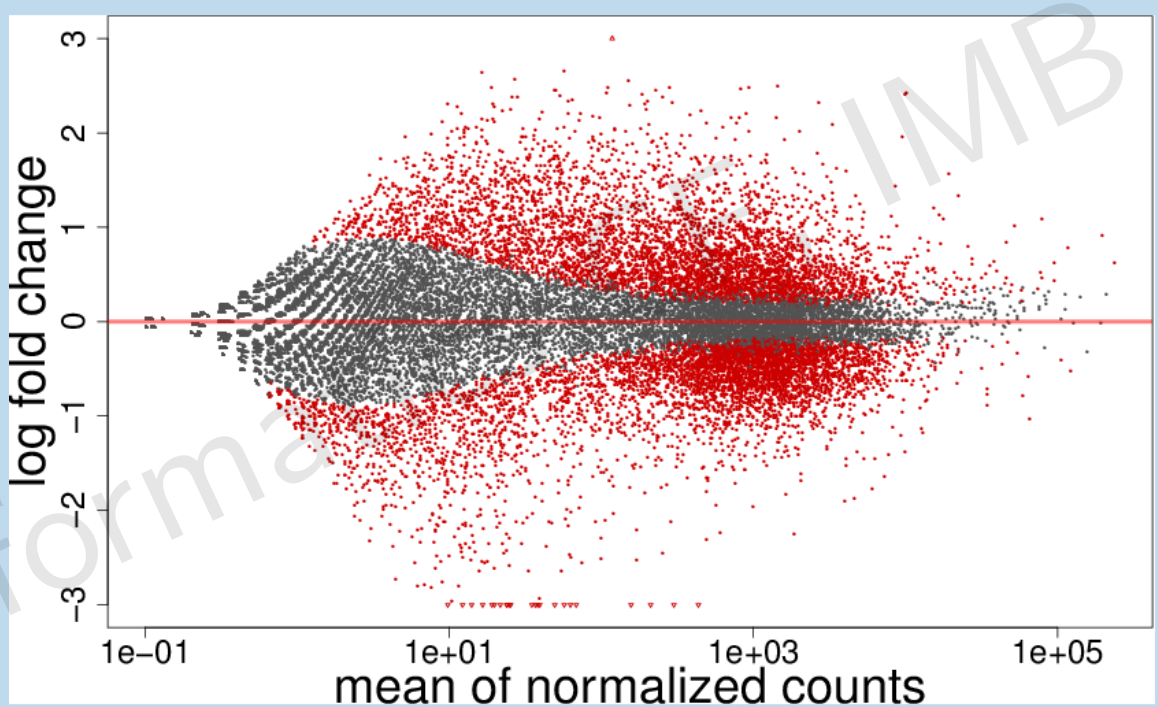
Data modeling / distribution estimation

- read counts modeled as following a negative binomial distribution (=poisson-like distribution with variance not only depending on mean)
- for each gene: calculate **p-value for differential expression**:
  - model read counts within conditions
  - assumption: genes with similar expression have similar variance
  - compare variation within conditions to that between conditions
- correct p-value for multiple testing (FDR)

Tools and plots

Example results from DESeq2 package

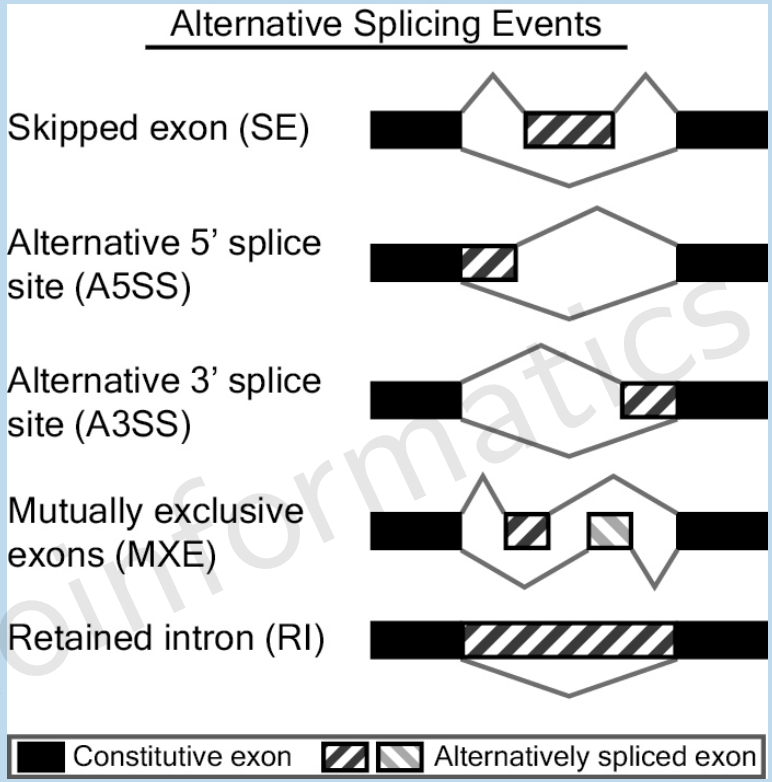
(Love et al., Genome Biology, 2014)



- variance within/between groups especially high for low read counts
- DESeq2: moderation of fold changes

Other downstream analyses

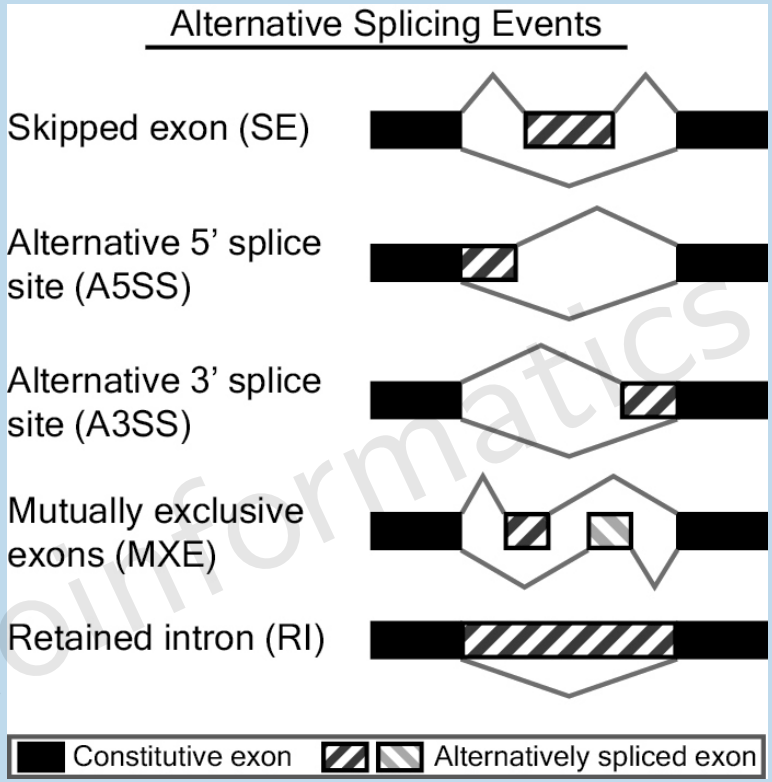
Differential alternative splicing analysis



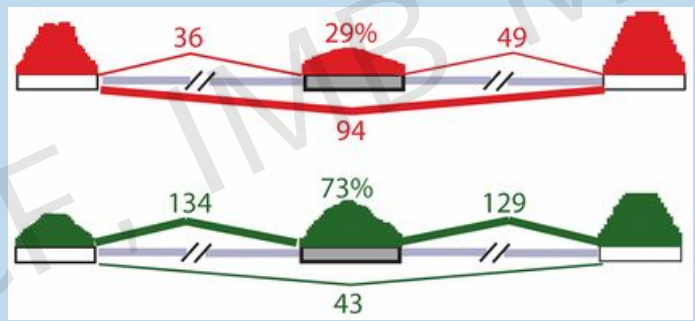
http://rnaseq-mats.sourceforge.net/

Other downstream analyses

Differential alternative splicing analysis



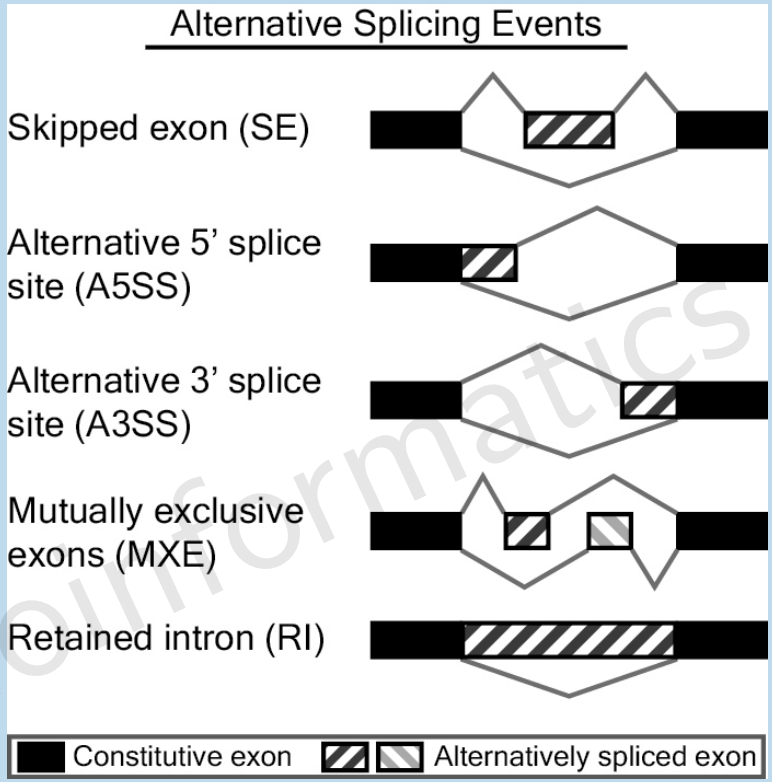
http://rnaseq-mats.sourceforge.net/



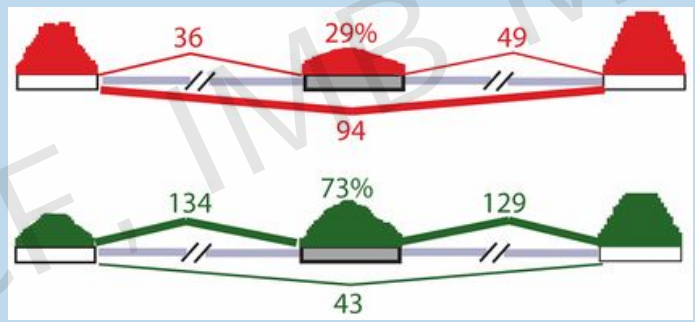
(Shen et al., PNAS, 2014.)

Other downstream analyses

Differential alternative splicing analysis



http://rnaseq-mats.sourceforge.net/



(Shen et al., PNAS, 2014.)

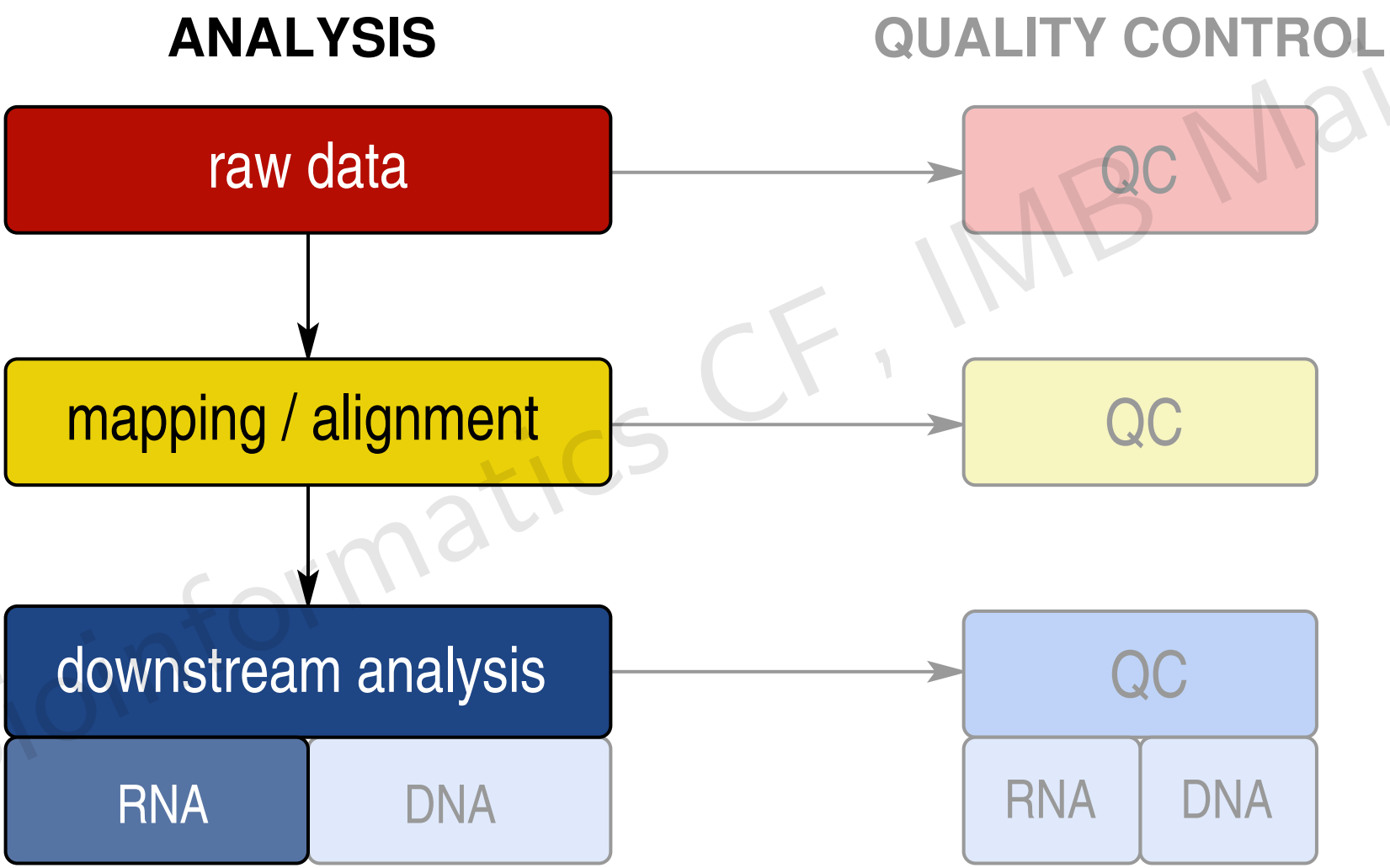
Tools (e.g.):

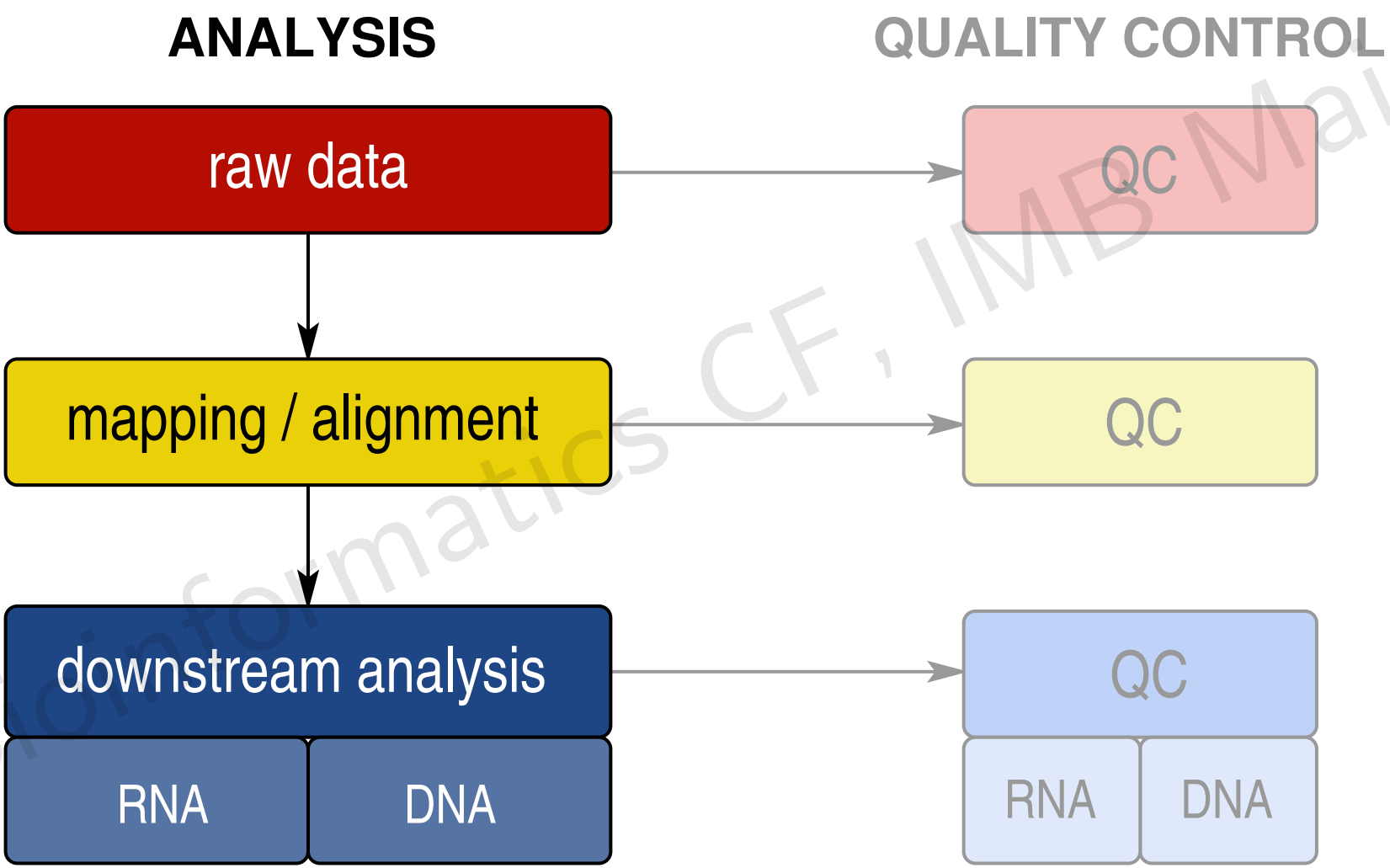
- `rMATS`
  (Shen et al., PNAS, 2014.)
- `MAJIQ`
  (Vaquero-Garcia et al., eLife, 2016.)

Other downstream analyses

GO term analysis

- assign GO terms to each significantly differentially expressed gene
- look for enriched GO terms in sign. up- or down-regulated genes
- tools (e.g.):
  - `clusterProfiler` (Yu et al., OMICS, 2012.)
  - `DAVID` (Huang et al., Nature Protocols, 2009.)



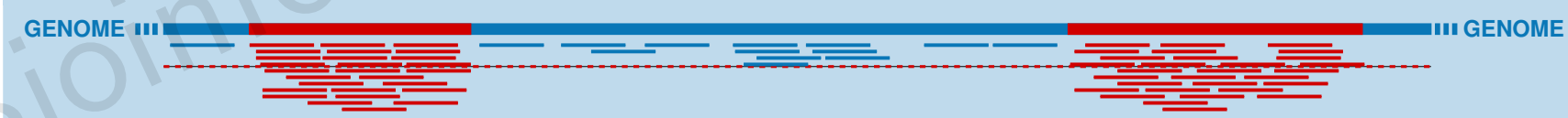


Quantification of reads

Quantify/count reads on **RNA**

- count reads in annotated regions
- #reads on gene A: 64

Quantify/count reads on **DNA**

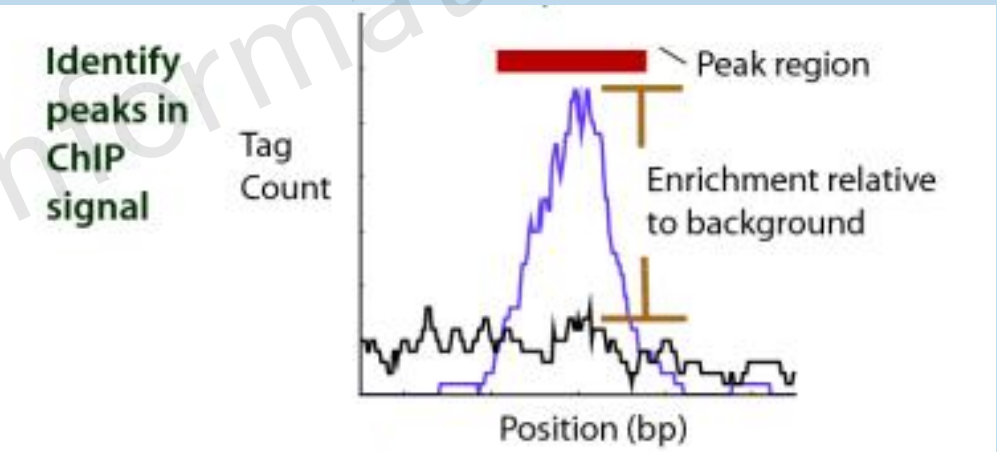


- search for enriched areas: peak calling
- count reads in enriched areas

Peak calling: general idea

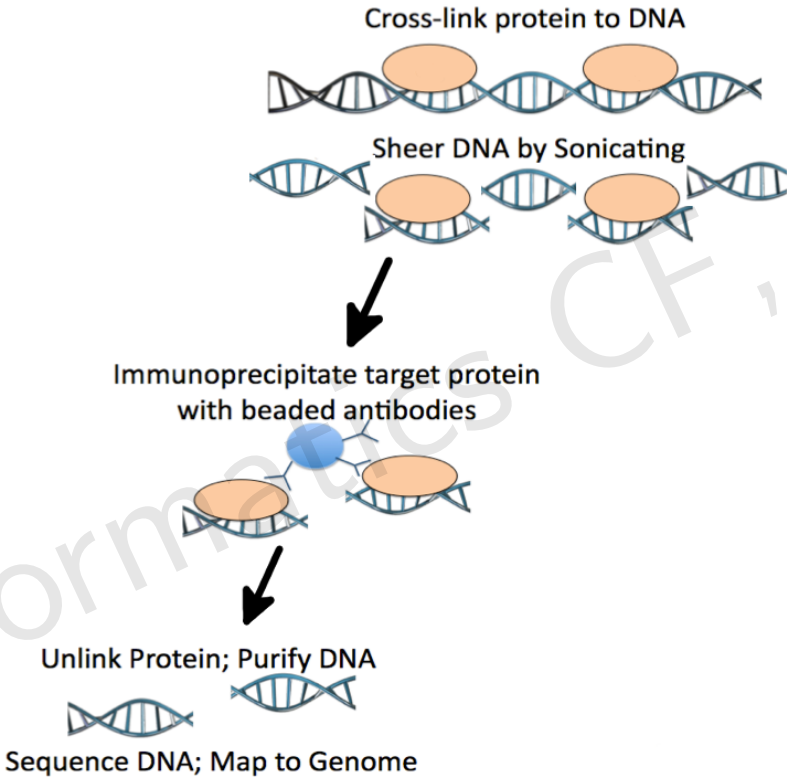
Peak calling

- identification of **regions (peaks)** that are **enriched in the ChIP sample** relative to the **control** with statistical significance
- most common approach: **sliding window**, count reads per window
- **peak** = enriched relative to control:



Pepke et al., Nat. Methods, 2009

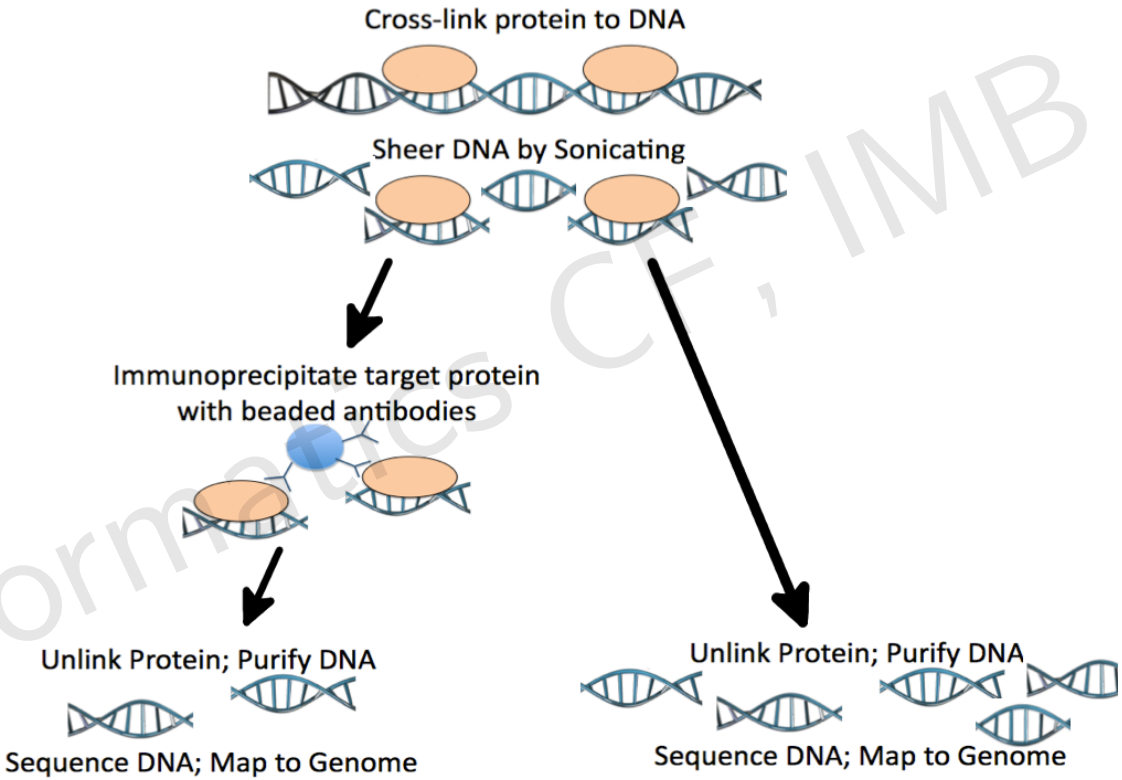
ChIP vs. input control



ChIP

adapted from
http://saltmanquarterly.wordpress.com/2013/06/16/epigenetics-changing-how-we-interpret-genomes/

ChIP vs. input control



adapted from
http://saltmanquarterly.wordpress.com/2013/06/16/epigenetics-changing-how-we-interpret-genomes/

Recommended controls

Input (most popular)

cross-linked and sonicated (fragmented) DNA, but not IP'd

Recommended controls

Input (most popular)

cross-linked and sonicated (fragmented) DNA, but not IP'd

IgG / mock IP

Immunoglobulin G (IgG) used as a control (unspecific) antibody, which does not recognize DNA or chromatin associated proteins

Recommended controls

Input (most popular)

cross-linked and sonicated (fragmented) DNA, but not IP'd

IgG / mock IP

Immunoglobulin G (IgG) used as a control (unspecific) antibody, which does not recognize DNA or chromatin associated proteins

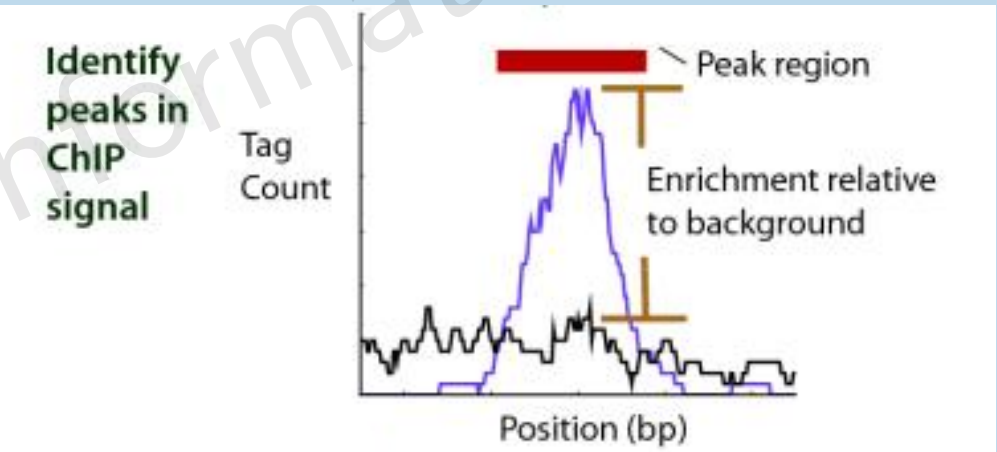
untagged epitope

in case of epitope tagged constructs, perform ChIP on cells lacking epitope tag

Peak calling: general idea

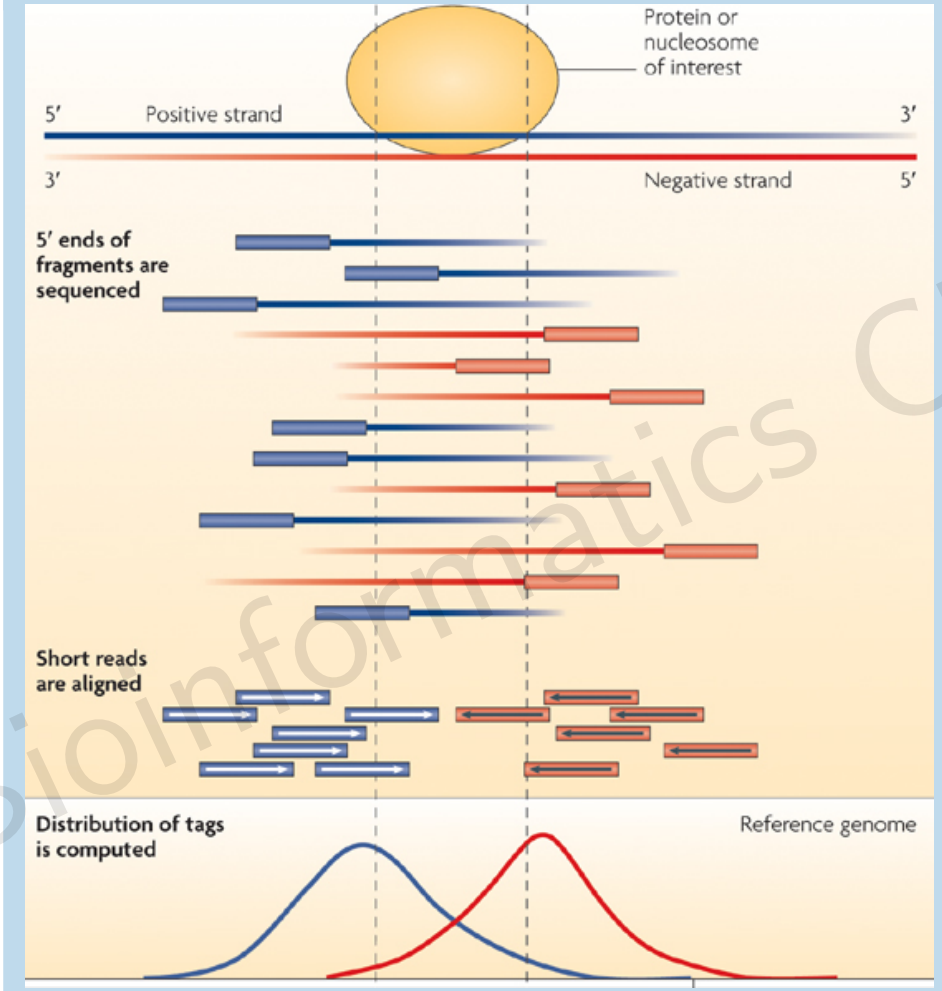
Peak calling

- identification of **regions (peaks)** that are **enriched in the ChIP sample** relative to the **control** with statistical significance
- most common approach: **sliding window**, count reads per window
- **peak** = enriched relative to control:



Pepke et al., Nat. Methods, 2009

Peak calling: Strand specific profiles at enriched sites



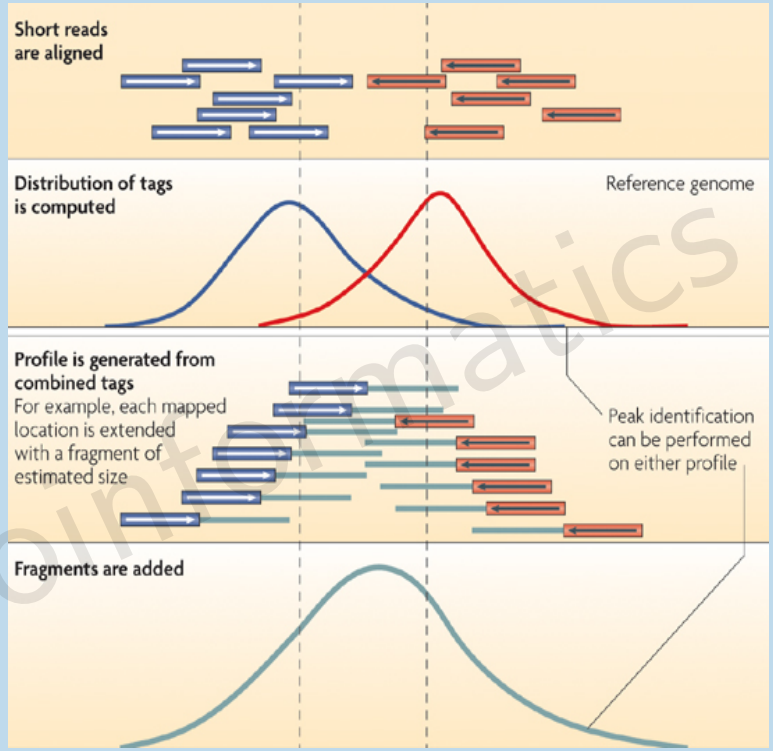
Park, Nat. Rev. Genetics, 2009

- DNA sequences are sequenced from the 5' end
- alignment to genome results in two peaks (one on each strand)
- peaks are flanking the binding location of the protein of interest

Peak calling: construction of combined signal profiles

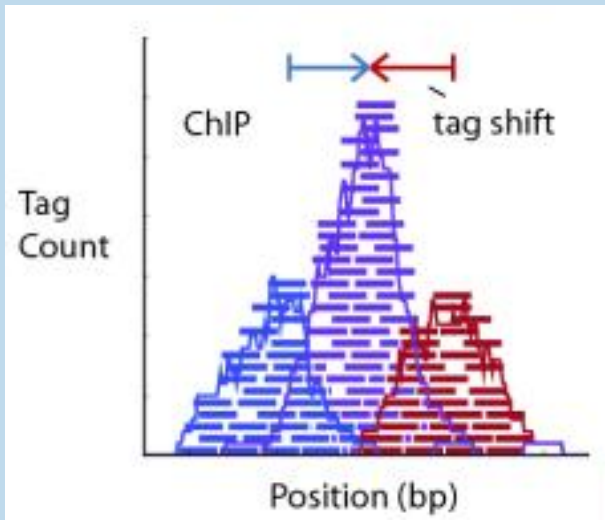
Estimation of local density: for both strands and individually

Park, Nat. Rev. Genetics, 2009



extending and combining fragments

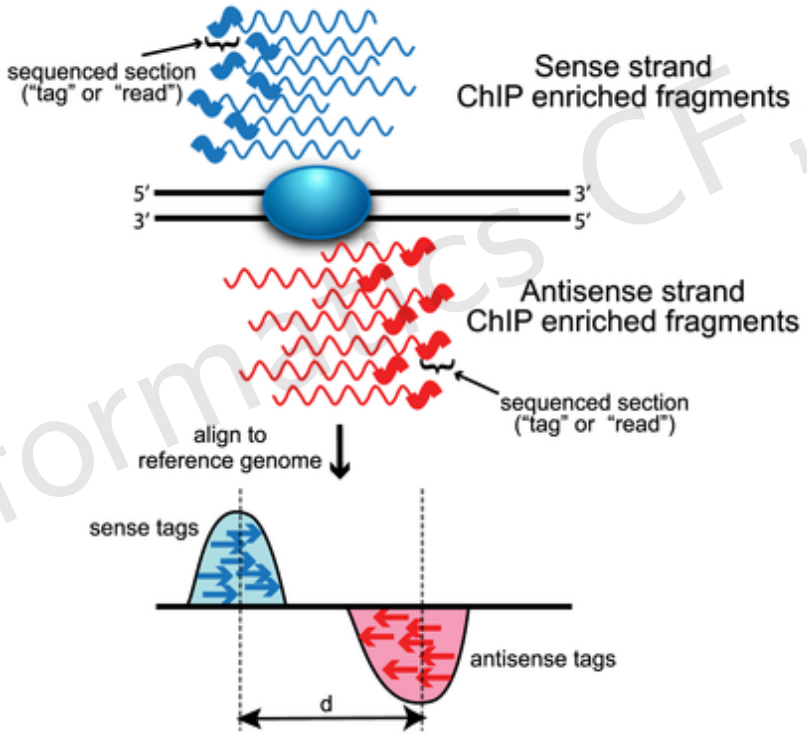
Pepke et al., Nat. Methods, 2009



shifting reads towards center

Peak calling: enrichment for TFs and histone modifications

sequence specific binding
(e.g. transcription factors)

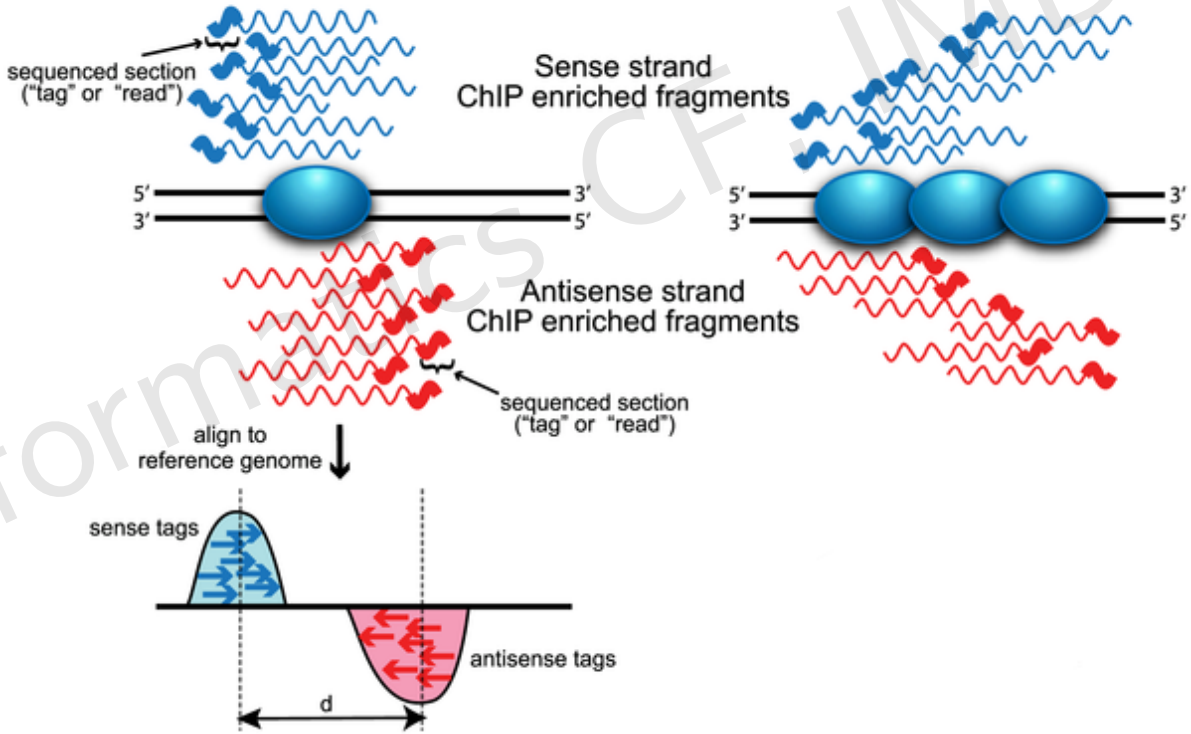


Wilbanks & Facciotti, PLOS ONE, 2010

Peak calling: enrichment for TFs and histone modifications

sequence specific binding (e.g. transcription factors)

distributed binding events (e.g. histones)

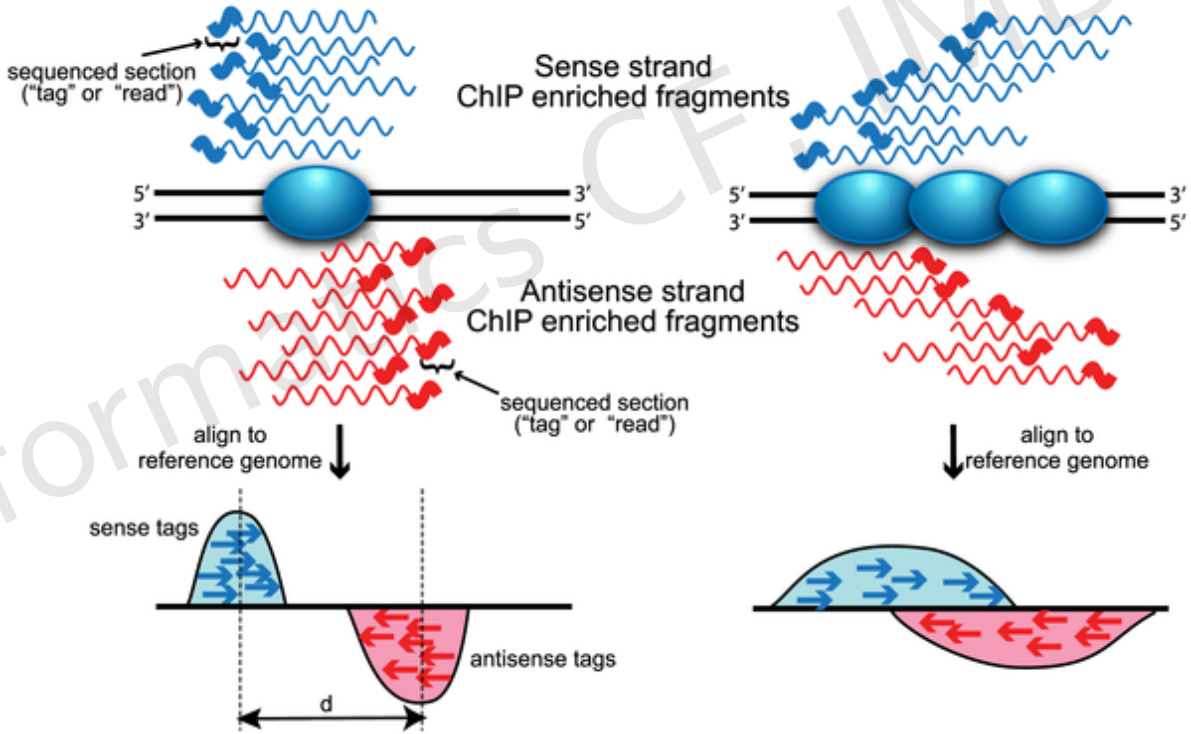


Wilbanks & Facciotti, PLOS ONE, 2010

Peak calling: enrichment for TFs and histone modifications

sequence specific binding (e.g. transcription factors)

distributed binding events (e.g. histones)



Wilbanks & Facciotti, PLOS ONE, 2010

Peak calling: tool comparison

Which peak finder should I use?

- dozens of different peak finders published
- some optimized for either TFs or histone marks
- sensitive to parameter settings
- e.g. `MACS2` (https://github.com/macs3-project/MACS)

Reviews

TF ChIP-seq:

- Laajala et al., BMC Genomics, 2009
- Wilbanks & Facciotti, PLOS ONE, 2010

histone ChIP-seq:

- Micsinai et al., NAR, 2012

General:

- Pepke et al., Nat. Methods, 2009

After peak calling: annotation / functional analysis

PeakAnalyzer (PeakAnnotator)

(Salmon-Divon et al., BMC Bioinformatics, 2010)

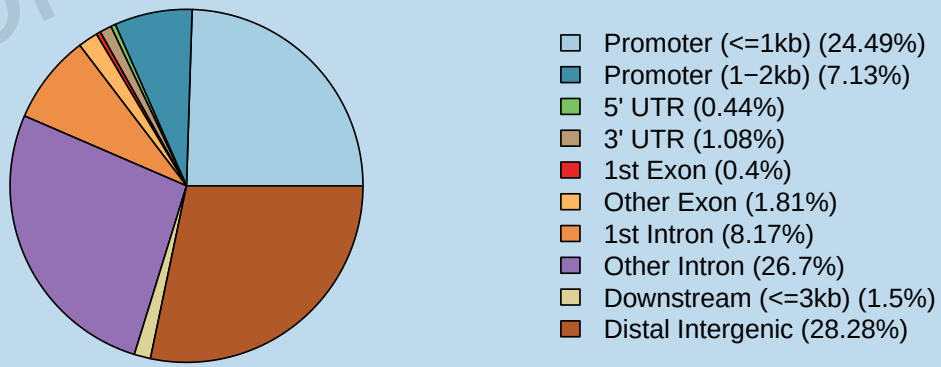
For each peak:

- downstream forward gene + distance
- downstream reverse gene + distance
- overlapped genes + overlap start (feat.) + overlap center (feat.) + overlap end (feat.)

ChIPseeker

(https://github.com/GuangchuangYu/ChIPseeker)

Peaks overlapping gene features



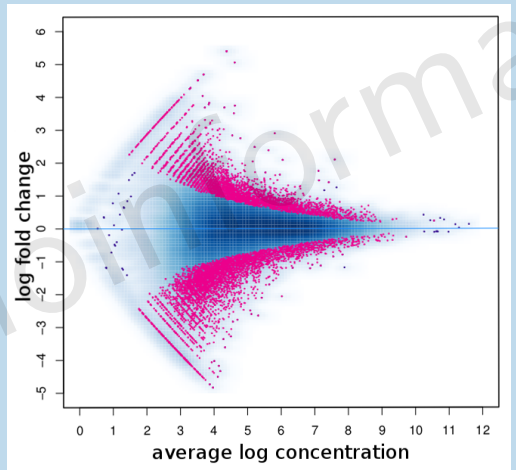
After peak calling: differential binding sites

DiffBind

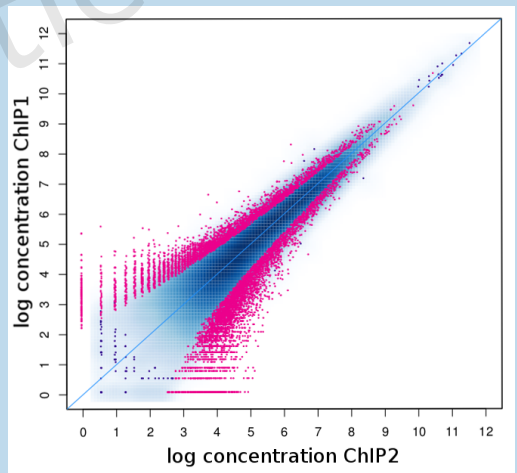
(Ross-Innes et al., Nature, 2012)

- comparison of two or more conditions
- runs edgeR, DESeq, or DESeq2 internally

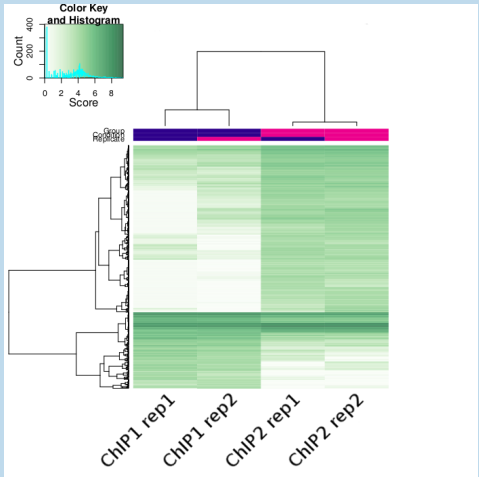
MA plot



binding affinity
ChIP1 vs. ChIP2



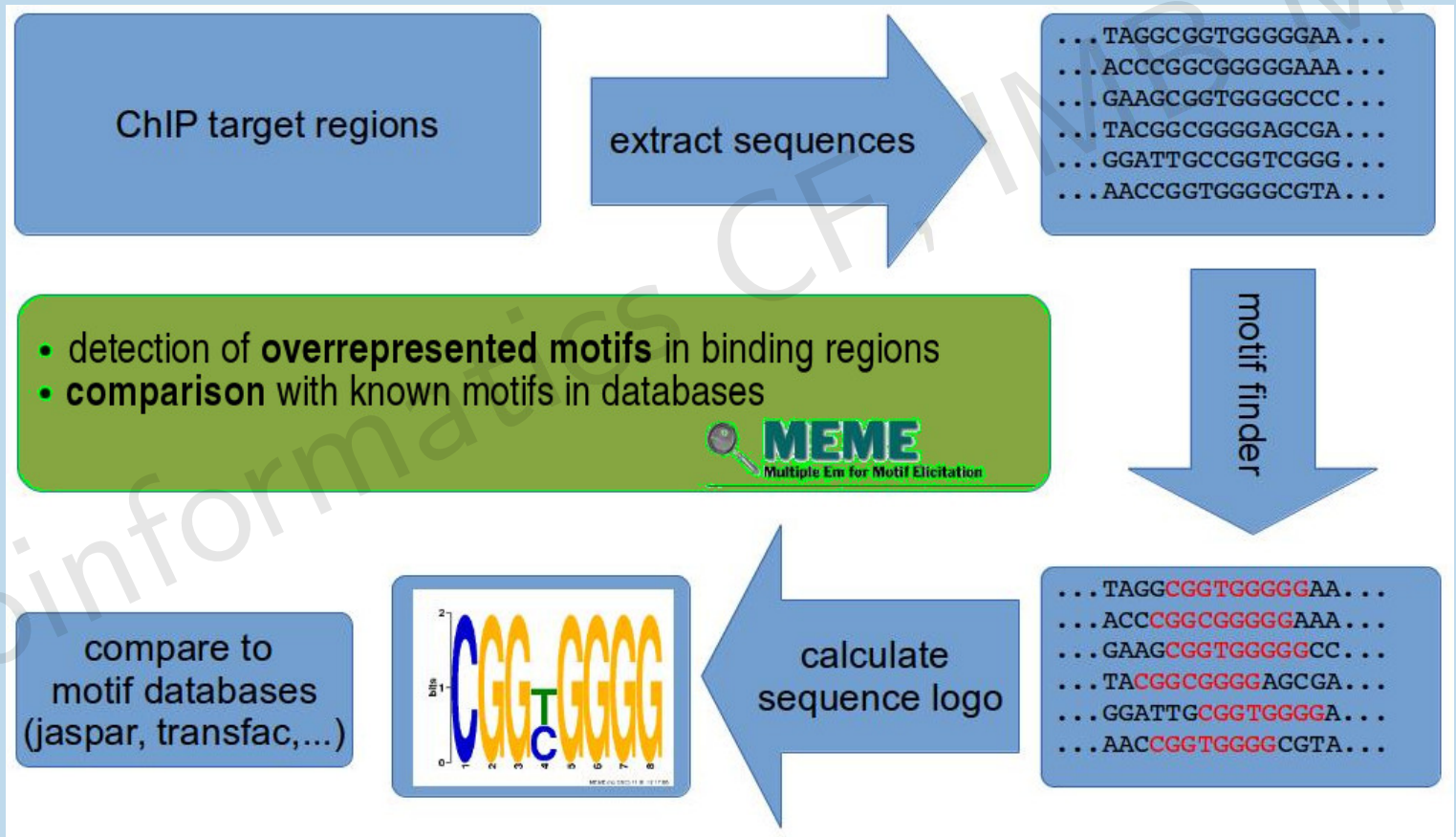
affinity heatmap



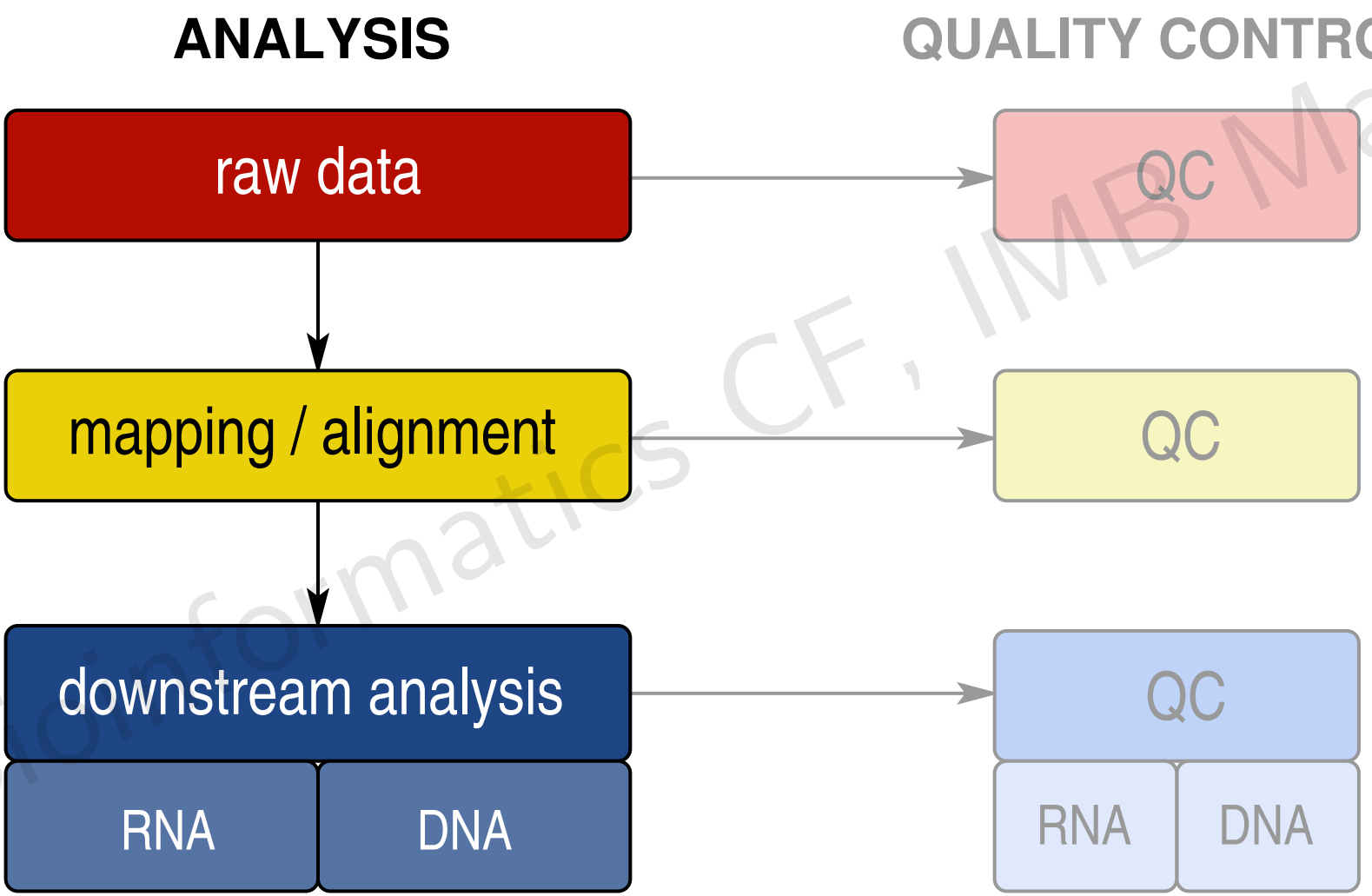
After peak calling: detection of (novel) binding motifs

MEME (Multiple EM for Motif Elicitation)

(Bailey & Elkan, Proc. of the ISMB, 1994)

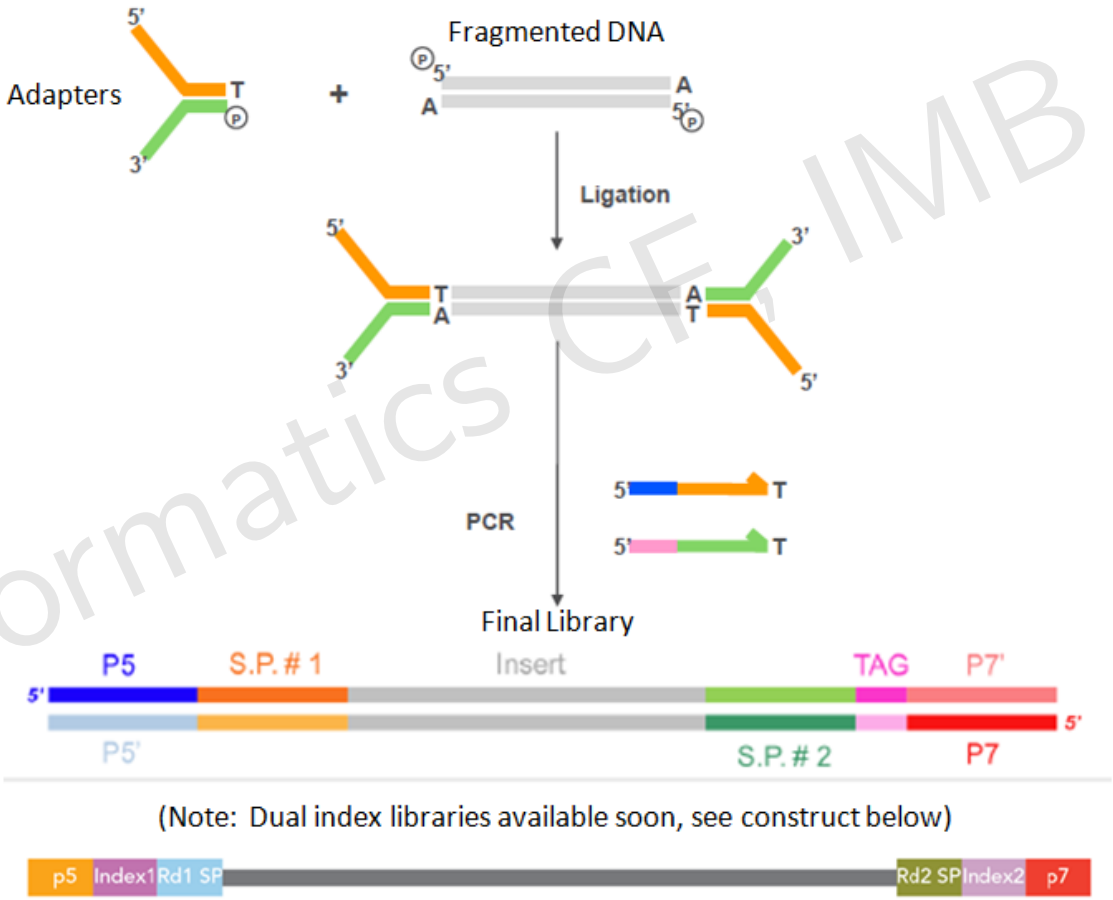


(scheme by Holger Klein)

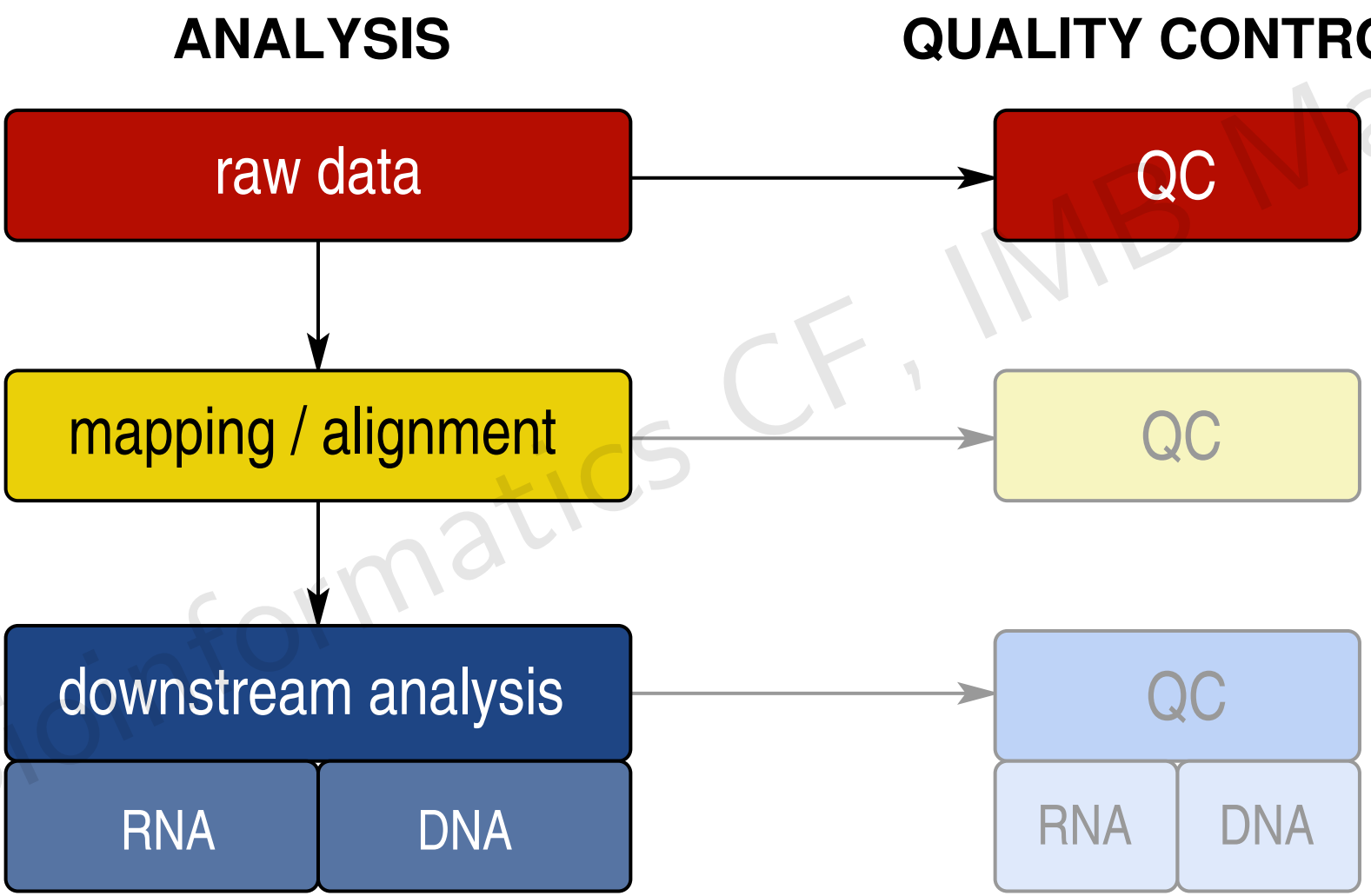


Quality control of next generation sequencing data

Library Preparation



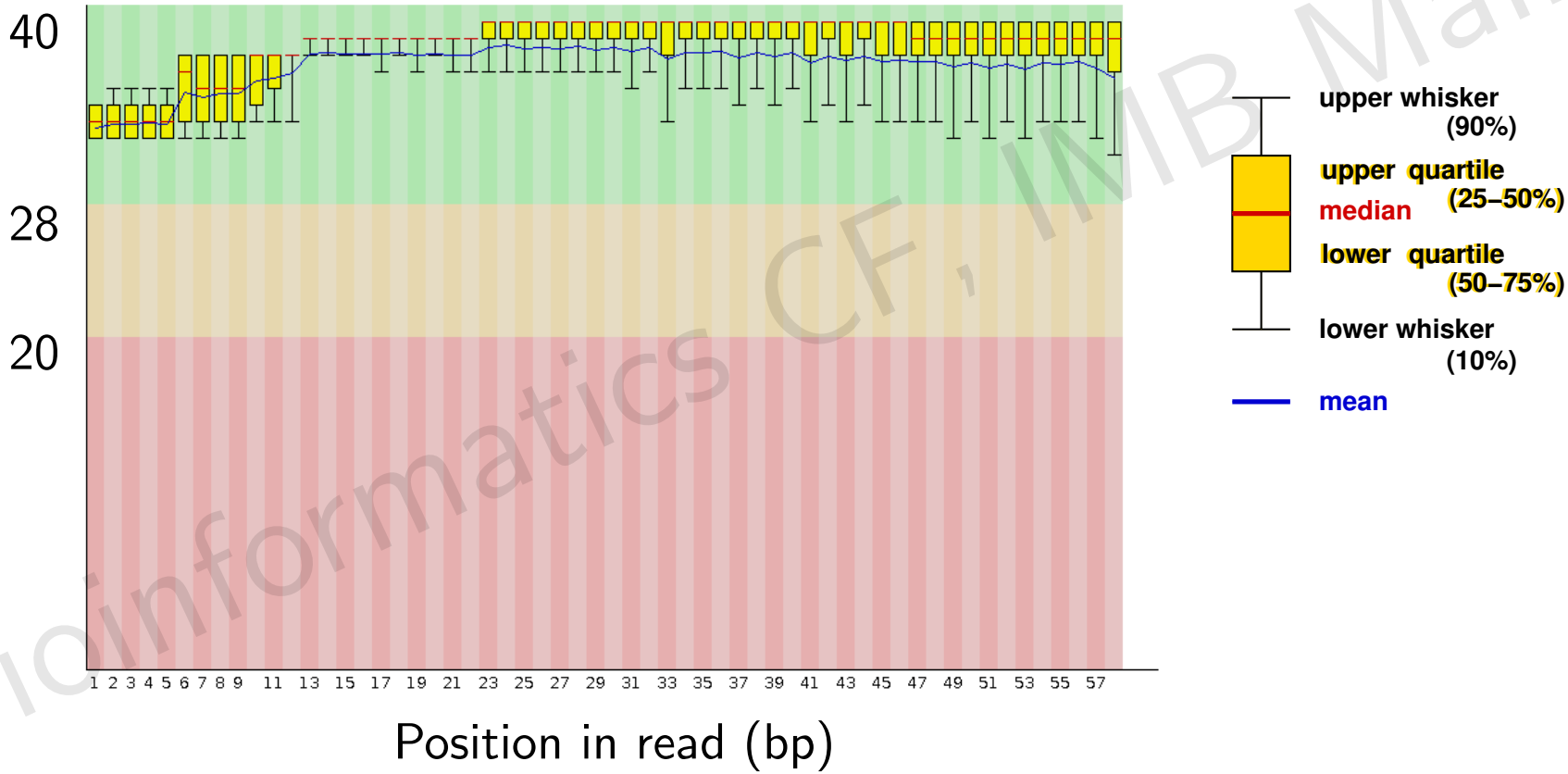
Illumina



Raw data quality control

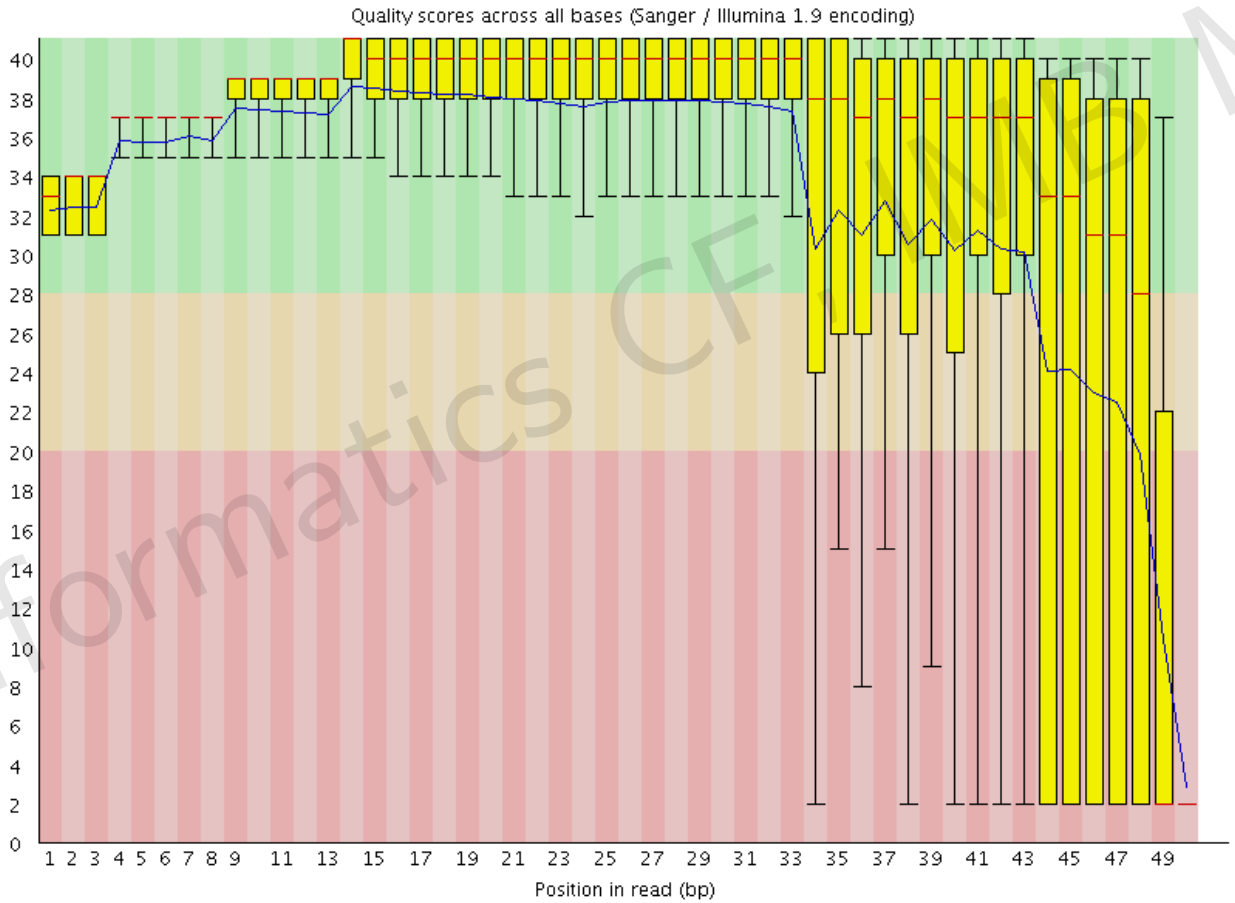
- quality score distribution
- base composition distribution
- read length distribution
- distribution of reads over samples
- overrepresented sequences

Distribution of quality values along reads



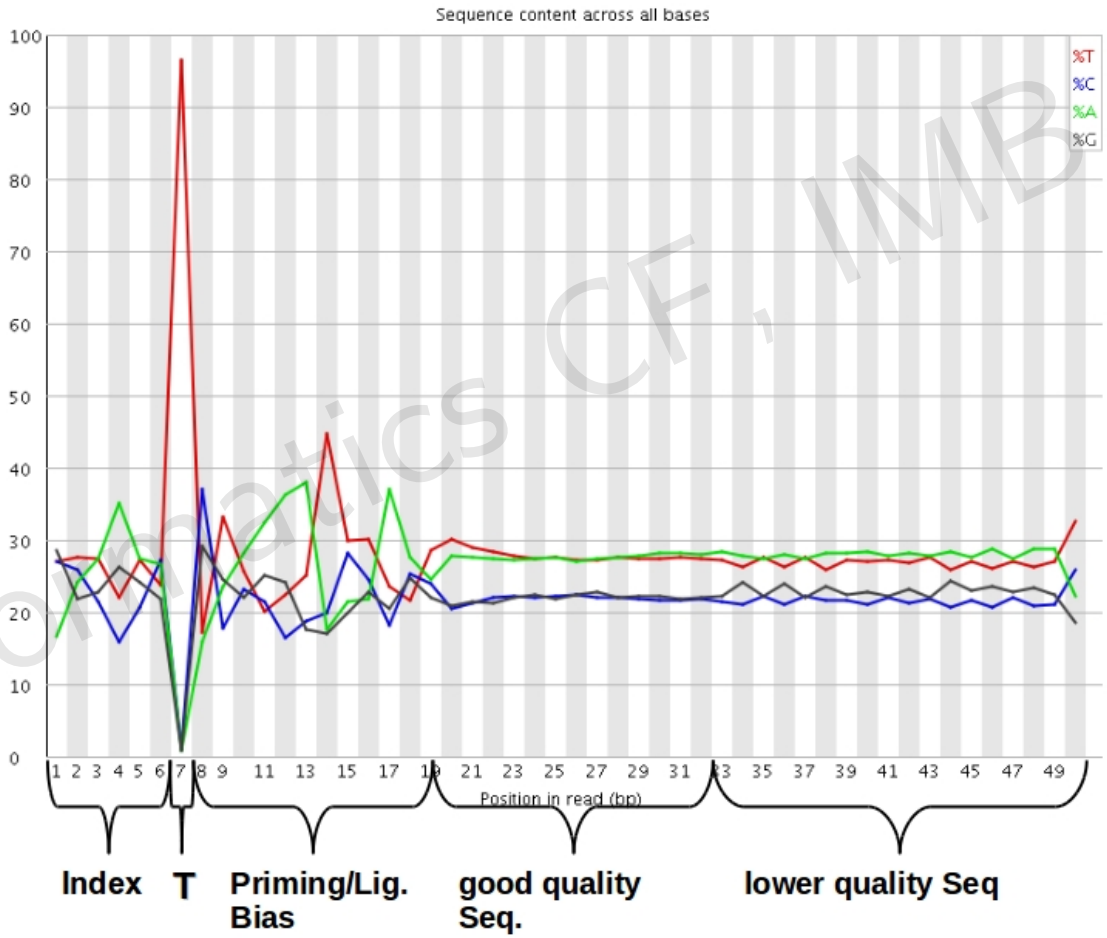
plot produced by fastqc: Andrews, S. "FASTQC. A quality control tool for high throughput sequence data." URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2010).

Quality score distribution



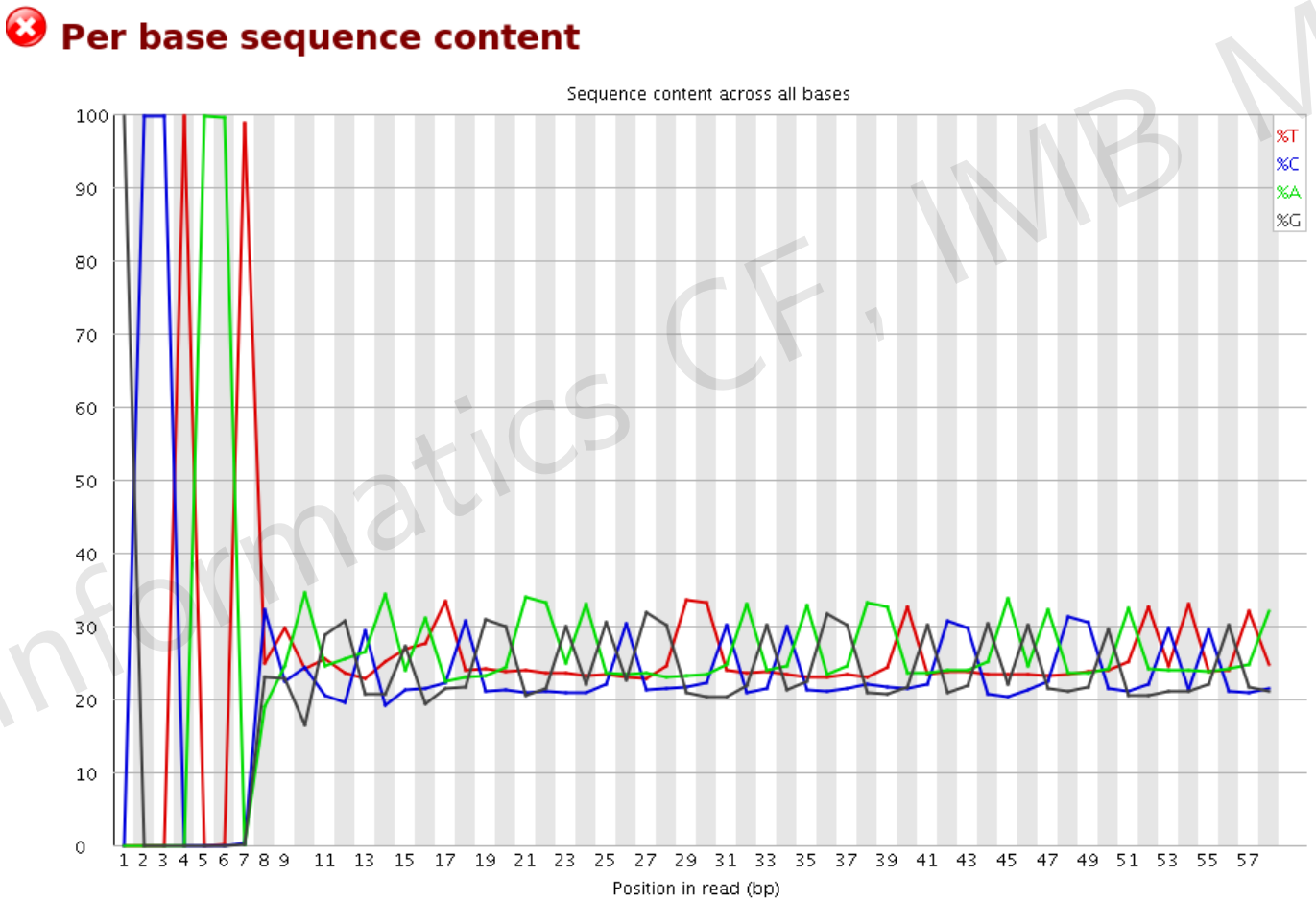
plot produced by fastqc: Andrews, S. "FASTQC. A quality control tool for high throughput sequence data." URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2010).

Per base sequence content



plot produced by fastqc: Andrews, S. "FASTQC. A quality control tool for high throughput sequence data." URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2010).

Per base sequence content



plot produced by fastqc: Andrews, S. "FASTQC. A quality control tool for high throughput sequence data." URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2010).

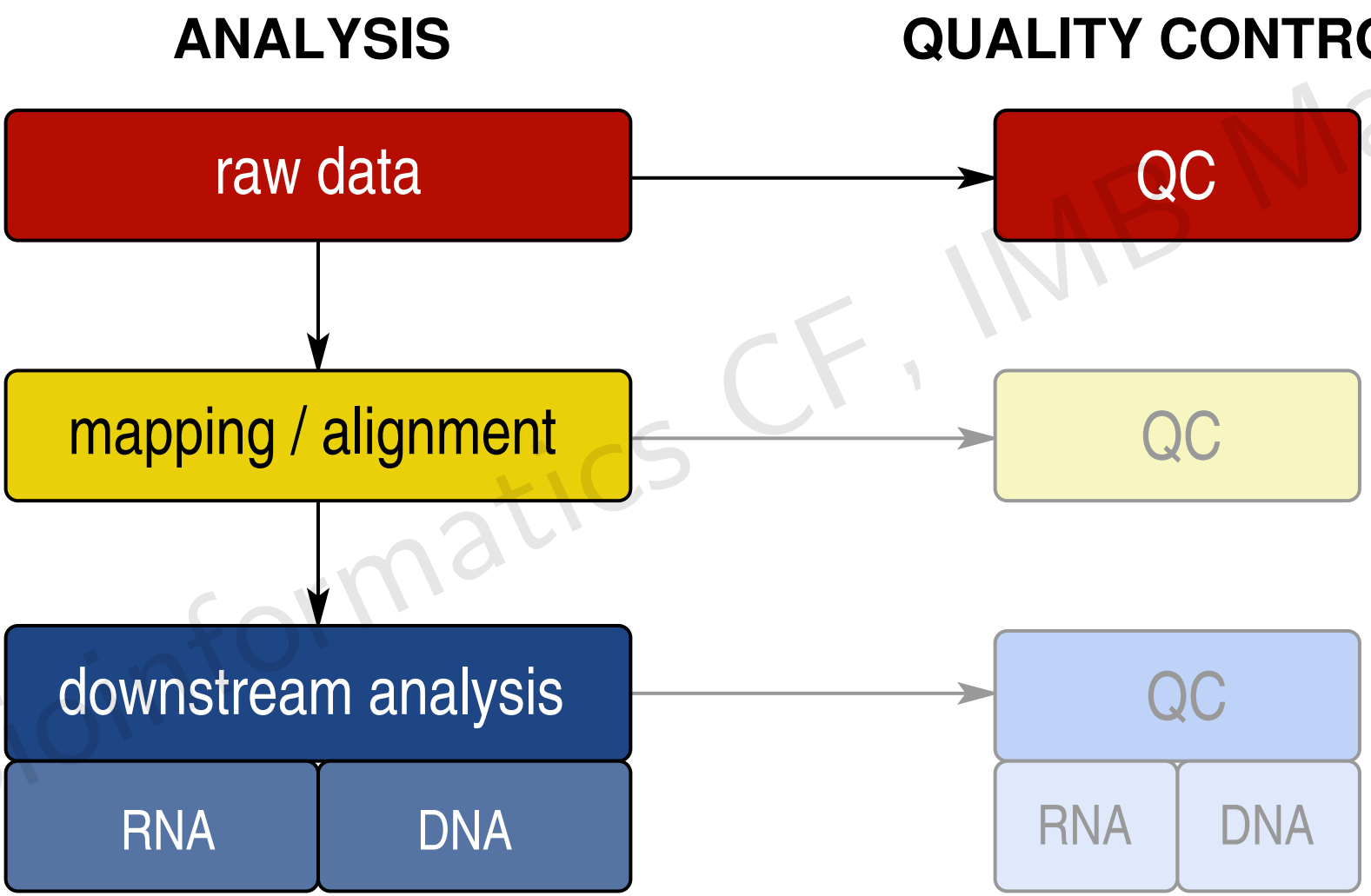
Overrepresented sequences

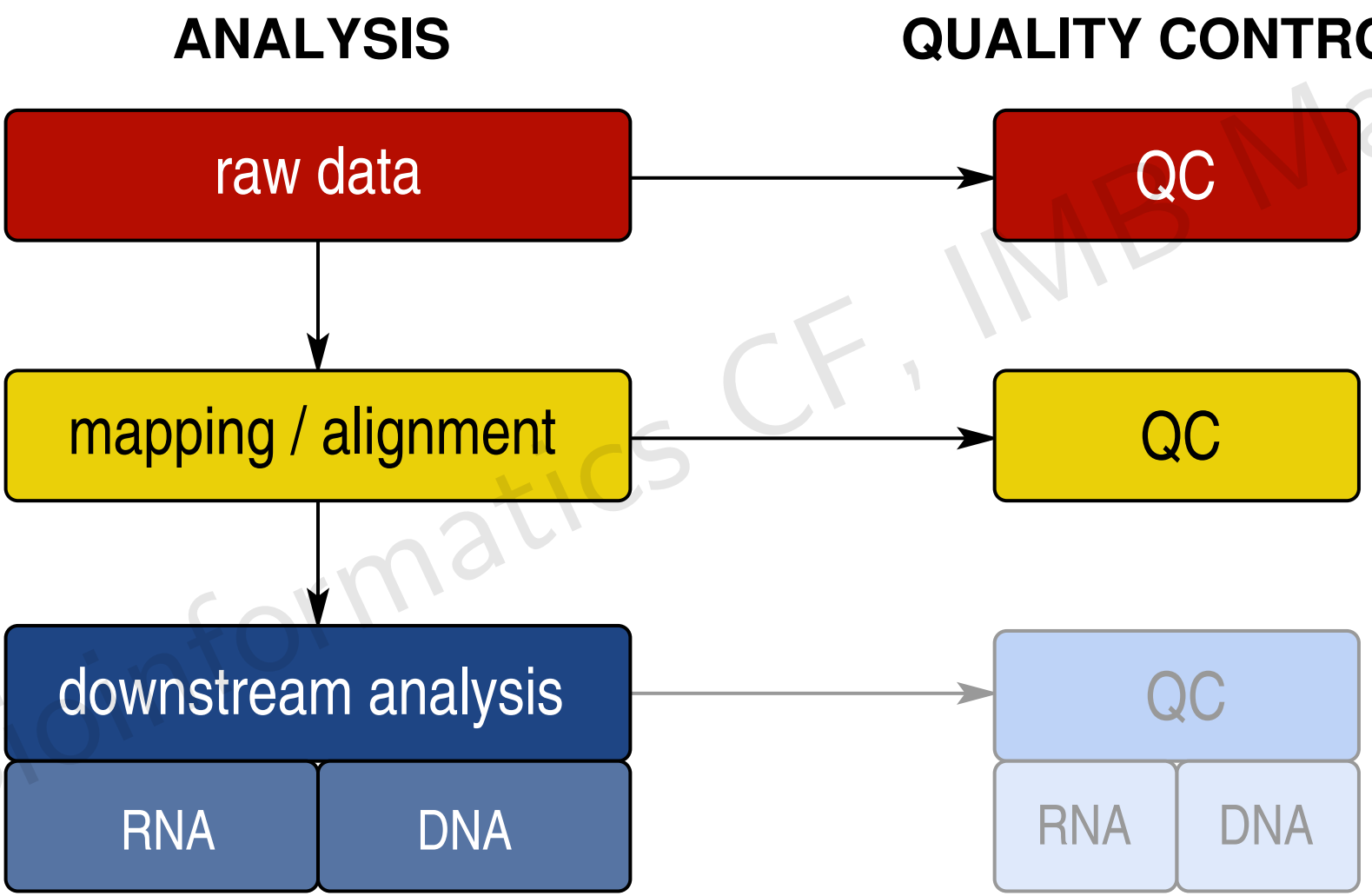
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GCCTAATTTAGGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA	834146	4.346867007222822	Illumina Paired End PCR Primer 2 (100% over 45bp)
GNCTAATTTAGGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTA	515410	2.685883195738773	Illumina Paired End PCR Primer 2 (100% over 45bp)
GCCTAATTTAGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTAT	27500	0.14330685839005114	Illumina Paired End PCR Primer 2 (100% over 46bp)

plot produced by fastqc: Andrews, S. "FASTQC. A quality control tool for high throughput sequence data." URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc (2010).

Institute of Molecular Biology

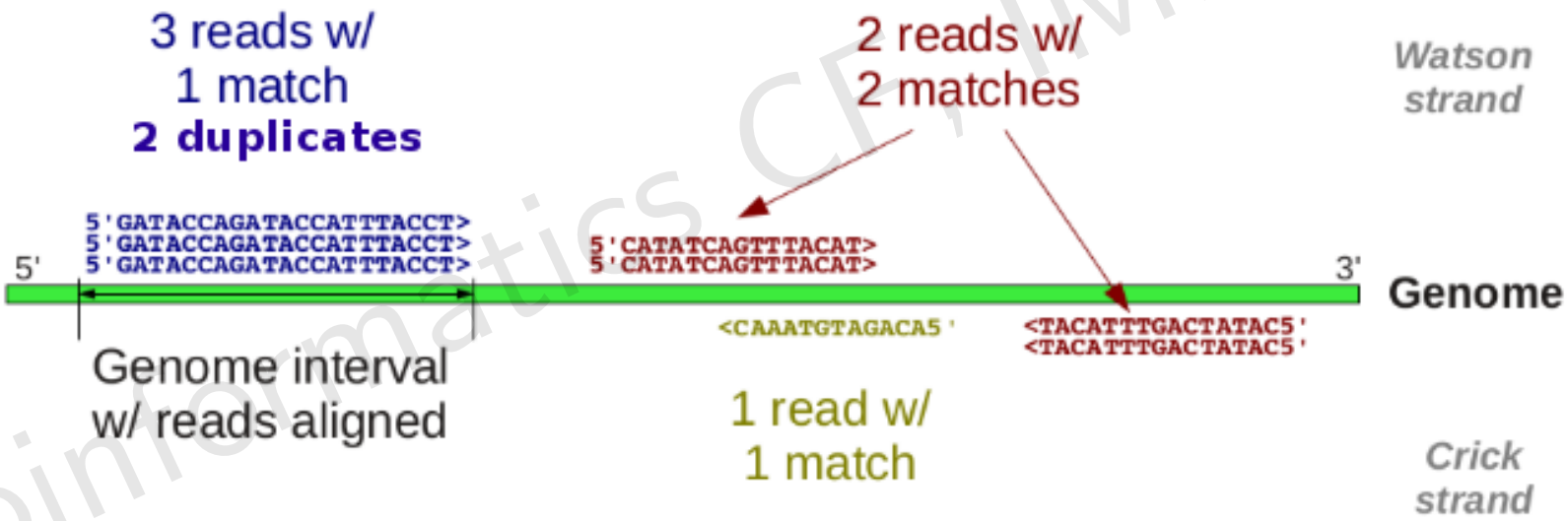




Alignment / Mapping quality control

- # reads mapped
- # reads unmapped
- # reads mapped to known contaminants
- # of uniquely mapped reads
- # duplicates
- expected read distribution pattern

Terminology



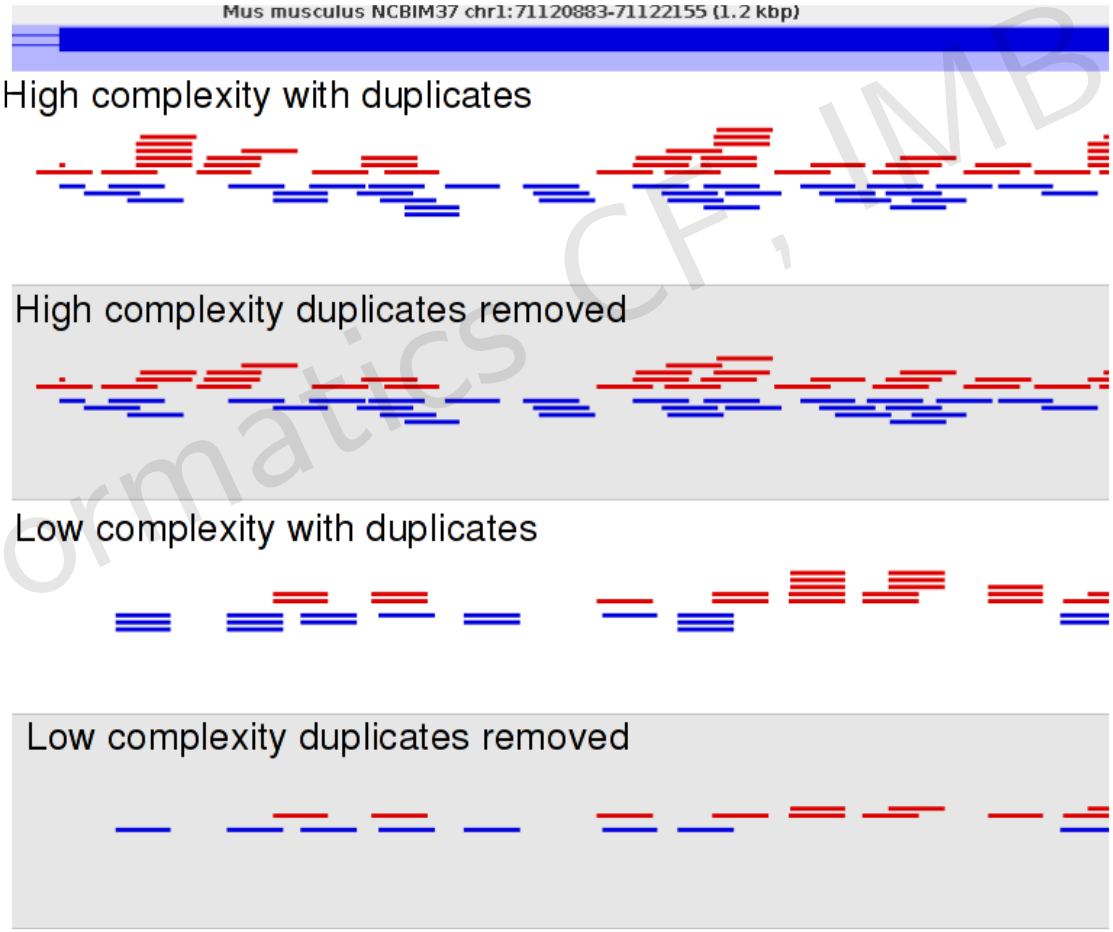
Graphic by Joern Toedling

Read duplication

Origins for Read Duplication

- biological
- technical (e.g. PCR amplification, optical duplicates)

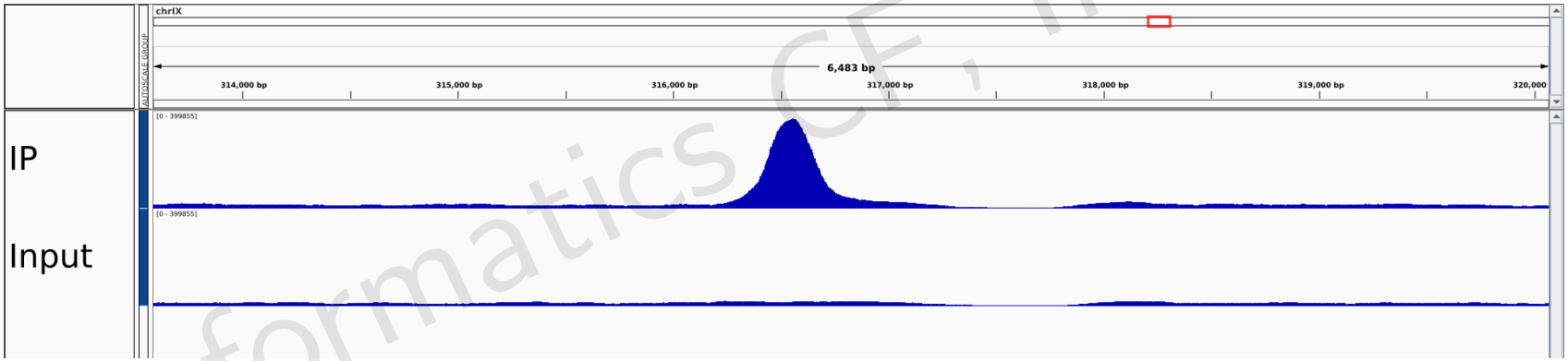
Visualising duplication



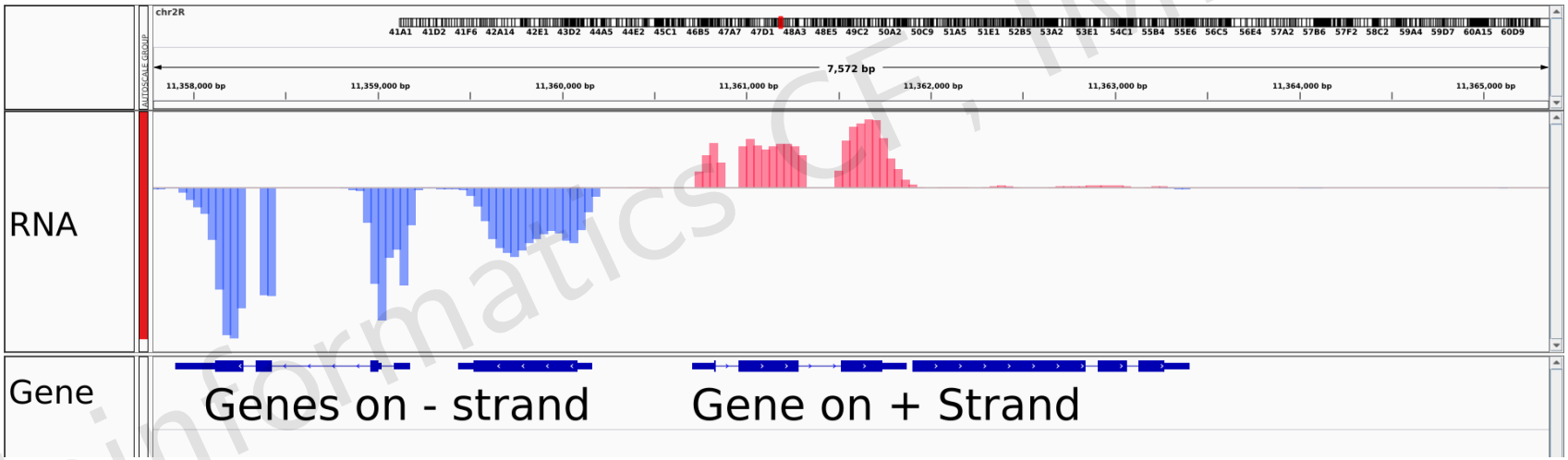
How to handle duplicate reads?

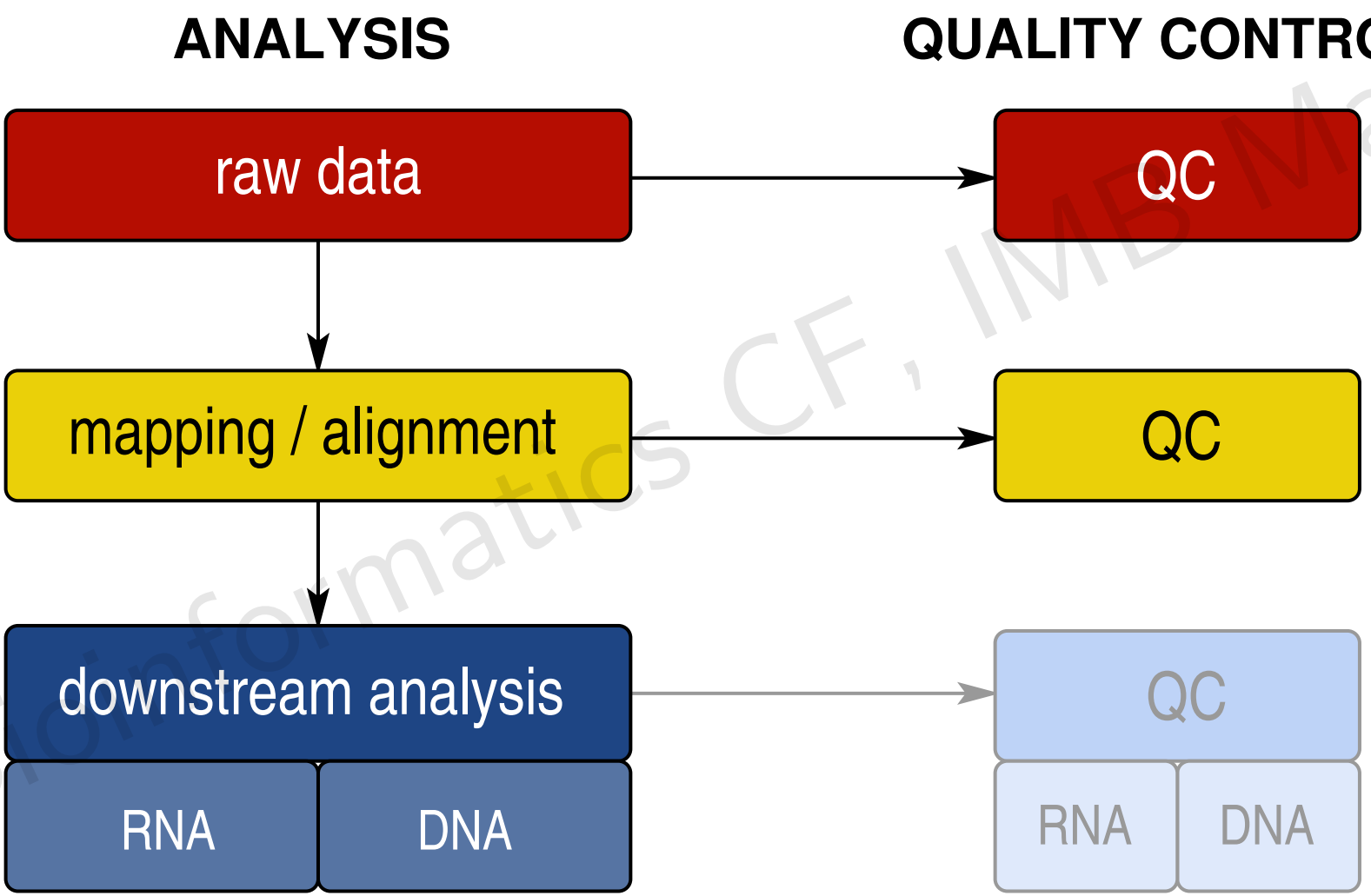
- DNA/ChIP-seq duplicate removal or estimation of biological duplication rate
- RNA-seq no duplication removal before analysis

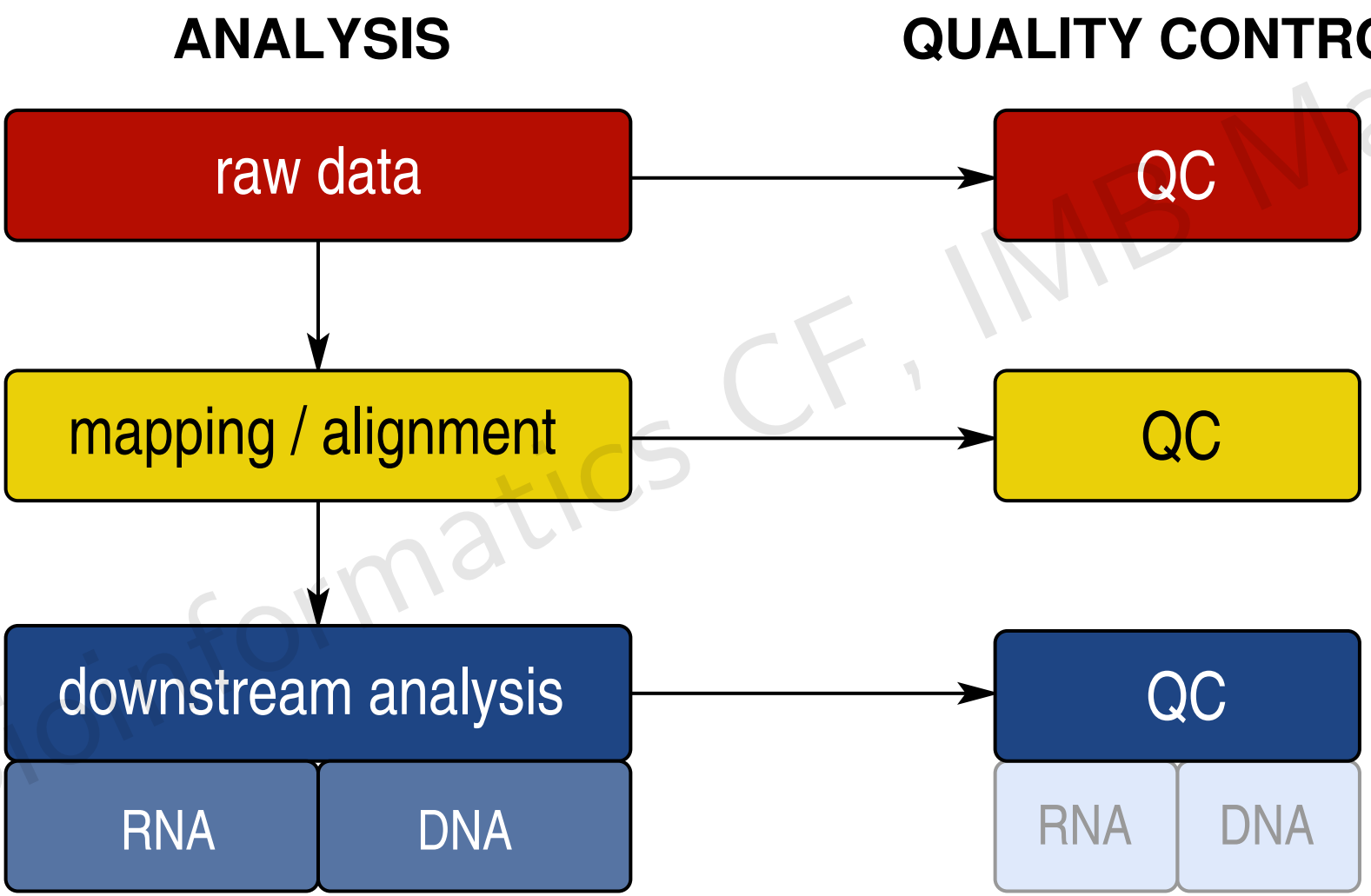
Read distribution pattern: ChIP

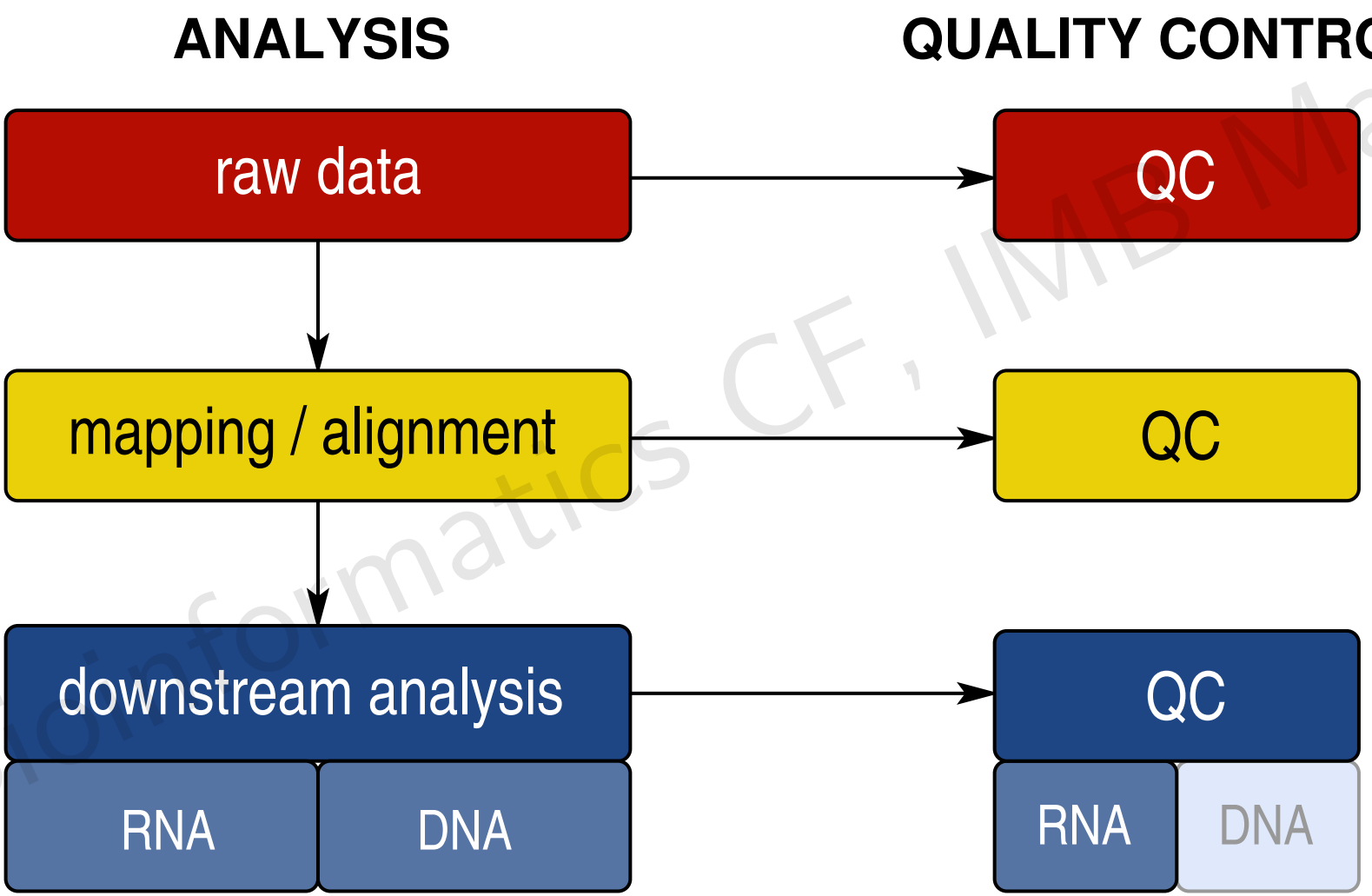


Read distribution pattern: RNA-Seq





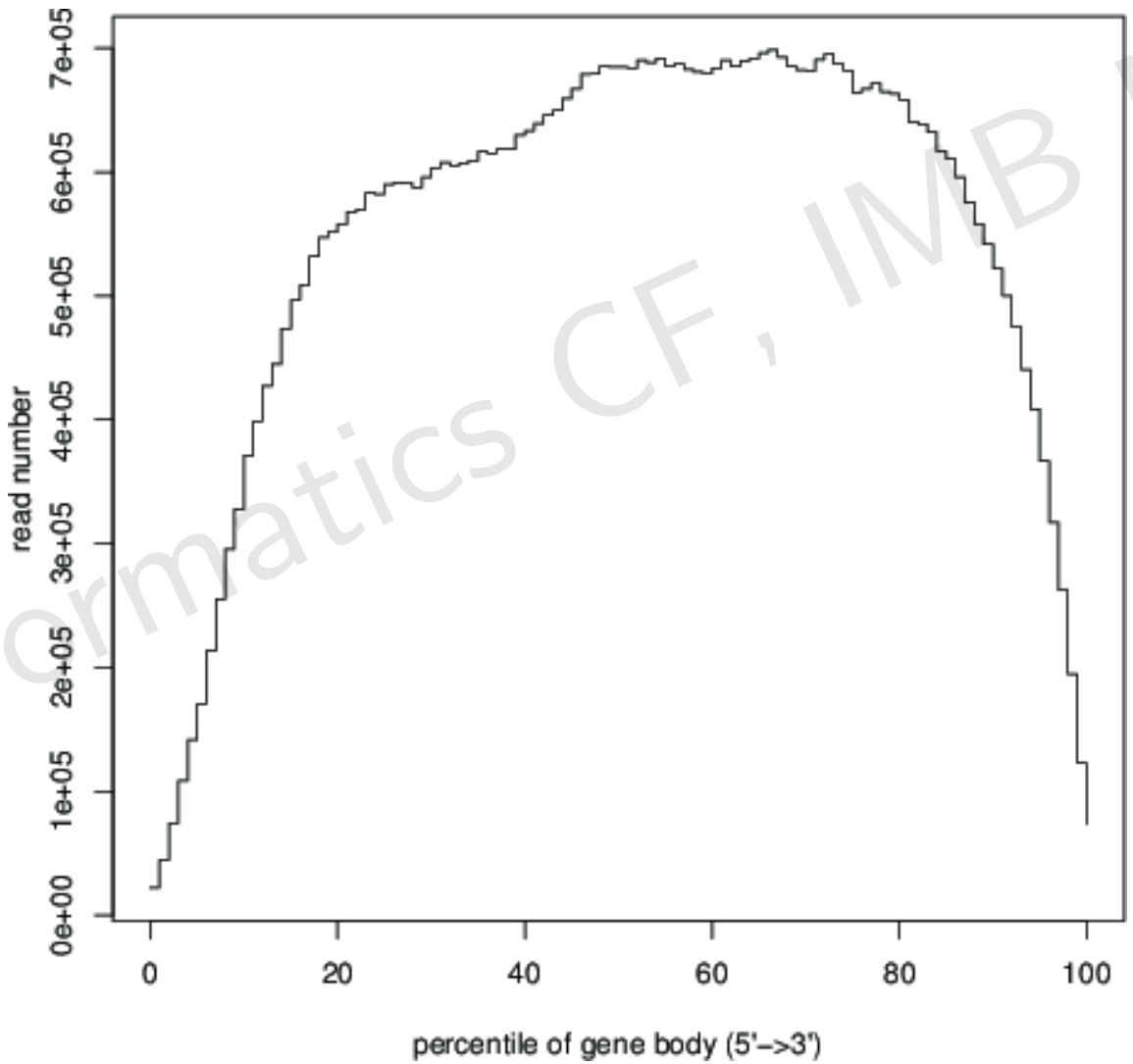




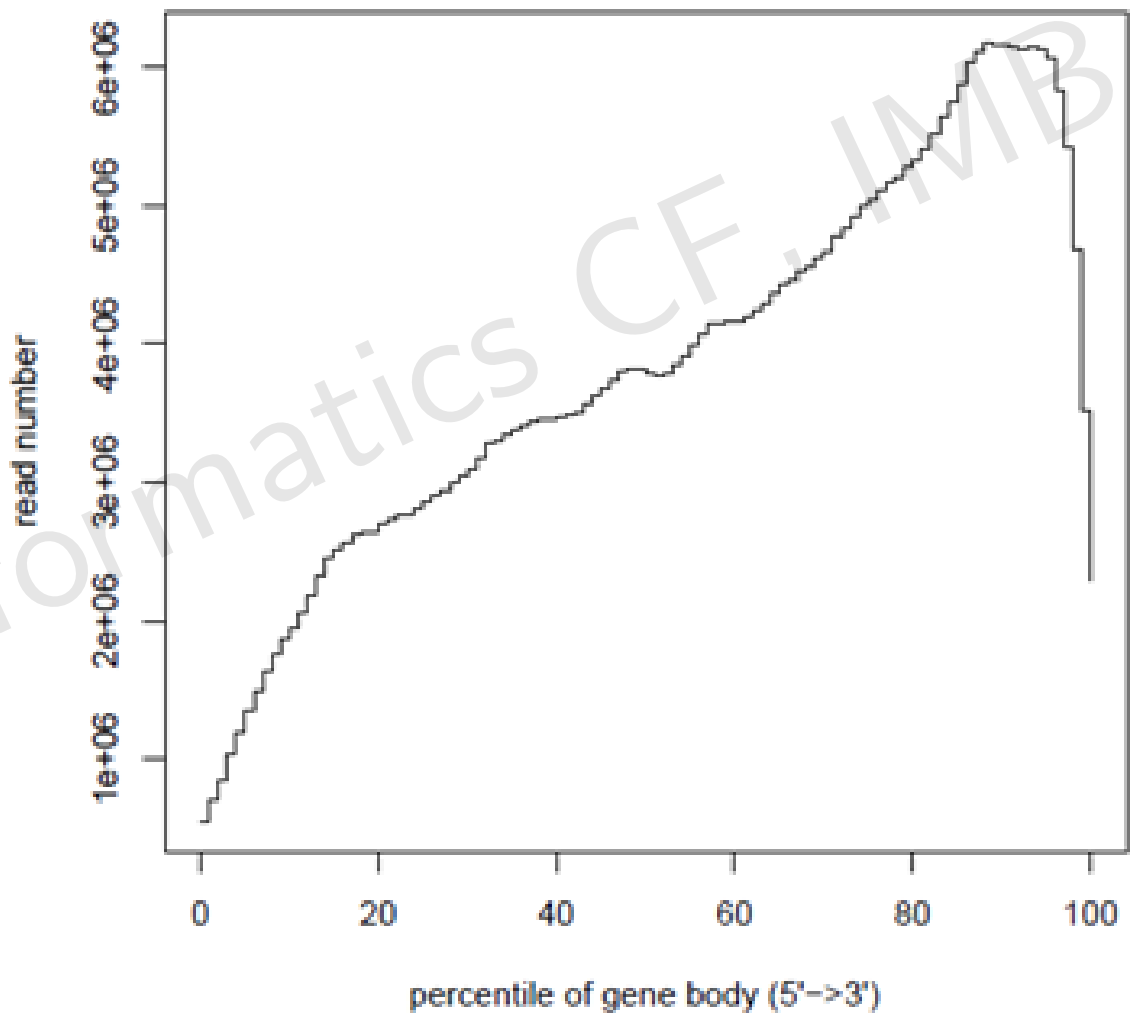
Quality control of RNA-seq

- sequencing depth (unique mapping reads)
- rRNA content
- 5' to 3' distribution of reads (gene body coverage)
- strand specificity
- duplication rate
- distribution of reads over different gene classes

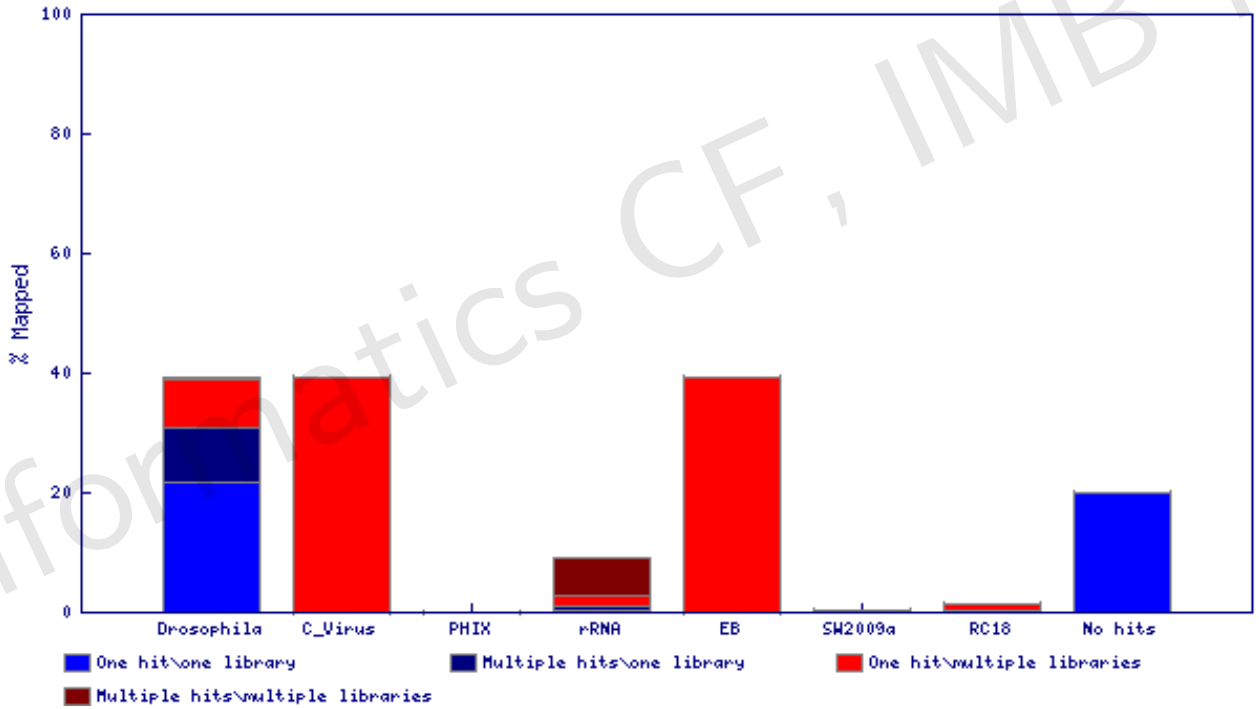
RNA-seq: 5' to 3' coverage



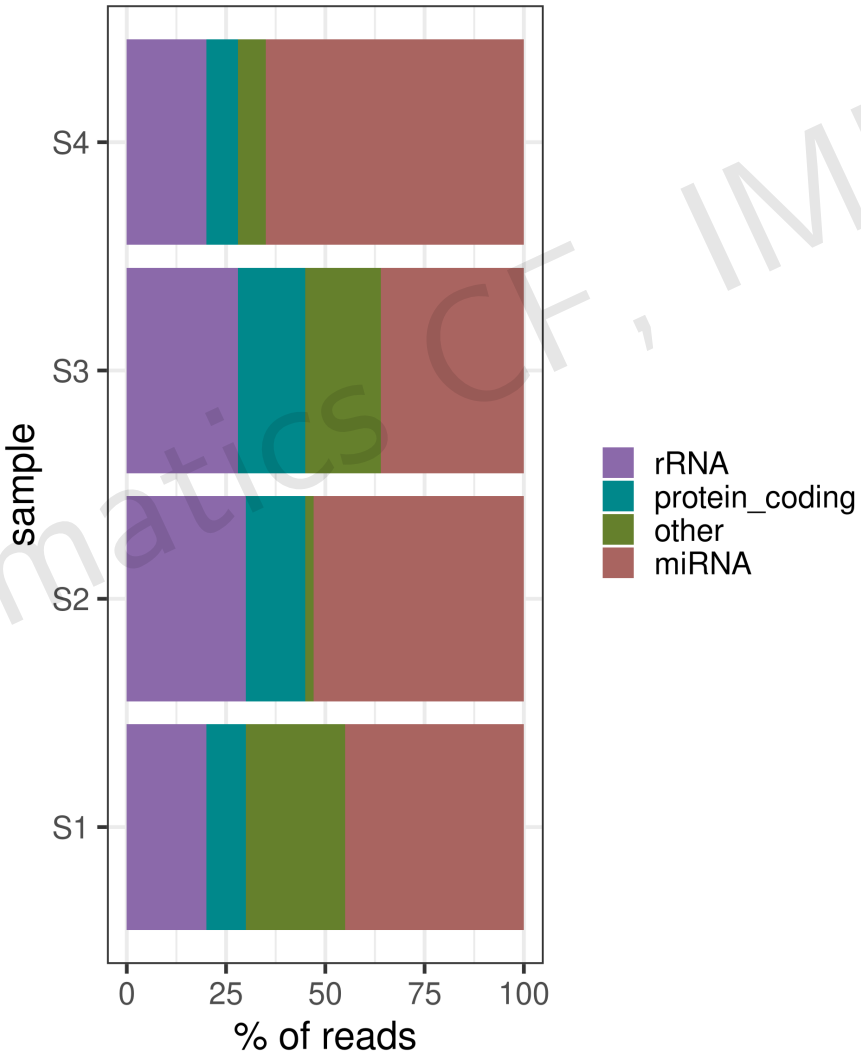
RNA-seq: 5' to 3' coverage



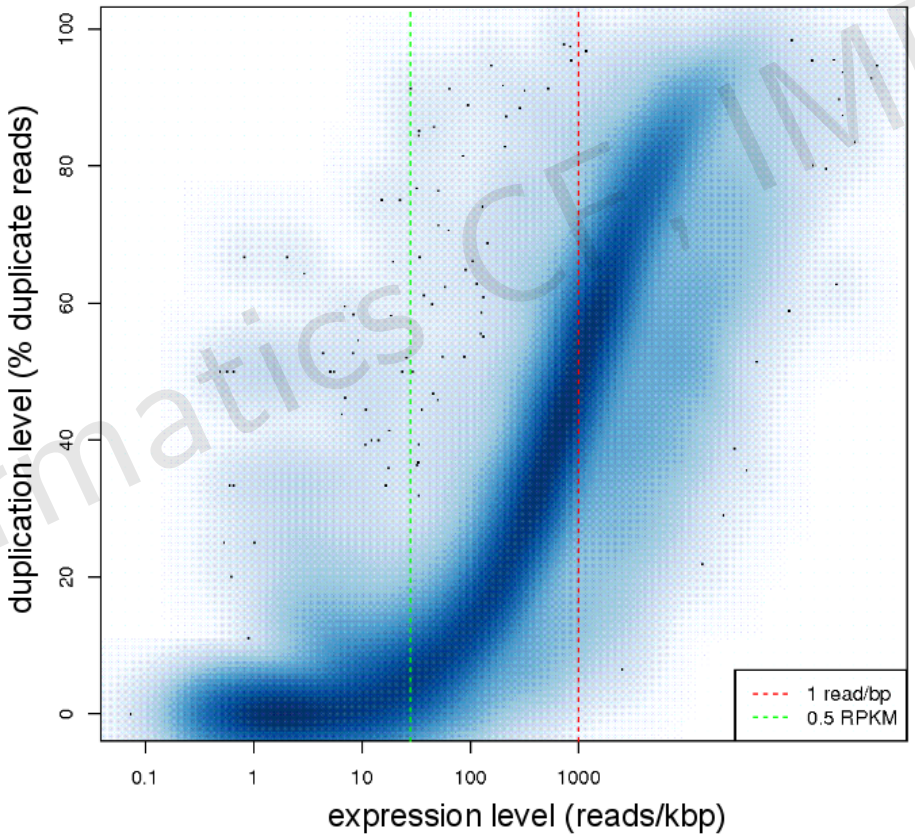
RNA-seq: Contamination Screening



RNA-seq: Counts on Features

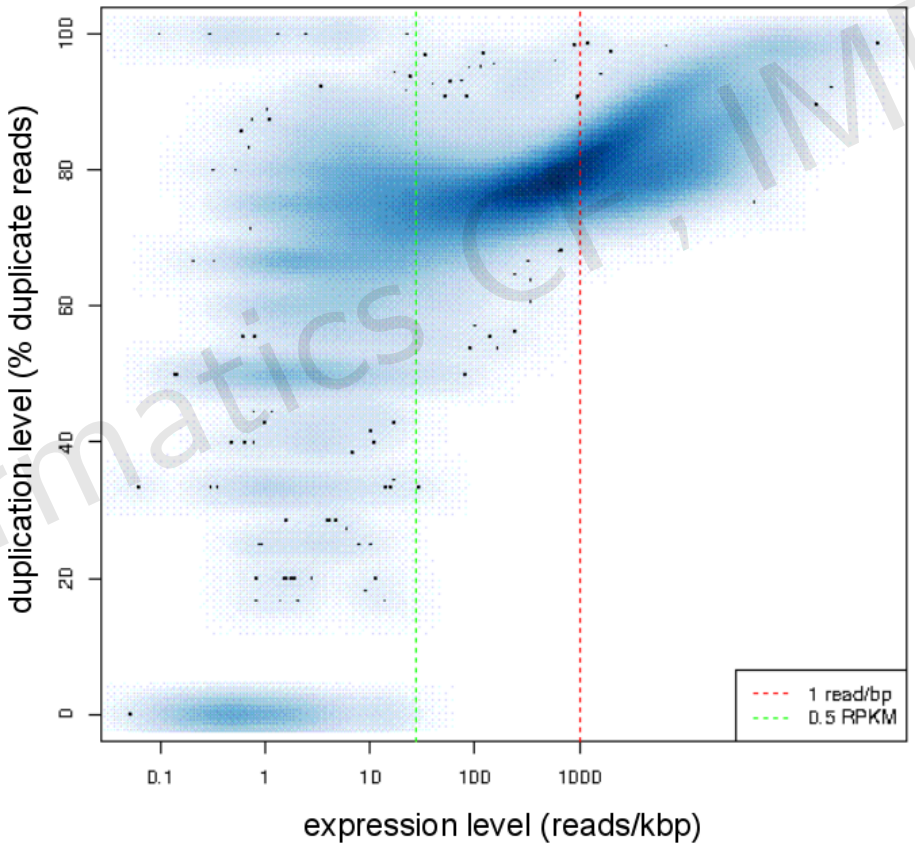


RNA-seq: duplication rate



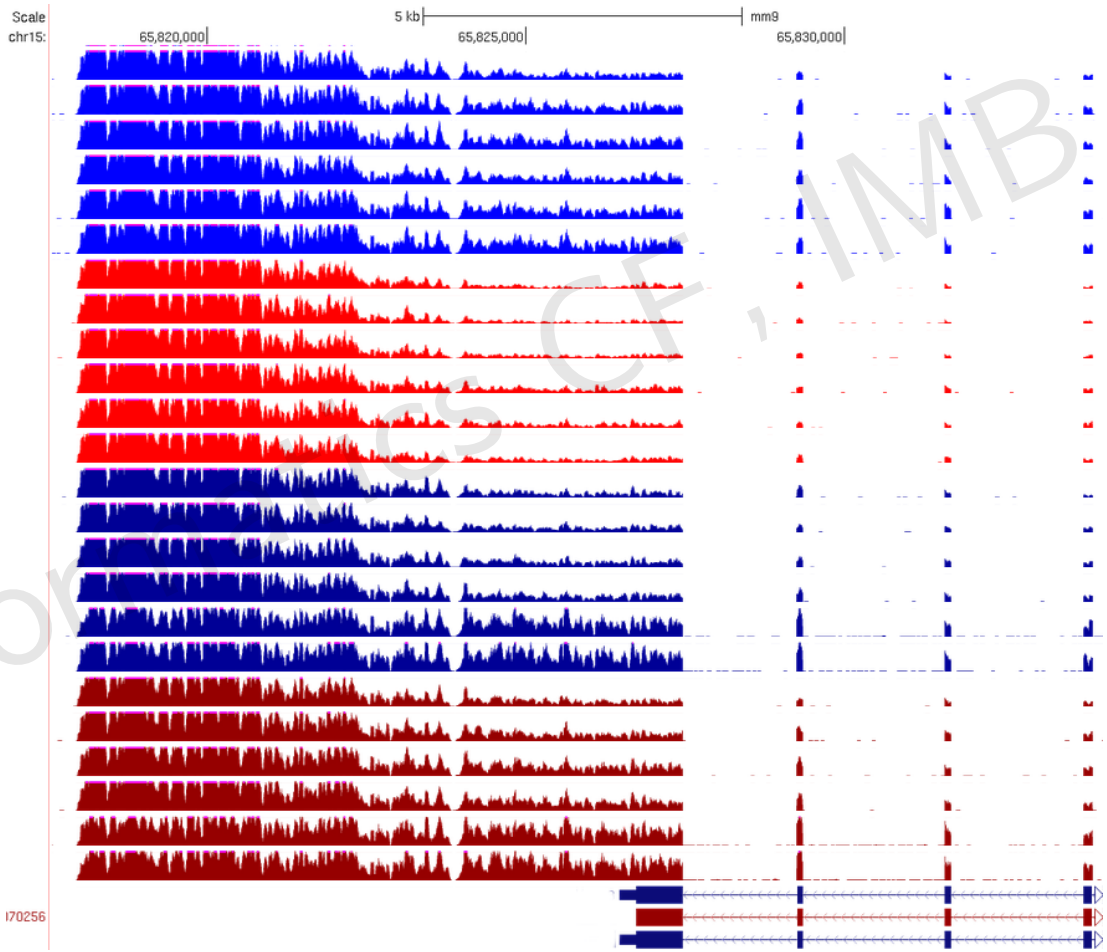
graphic from Holger Klein based on DupRadar

RNA-seq: low complexity library



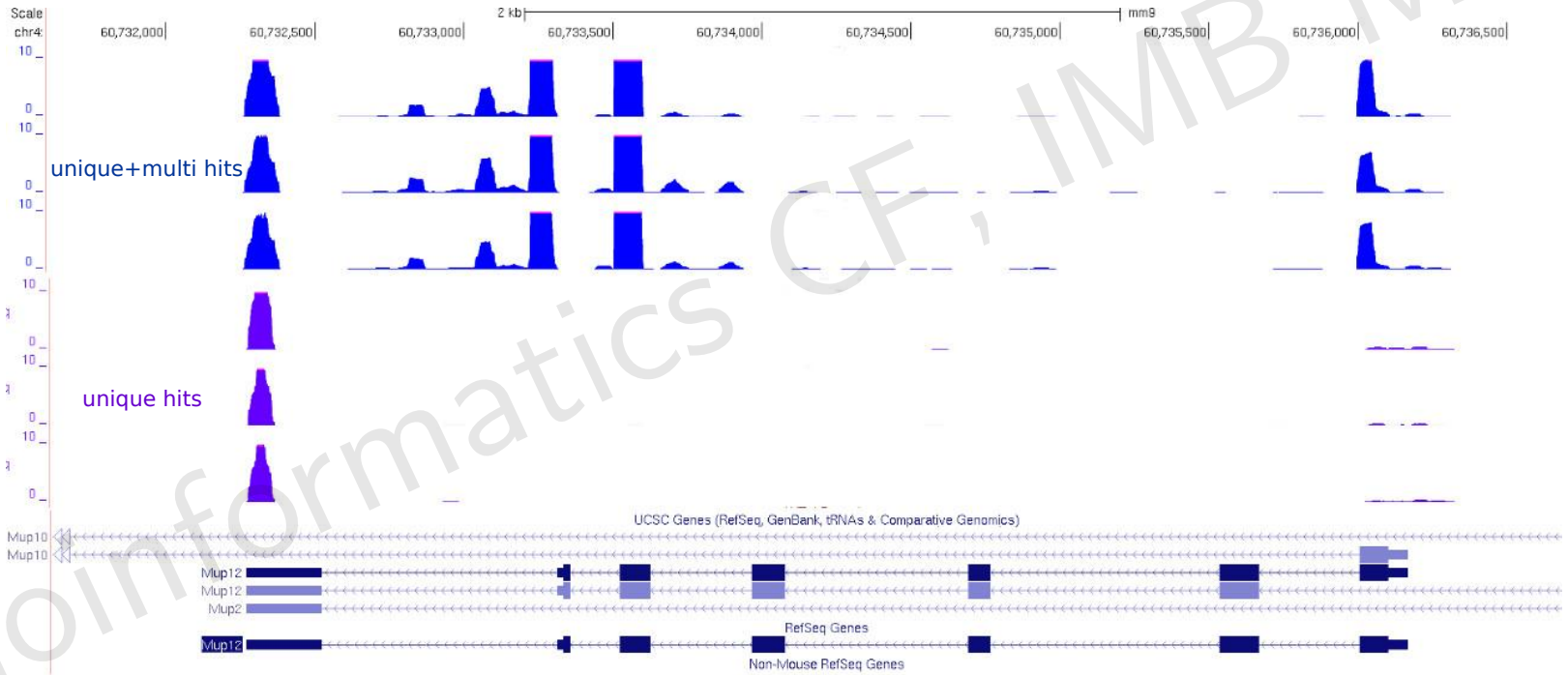
graphic from Holger Klein based on DupRadar

RNA-seq: annotation issues

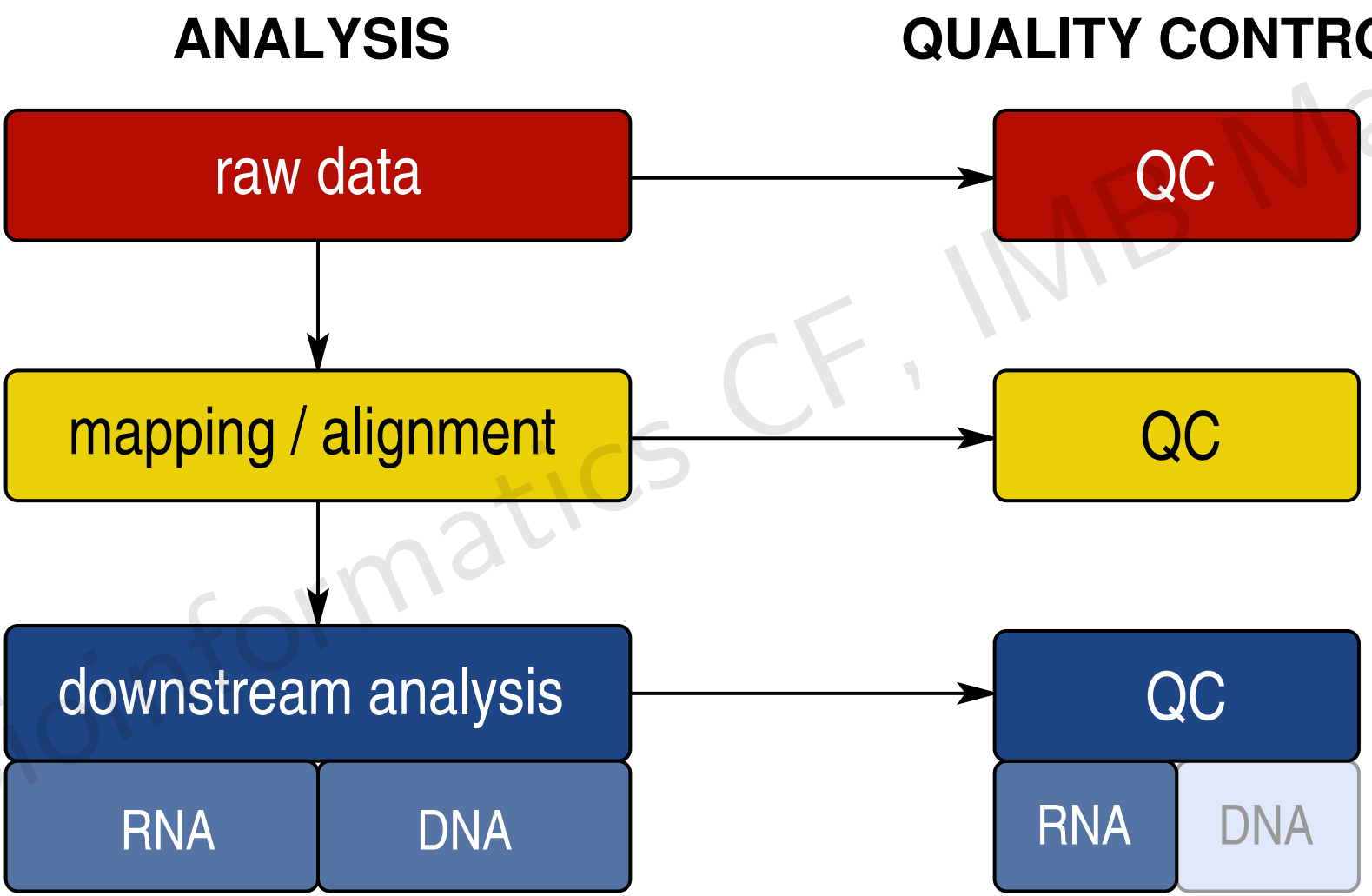


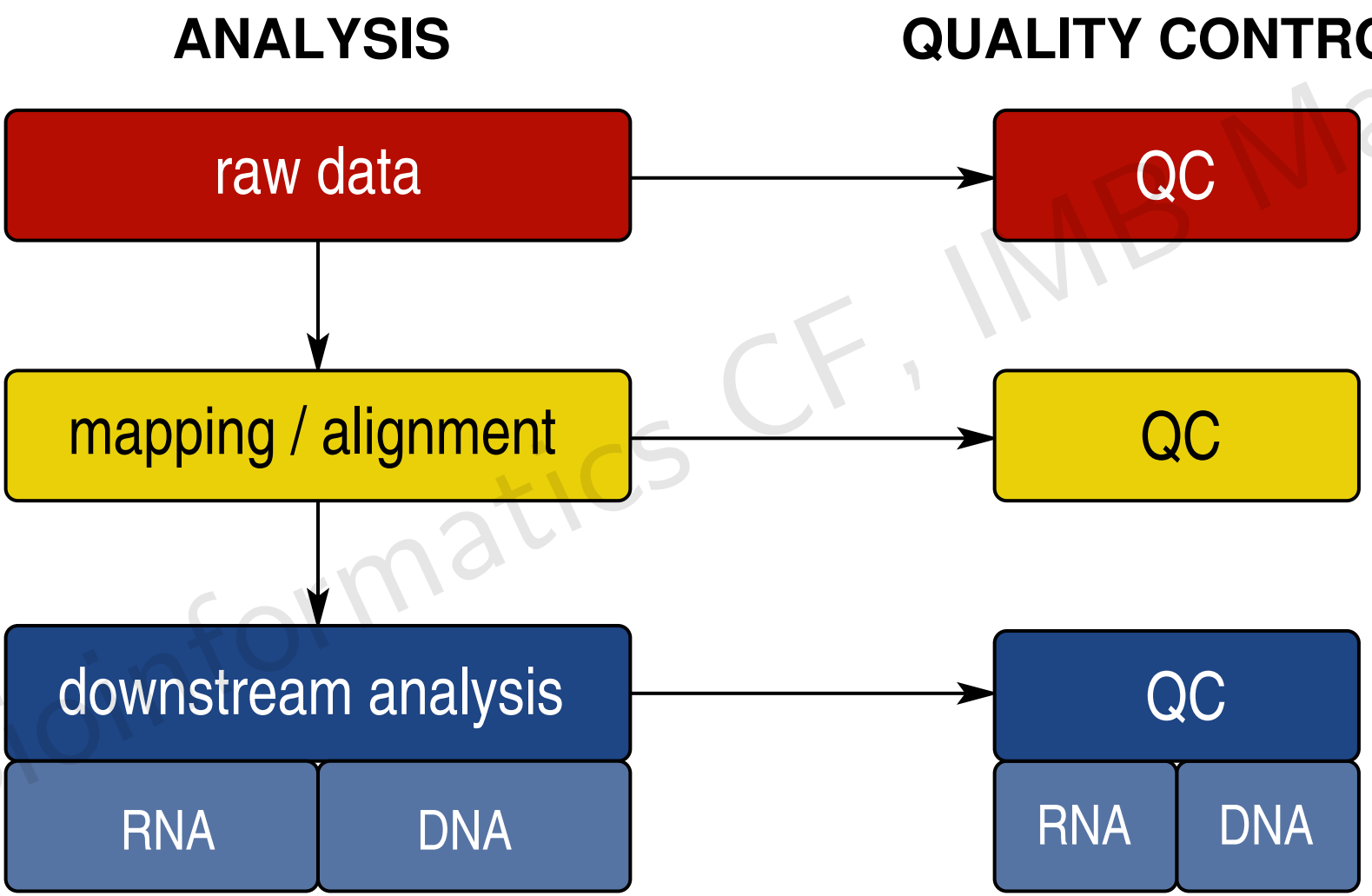
in house data (mm9), figure adapted from Emil Karaulanov

RNA-seq: multi-mapping reads

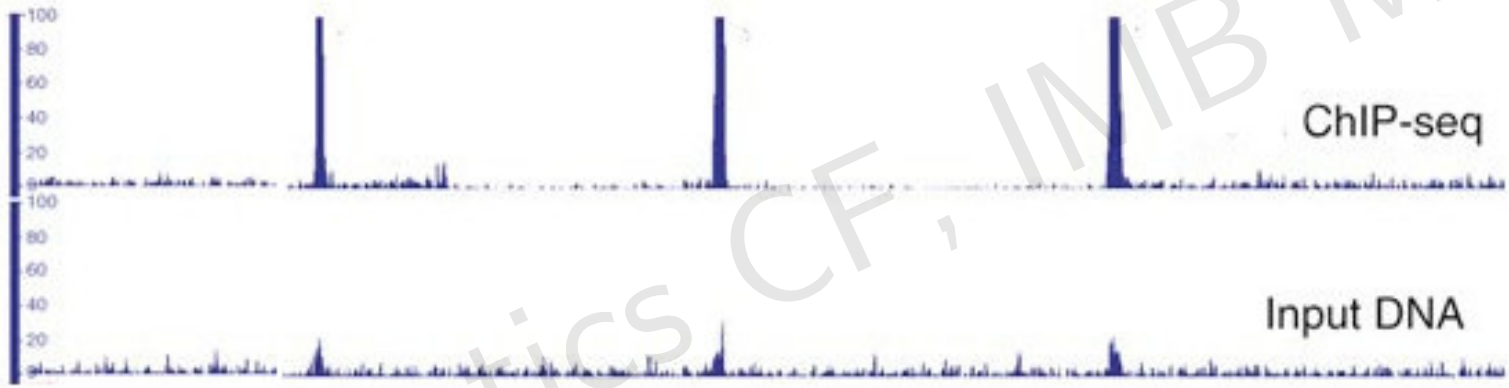


in house data (mm9), figure adapted from Emil Karaulanov





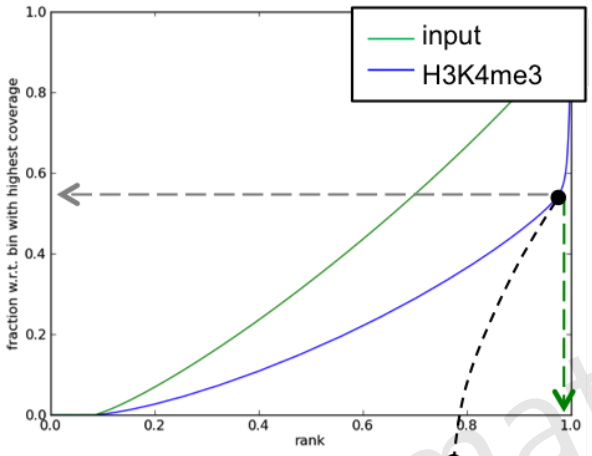
ChIP: input control



- biases by sonication
- large genomic variation (e.g. Aneuploidy, large InDels, CNV)
- artefacts of preparation

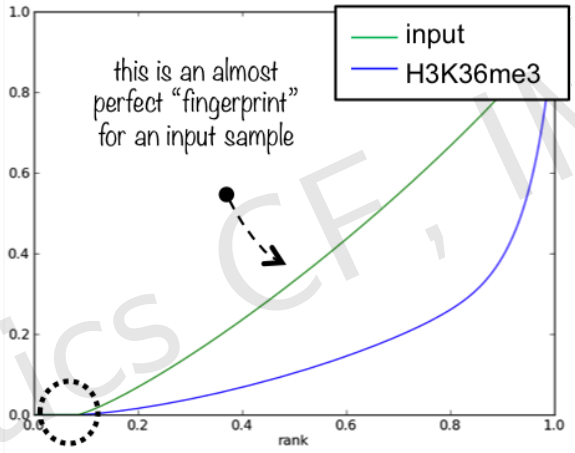
Rozowsky et al., Nature Biotech, 2009

ChIP-seq: enrichment quality control IPStrength

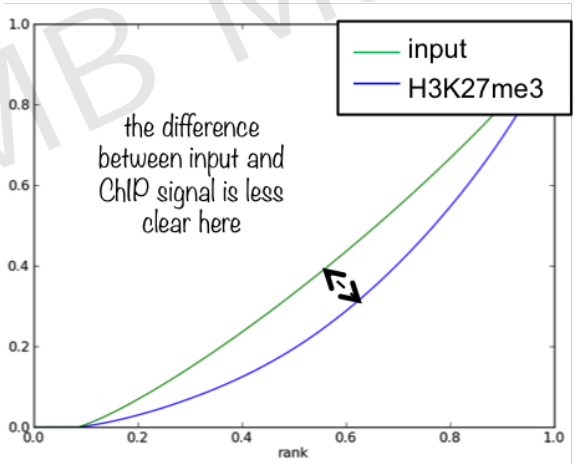


when counting the reads contained in 97% of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

this indicates very localized, very strong enrichments!
(as every biologist hopes for in a ChIP for H3K4me3)



pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)

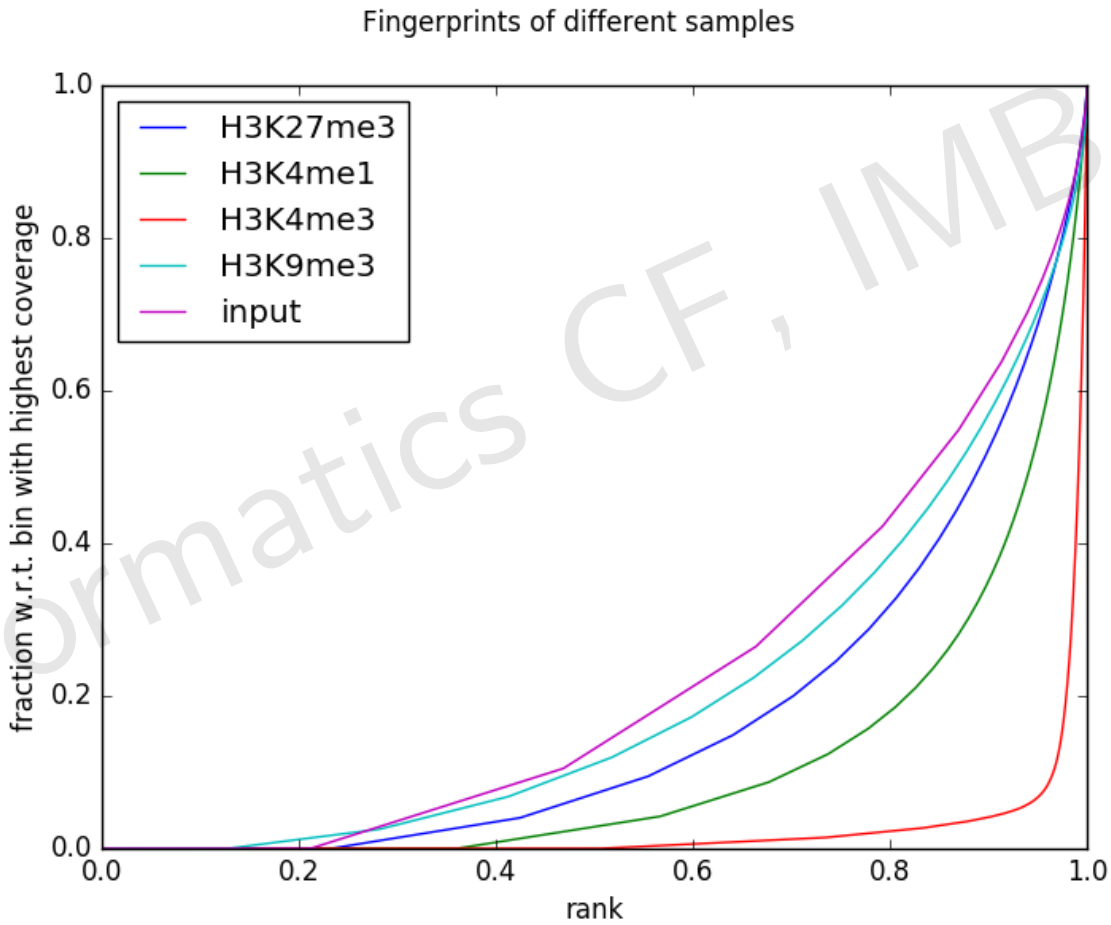


H3K27me3 is a mark that yields broad domains instead of narrow peaks

it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

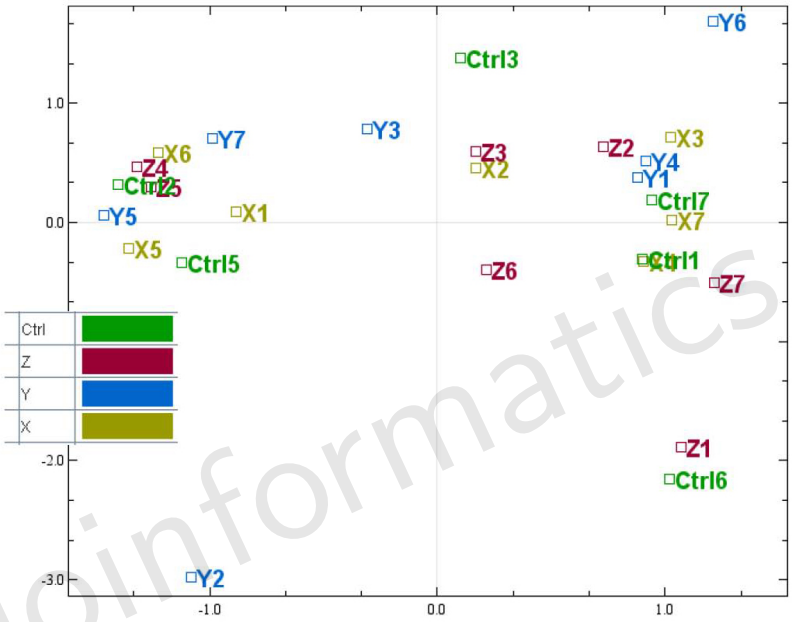
Fidel et al deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. Nucleic Acids Research (2016).

ChIP-seq: enrichment quality control IPStrength



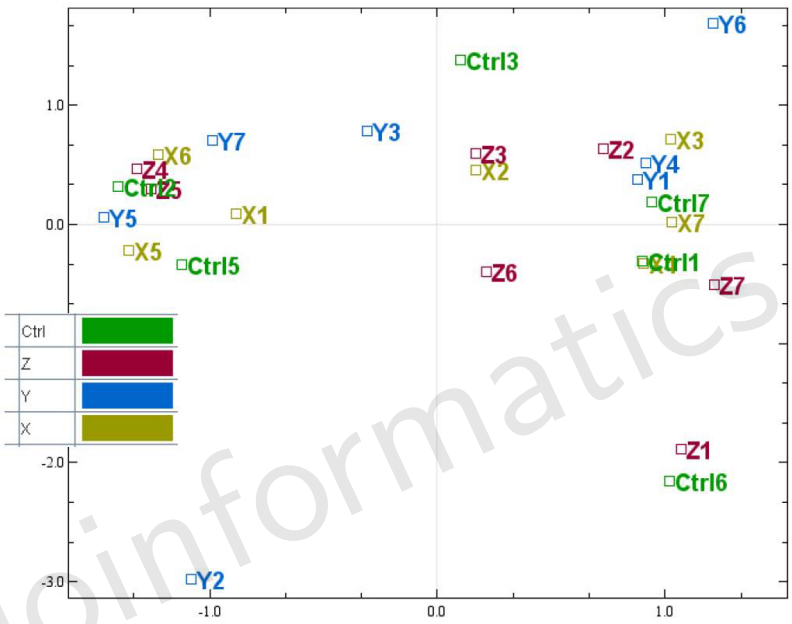
Fidel et al deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. Nucleic Acids Research (2016).

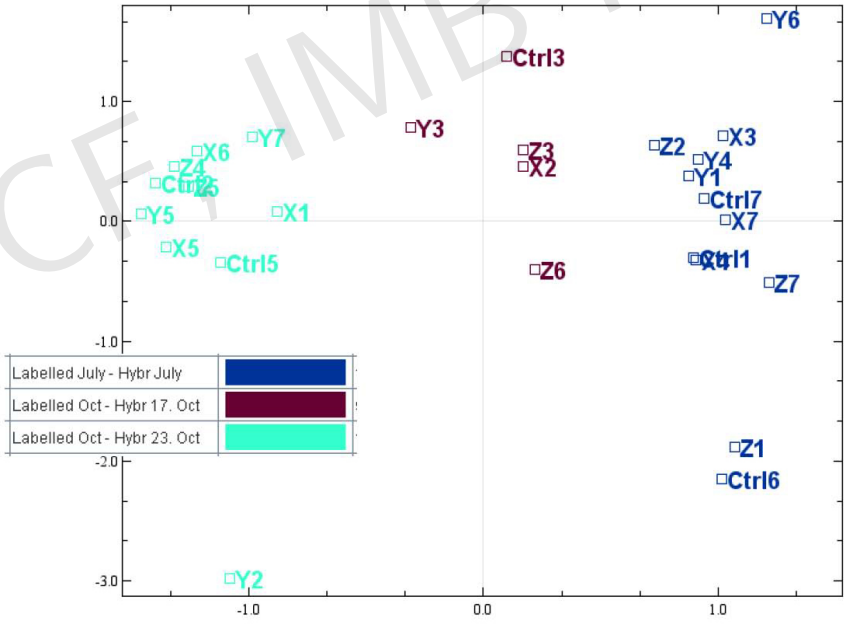
Batch effects



Source: http://www.molmine.com/magma/global_analysis/batch_effect.html

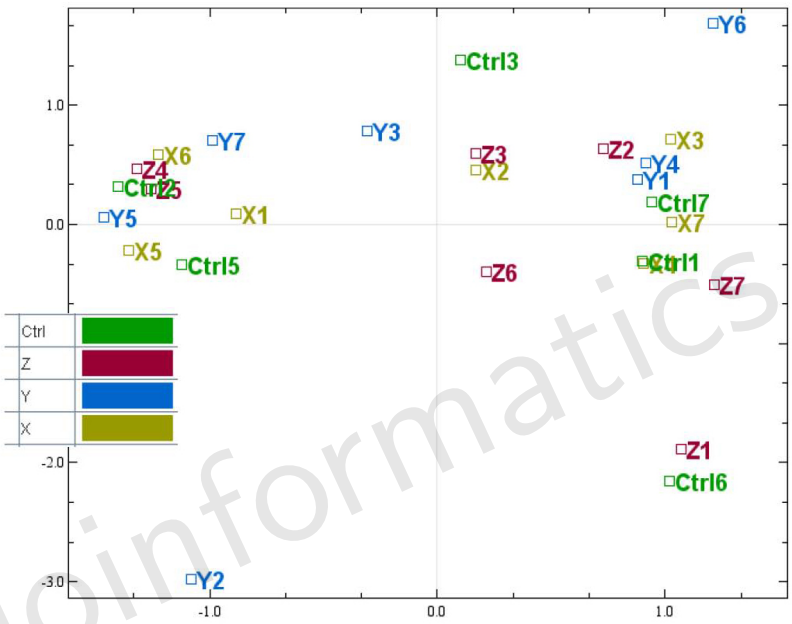
Batch effects

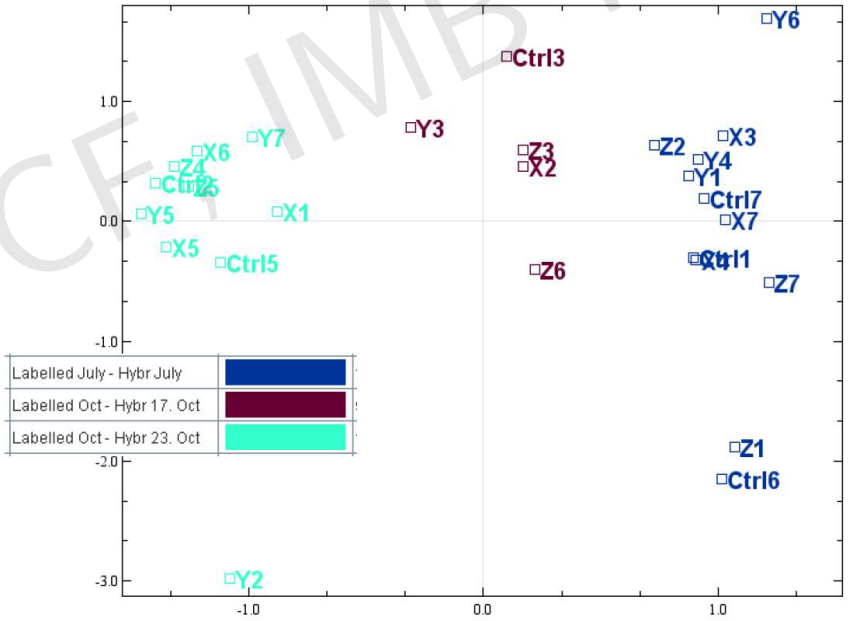




Source: http://www.molmine.com/magma/global_analysis/batch_effect.html

Batch effects

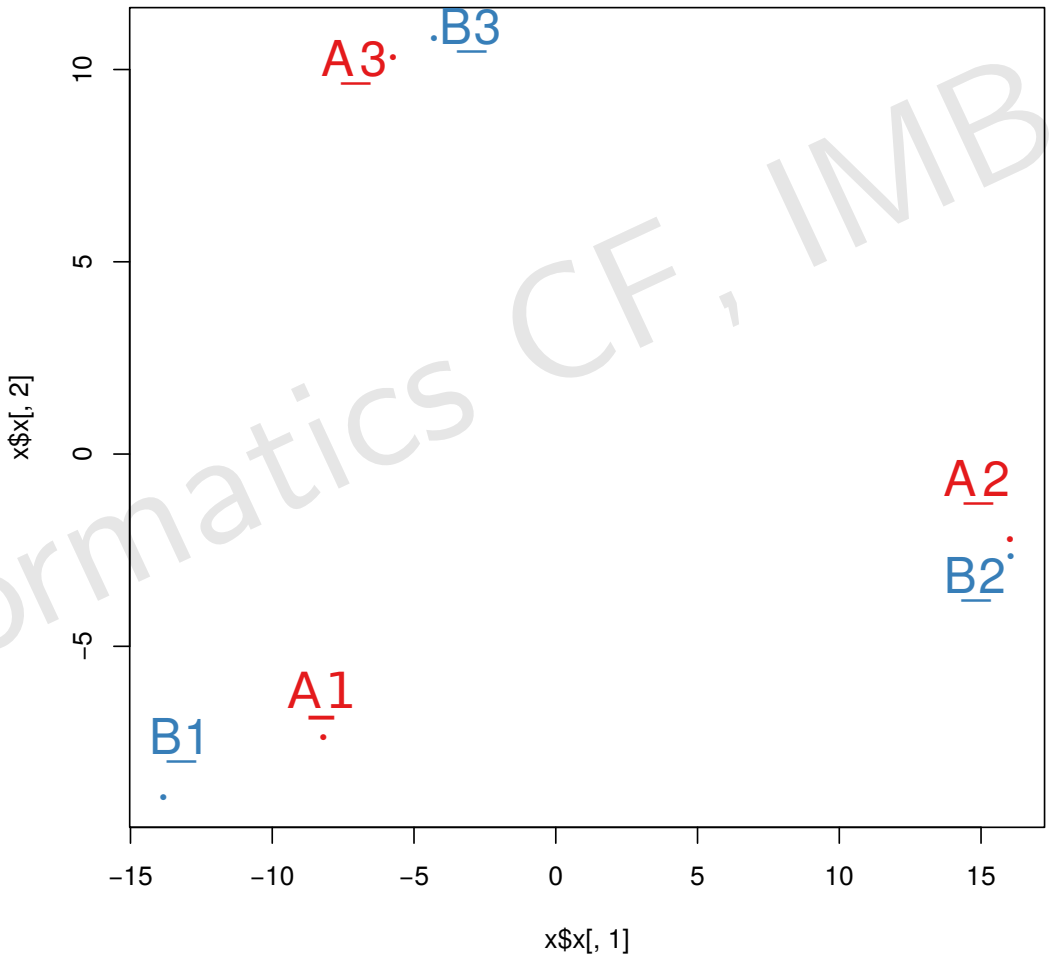




$\Rightarrow$ if batch processing is unavoidable each sample group should be represented within each batch.

Source: `http://www.molmine.com/magma/global_analysis/batch_effect.html`

Batch effects



Guidelines for Experiments

- Encode ChIP-seq, DNA-seq Guidelines
- Encode RNA-seq Standards
- ...

Encode Guidelines: `https://www.encodeproject.org/data-standards/`

Summary II

- exploratory data analysis to ensure data quality

Summary II

- exploratory data analysis to ensure data quality
- use known unbiased quality control methods on the different analysis levels

Summary II

- exploratory data analysis to ensure data quality
- use known unbiased quality control methods on the different analysis levels
- investigate the similarities between samples (principal component analysis and clustering)

Summary II

- exploratory data analysis to ensure data quality
- use known unbiased quality control methods on the different analysis levels
- investigate the similarities between samples (principal component analysis and clustering)
- visualise your data

Tools

- Rawdata: FastQC
- Mapper: Bowtie, BWA, STAR, HiSat2
- Mapping-QC: samtools, Picard tools, qualimap, deepTools
- Duplication: bamUtils, DupRadar
- RNA diff. expression: DESeq2, edgeR
- RNA-seq QC: RNASeqQc, RNAQC
- ChIP-seq QC: deepTools
- Peak calling: MACS2
- Contamination: BLAST, FastqScreen
- Misc Analysis: SeqMonk
- Visualisation: IGV, SeqMonk, UCSC genome browser, Washington epigenome browser

Acknowledgements

- Holger Klein
- Bioinformatics Core Facility:
  - Emil Karaulanov
  - Fridolin Kielisch
  - Martin Oti
  - Giuseppe Petrosino
  - Frank Rühle
  - Sergi Sayols Puig

Acknowledgements

- Holger Klein
- Bioinformatics Core Facility:
  - Emil Karaulanov
  - Fridolin Kielisch
  - Martin Oti
  - Giuseppe Petrosino
  - Frank Rühle
  - Sergi Sayols Puig

Thank you for your attention.

imb Institute of Molecular Biology