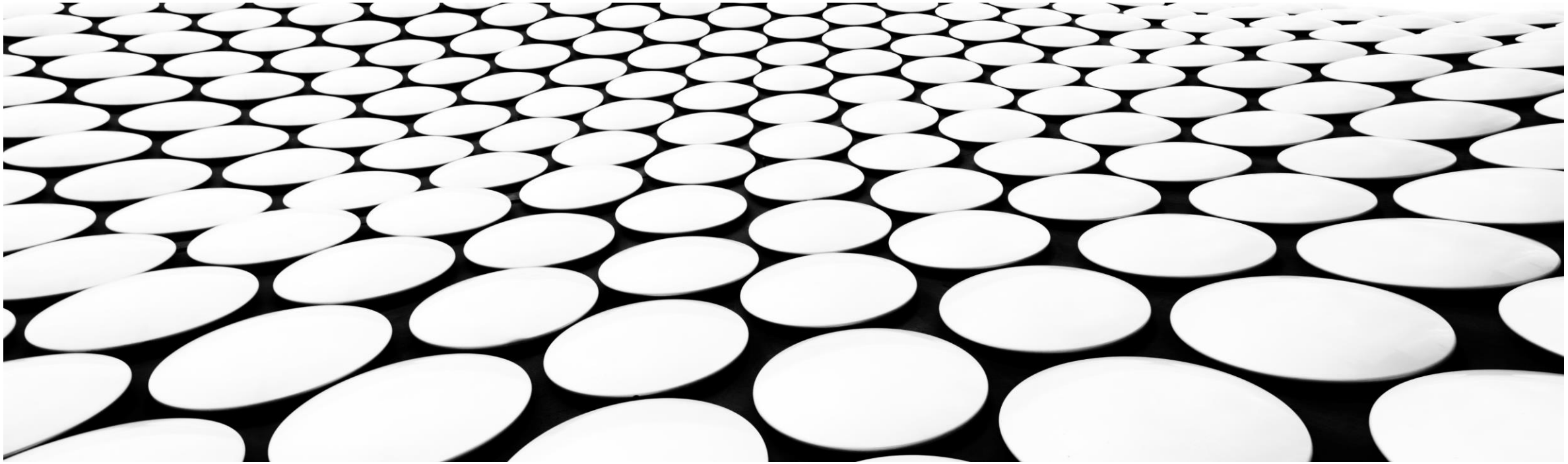
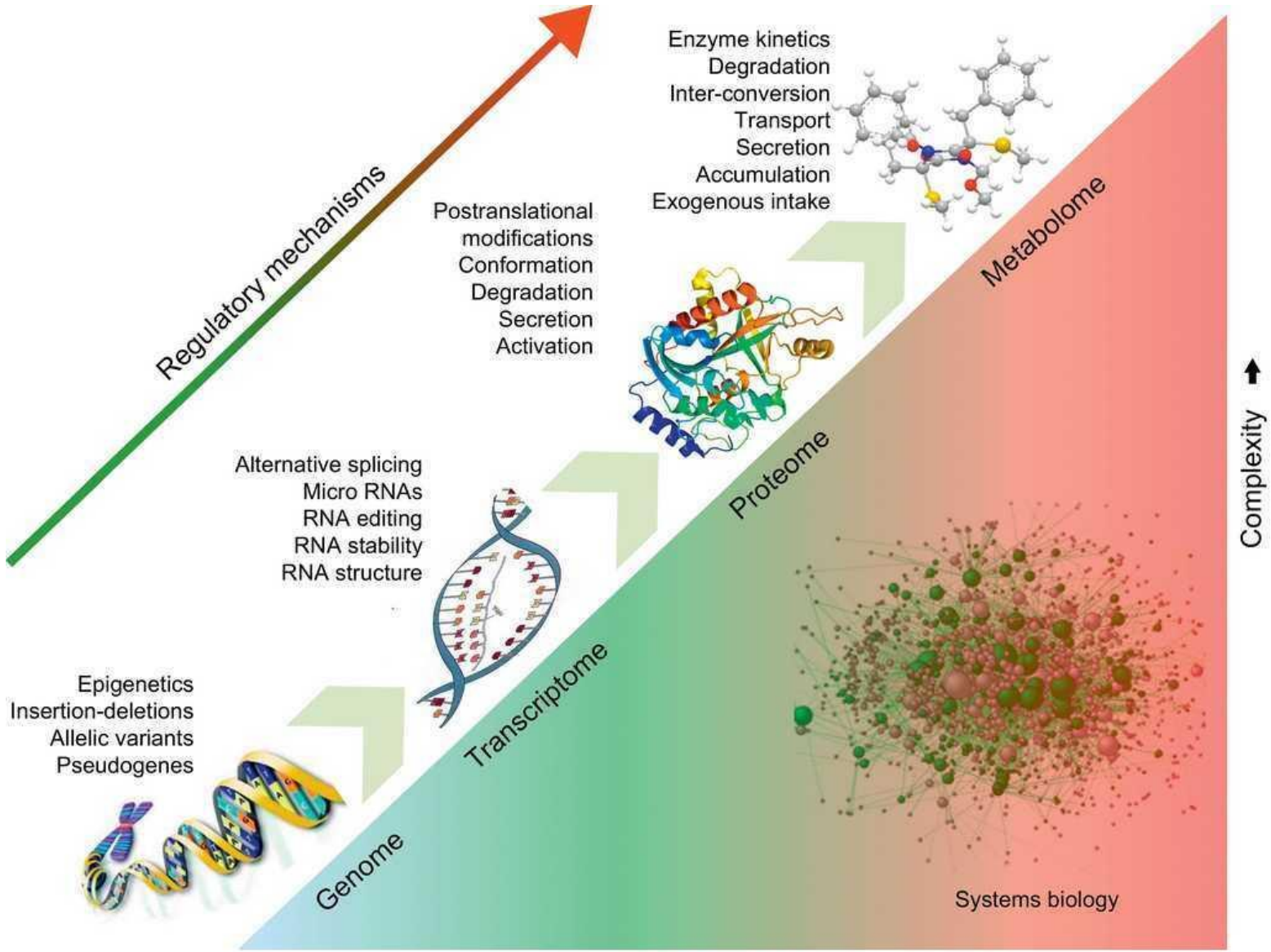
RNA-SEQ



TRREP: TRANSCRIPTÓMICA, REGULACIÓN GENÓMICA Y EPIGENÓMICA

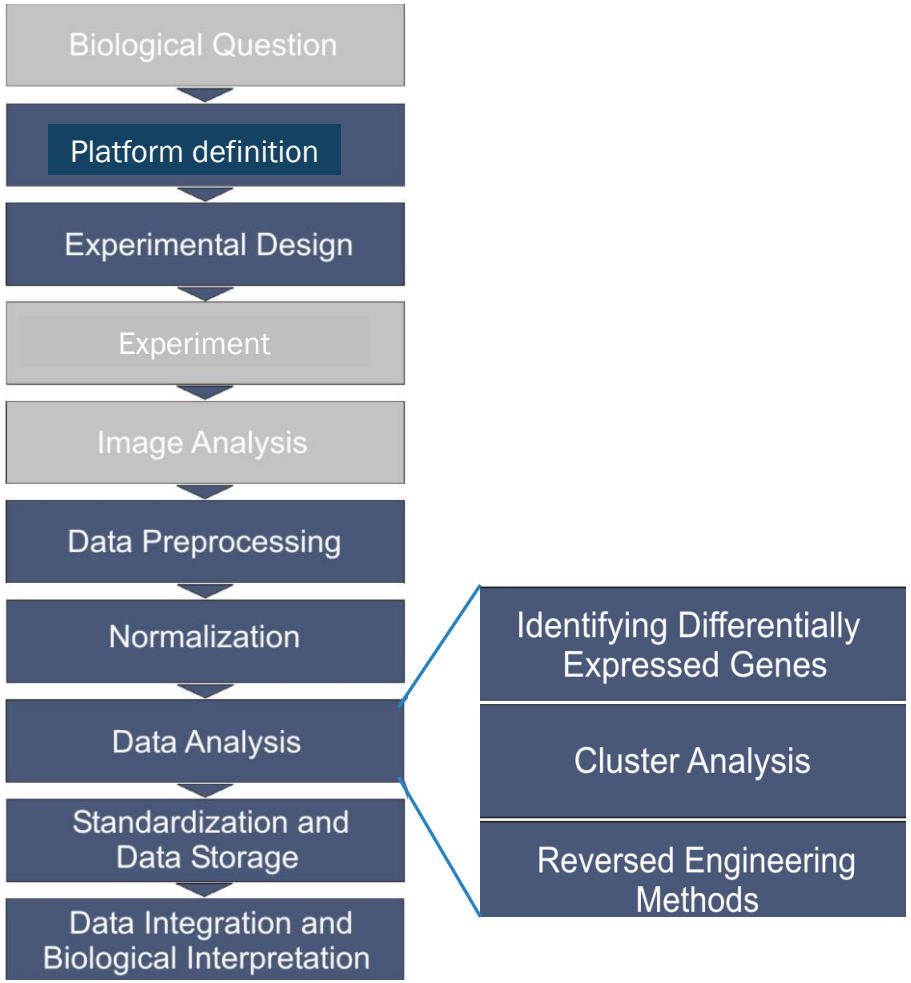
Calificación = 0.4*Trabajo + 0.5*Examen teórico+0.1*participación

Fecha	semana	Hora	Dia semana	Bloque	Clase	Horas	Profesor	Aula
29/01/2025	1	18:00-20:00	X	1	Experimental design	2	FSC	Seminario 1 (Decanato)
31/01/2025	1	18:00-20:00	V	1	Statistical principles of omics data	2	FSC	Seminario 3 (Decanato)
					BLOQUE 1: TRANSCRIPTOMICA			
05/02/2025	2	18:00-20:00	X	2	RNA-Seq pipeline (1): Pipeline general, alineadores	2	FSC	Seminario 1 (Decanato)
07/02/2025	2	18:00-20:00	V	2	RNA-Seq pipeline (2): Expresion diferencial	2	FSC	Seminario 1 (Decanato)
12/02/2025	3	18:00-20:00	X	2	RNA-Seq pipeline (3): Análisis funcional	2	FSC	Seminario 1 (Decanato)
14/02/2025	3	18:00-20:00	V	4	CHIP-SEQ	2	CTF	Seminario 1 (Decanato)
					BLOQUE 2: PROTEOMICA & Metabolomica			
19/02/2025	4	18:00-20:00	X	3	Intro: Proteomics and Mass spectometry	2	JMR	Seminario 1 (Decanato)
21/02/2025	4	18:00-20:00	V	3	Protein identification & PTMs (I): Practical exercise	2	JMR	Seminario 1 (Decanato)
26/02/2025	5	18:00-20:00	X	3	Protein identification & PTMs (II): Practical exercise	2	JMR	Seminario 1 (Decanato)
05/03/2025	6	18:00-20:00	X	3	Functional annotation and visualization: Practical exercise	2	JMR	Seminario 1 (Decanato)
07/03/2025	6	18:00-20:00	V	3	Introduction to metabolomics	2	JMR	Seminario 1 (Decanato)
					BLOQUE 3: REGULACION DE LA EXPRESION GENICA			
12/03/2025	7	16:00-19:00	X	4	Intro. Regulación Genómica	2	EC	Seminario 1 (Decanato)
14/03/2025	7	18:00-20:00	V	4	práctica chromatin states (I)	2	EC	Seminario 1 (Decanato)
21/03/2025	8	16:00-19:00	V	4	práctica chromatin states (II)	3	EC	Seminario 1 (Decanato)
26/03/2025	9	16:00-19:00	X	4	Methylation	3	EC	Seminario 3 (Decanato)
02/04/2025	10	16:00-19:00	X	4	Estructura del DNA	3	EC	Seminario 1 (Decanato)
					BLOQUE 4: scRNAseq			
09/04/2025	11	16:00-19:00	X	5	scRNA-Seq	3	CTF	Seminario 1 (Decanato)
23/04/2025	13	16:00-20:00	X	5	scRNA-Seq	4	CTF	Seminario 1 (Decanato)
25/04/2025	13	16:00-19:00	V	5	scRNA-Seq	3	CTF	Seminario 1 (Decanato)
08/05/2025	15	16:00-20:00	X	5	scRNA-Seq	4	CTF	Seminario 1 (Decanato)
13/05/2025		16:00-20:00		5	PRESENTACIONES Bloque 3	4	EC	Seminario 1 (Decanato)



Rev Esp Cardiol. 2013;66:657-61

OMICS EXPERIMENT PIPELINE

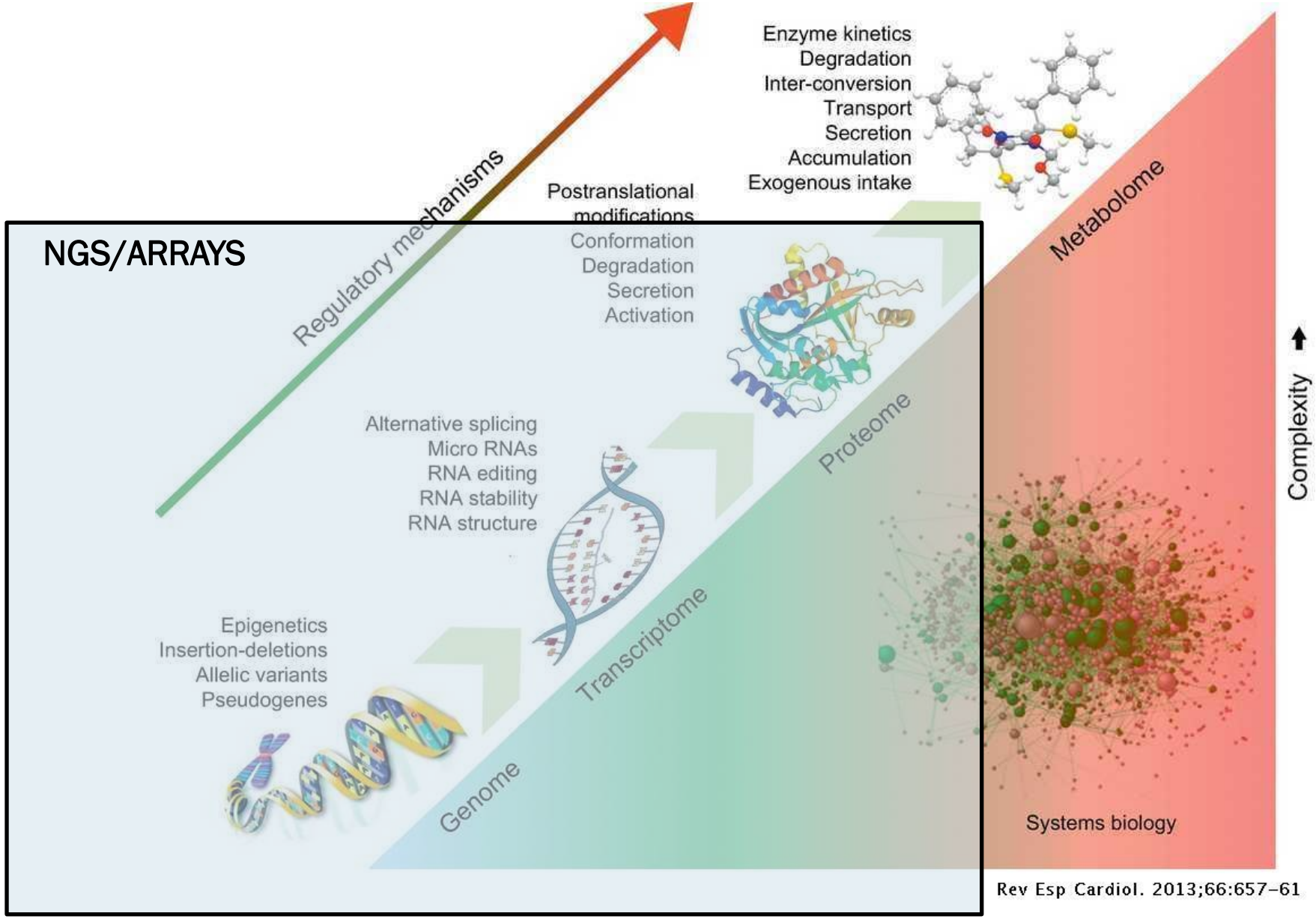


Analysis of DNA microarray data. Hackl H, Sanchez Cabo F, Sturn A, Wolkenhauer O, Trajanoski Z. Curr Top Med Chem. 2004;4(13):1357-70.

doi:10.2174/1568026043387773.

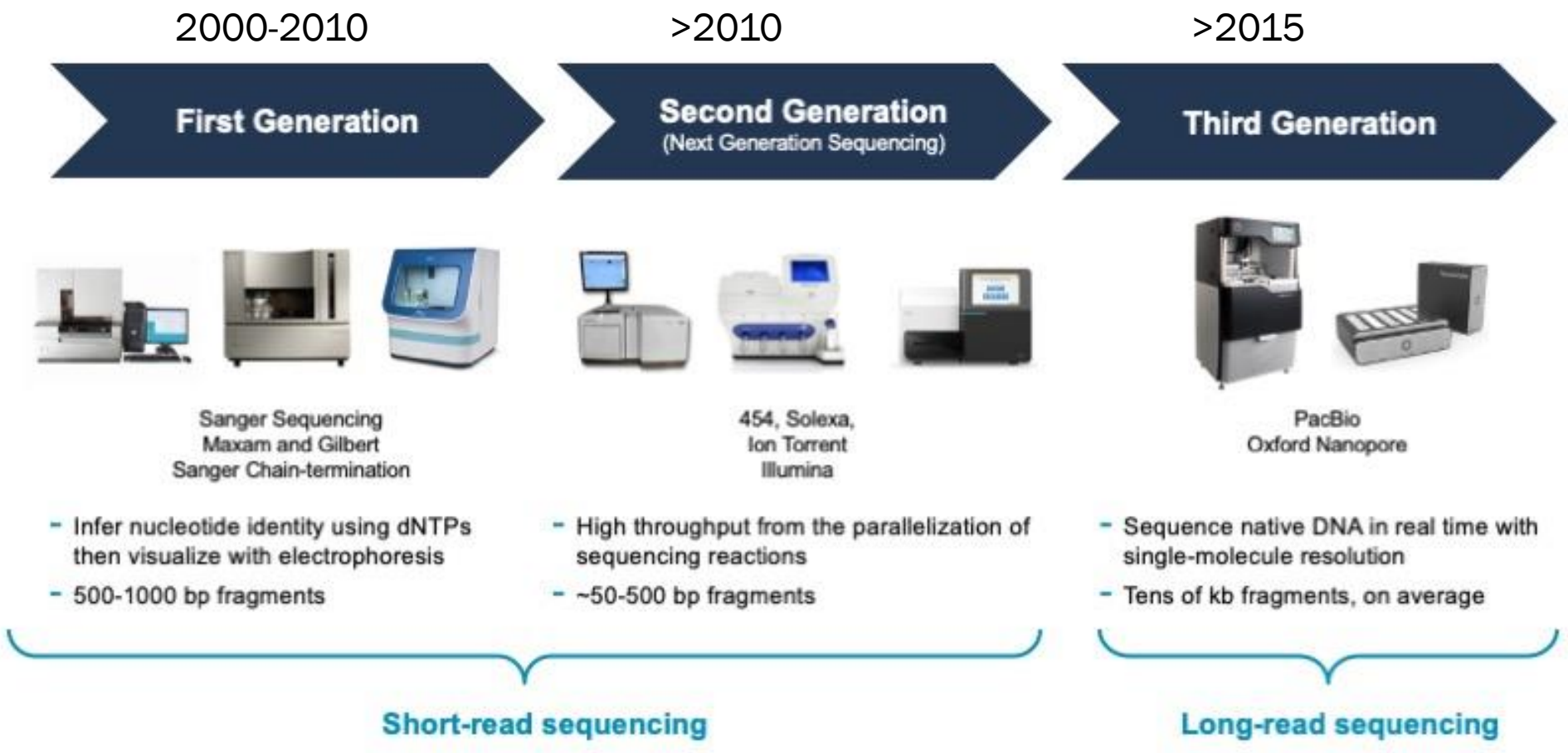
PREVIOUSLY IN TRREP…

- Experimental design in omics experiments
  - Replication:
    - Understand the difference between biological and technical replicates and its impact in estimates
    - What is pooling and when to use it
    - Why is three the magical number (in animal experiments)
  - Randomization
  - Blocking
    - Ways to control confounders
    - Batch effect
- Statistical peculiarities of omics data for quantification of expression:
  - Small sample size leads to a bad estimation of the variance per gene
    - Moderated t-tests “borrow” information from all genes to improve per-gene variability estimates
  - Large number of hypothesis tested simultaneously:
    - Need to correct the p-values…best using the FDR (probability to make X% errors from the total of detected trues)
  - Data is not normal
    - Large variability for poorly expressed genes -> variance estabilization normalization
    - Normalization per library size and gene length

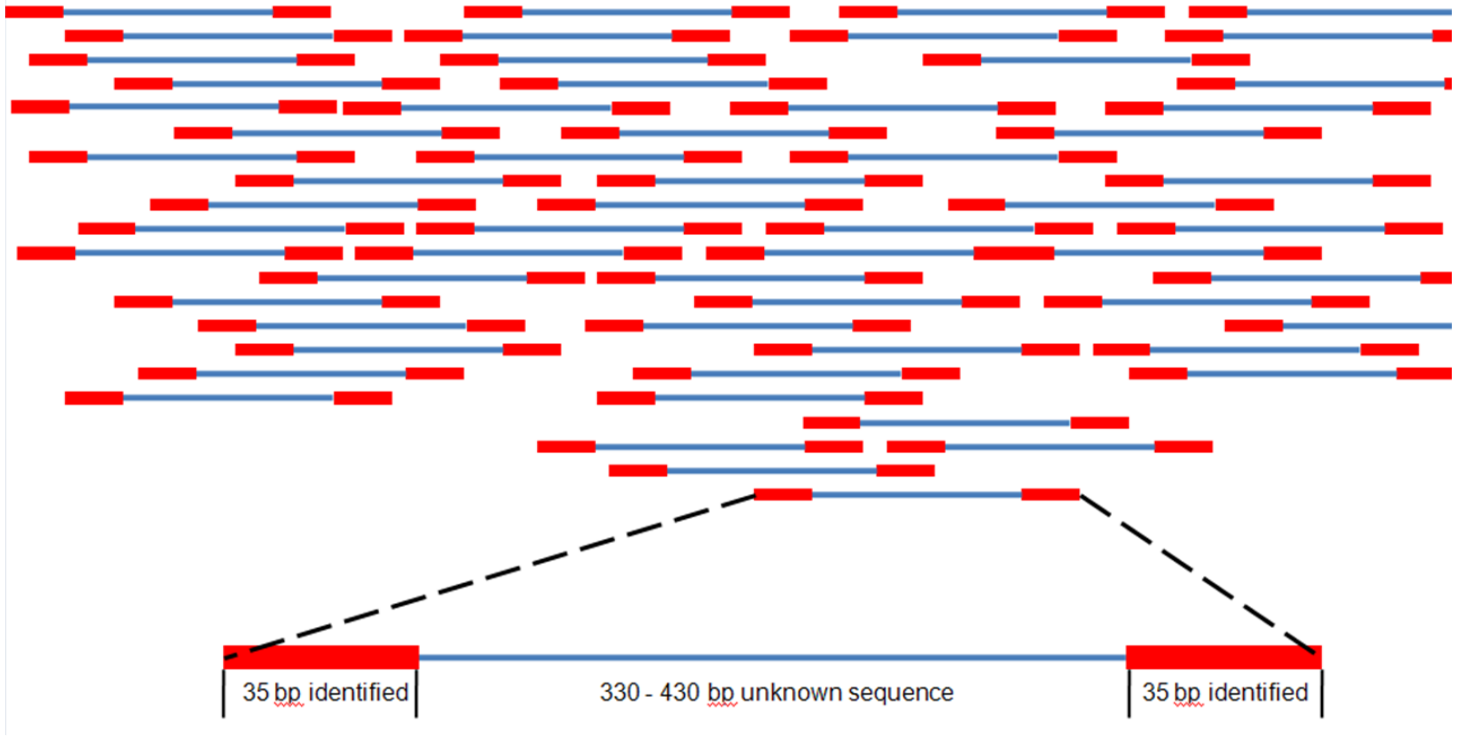


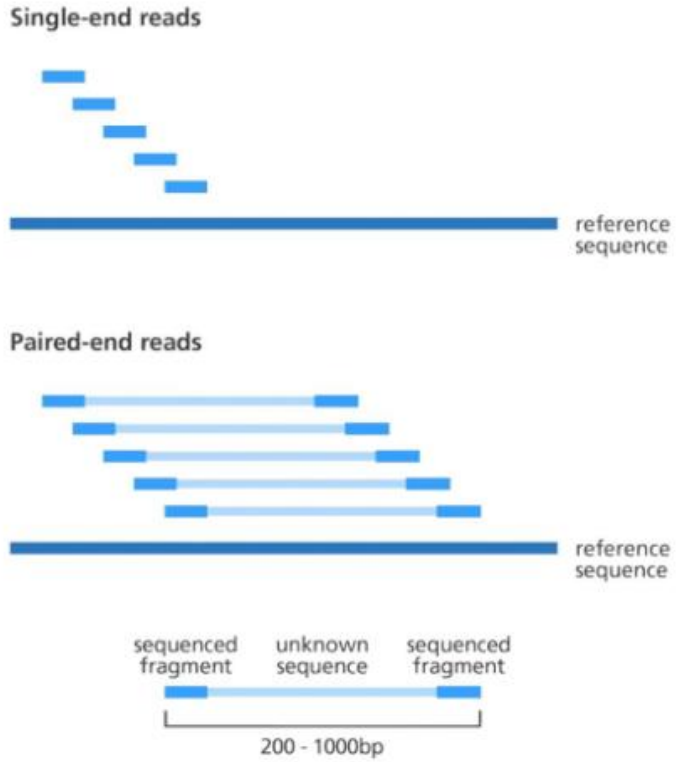
Rev Esp Cardiol. 2013;66:657-61

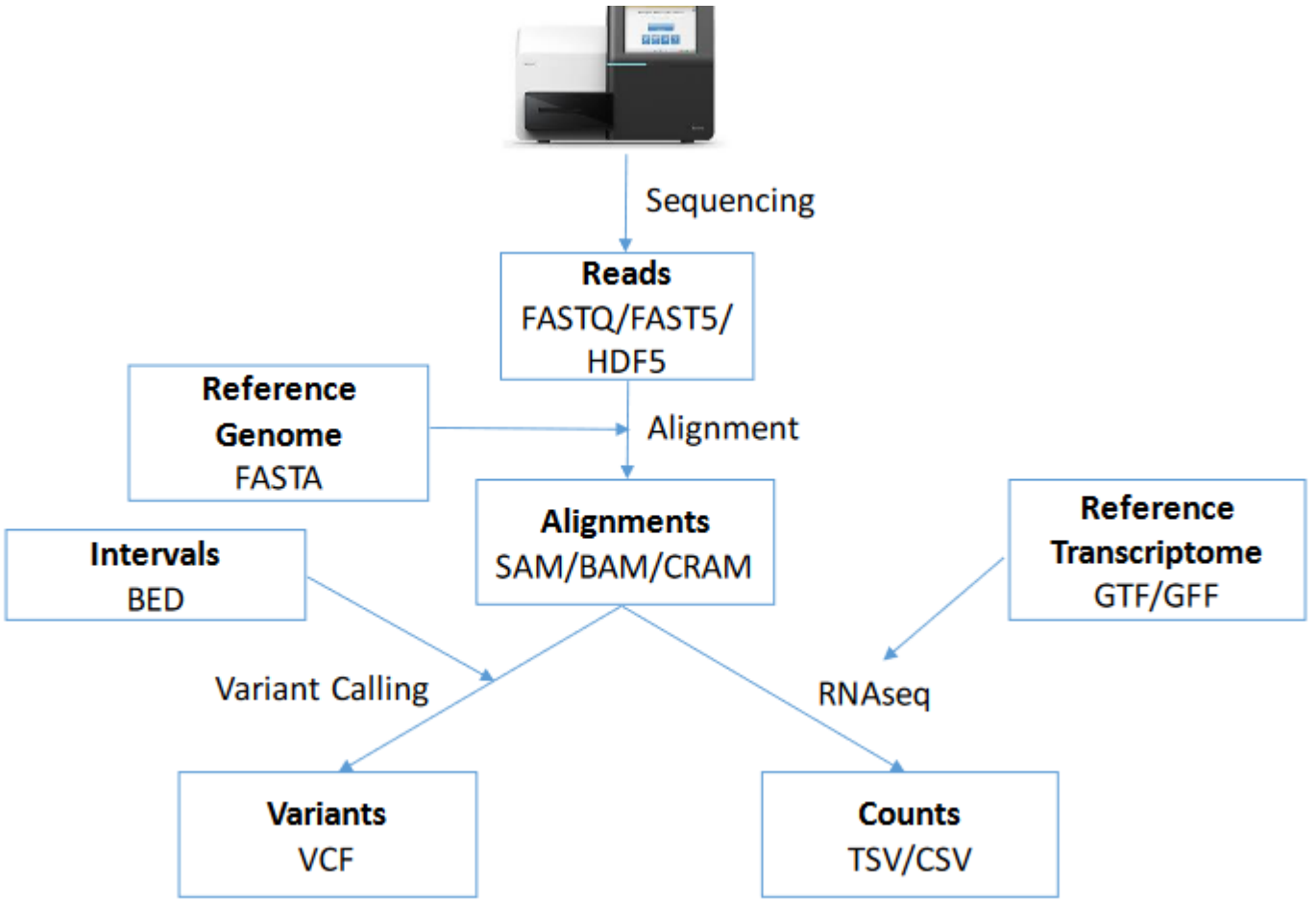
NEXT GENERATION SEQUENCING (NGS)



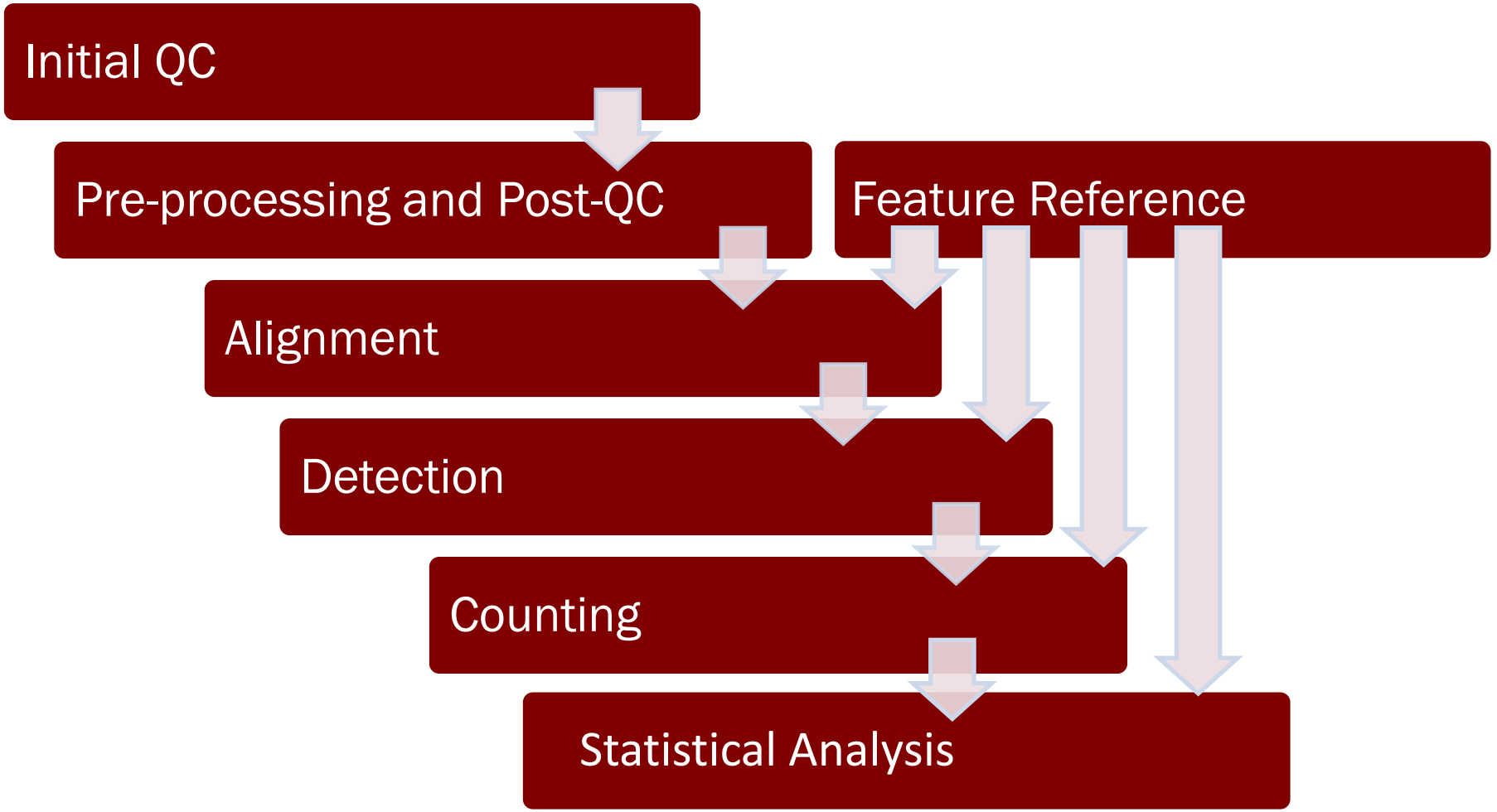
READS


35 bp identified
330 - 430 bp unknown sequence
35 bp identified


Single-end reads
reference sequence
Paired-end reads
reference sequence
sequenced fragment
unknown sequence
sequenced fragment
200 - 1000bp


Sequencing
Reads
FASTQ/FAST5/
HDF5
Reference
Genome
FASTA
Alignment
Alignments
SAM/BAM/CRAM
Intervals
BED
Reference
Transcriptome
GTF/GFF
Variant Calling
RNAseq
Variants
VCF
Counts
TSV/CSV

NGS ANALYSIS WORKFLOW

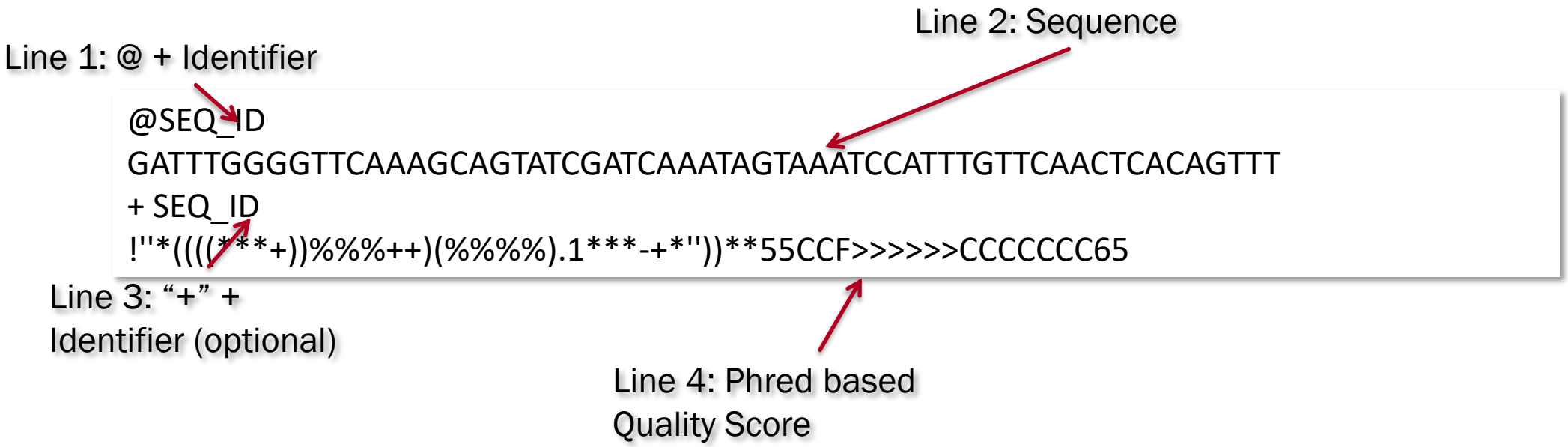


SEQUENCING RESULT: FASTQ FORMAT

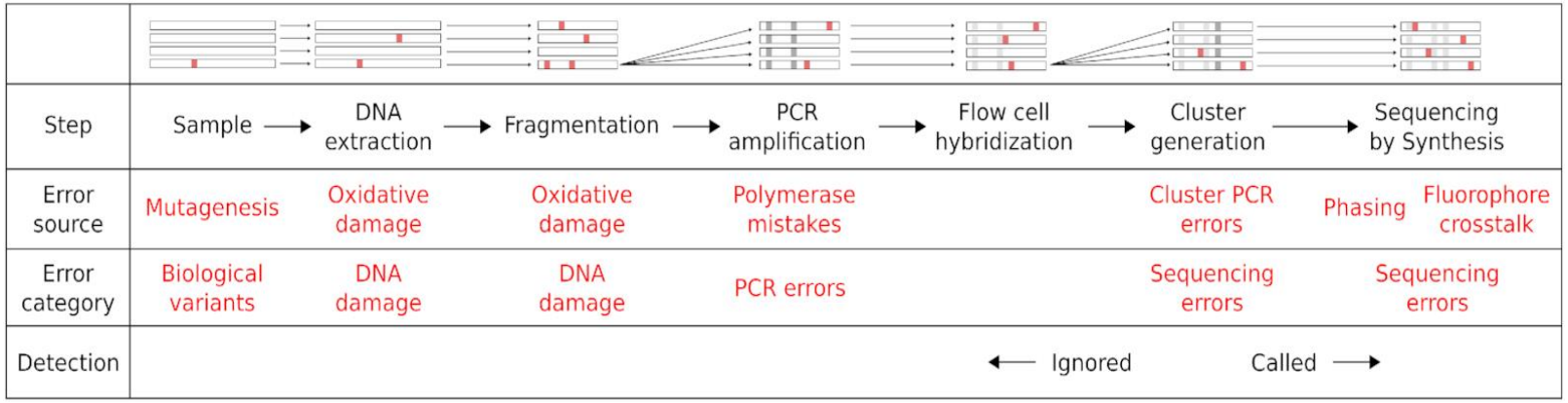
A file with Millions of reads

FASTQ file format: Millions of 4 line records

Record format:



SEQUENCING STEPS IN WHICH POLIMORPHISMS MIGHT ARISE



Step	Sample →	DNA extraction →	Fragmentation →	PCR amplification →	Flow cell hybridization →	Cluster generation →	Sequencing by Synthesis
Error source	Mutagenesis	Oxidative damage	Oxidative damage	Polymerase mistakes		Cluster PCR errors	Phasing / Fluorophore crosstalk
Error category	Biological variants	DNA damage	DNA damage	PCR errors		Sequencing errors	Sequencing errors
Detection					← Ignored	Called →	

FASTQ FORMAT

The Phred Quality Score

✓ Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P

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

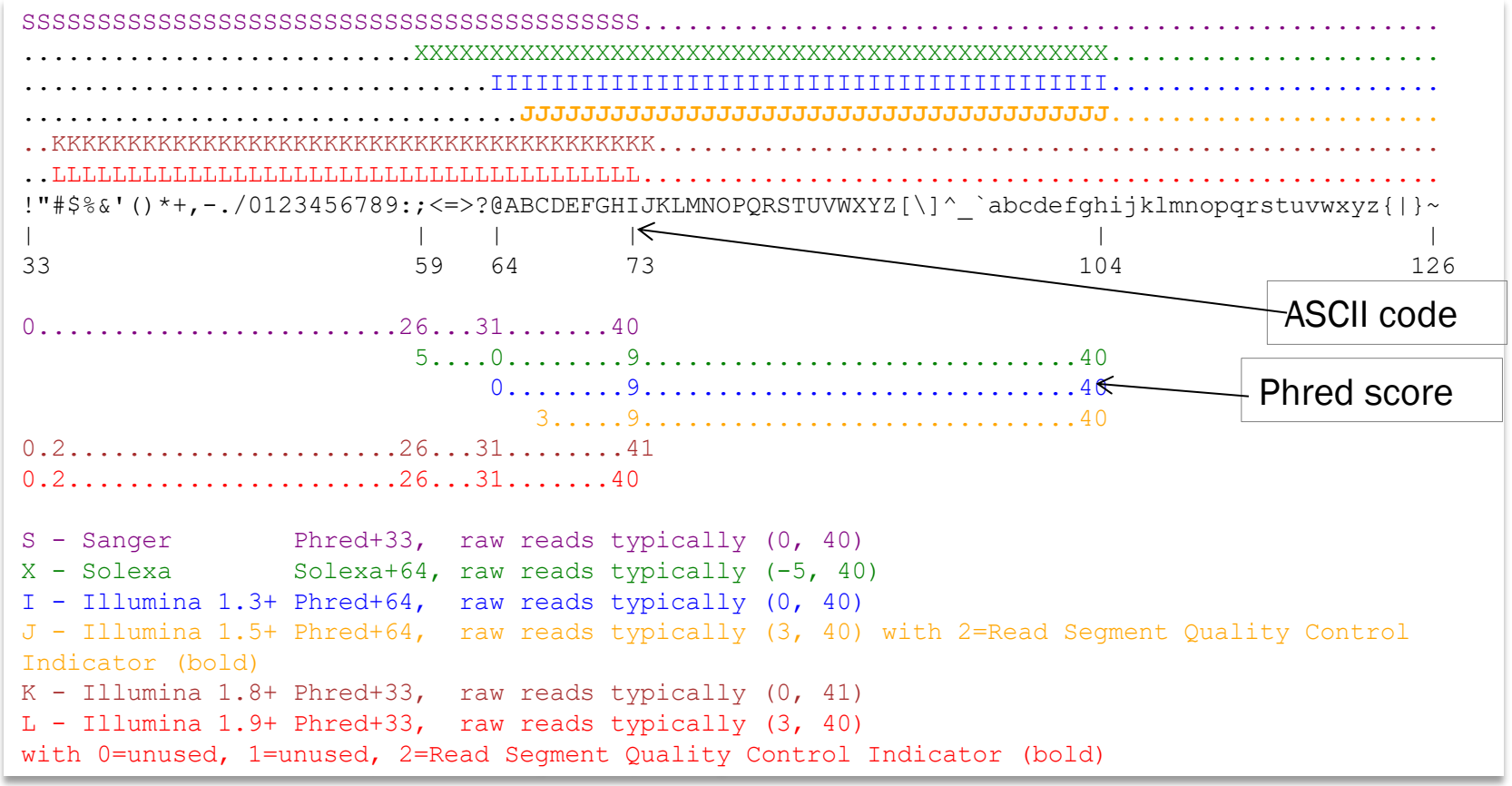
Q	*P*	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

$$P = 10^{-Q/10}$$

FASTQ FORMAT

The Phred Quality Score

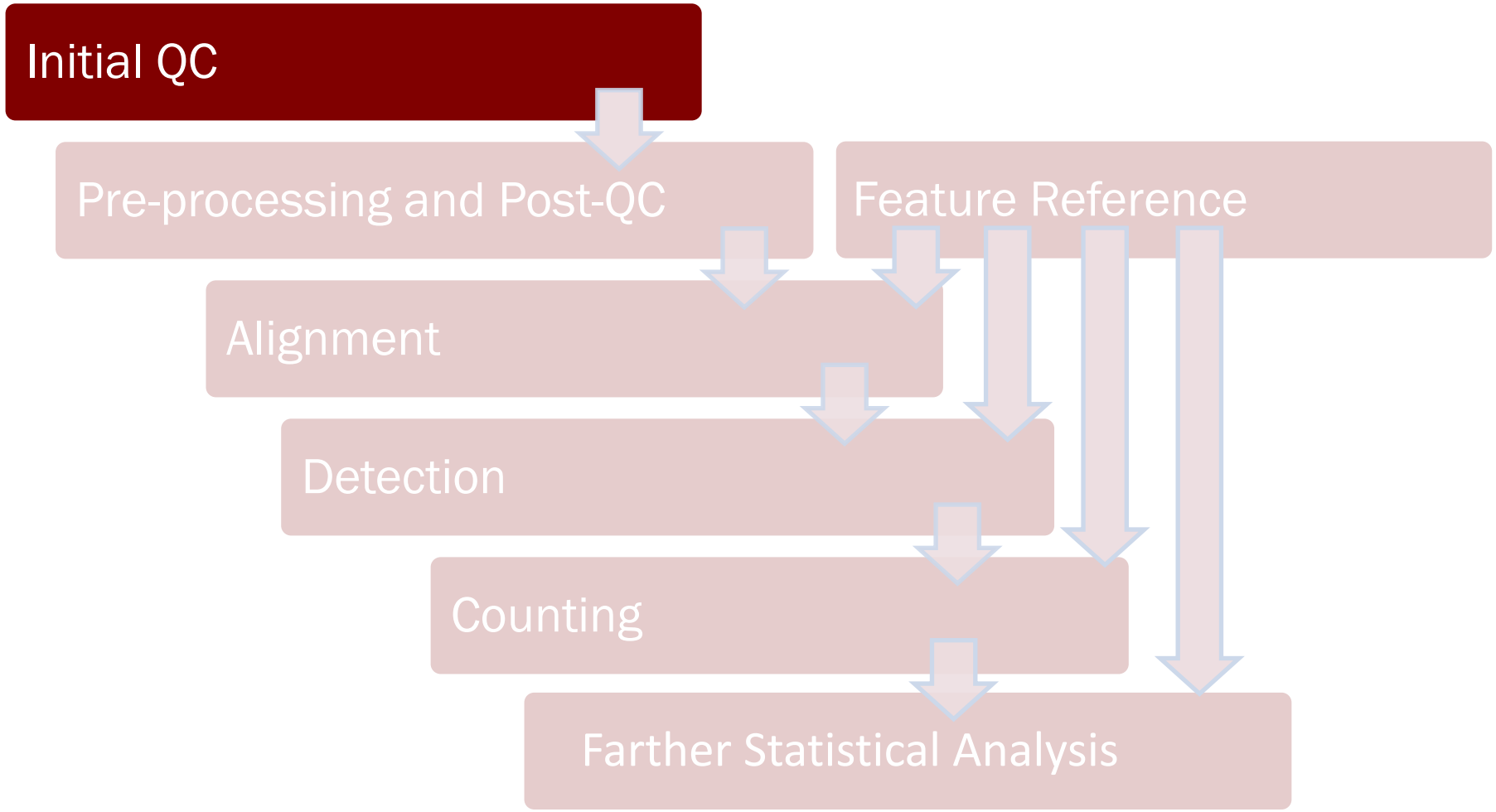
- ✓ The Quality Score is encoded as a letter using the Decimal ASCII code for the asignment



DECIMAL ASCII CODE

Code	Char	Code	Char	Code	Char	Code	Char	Code	Char	Code	Char
32	[space]	48	0	64	@	80	P	96	`	112	p
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(	56	8	72	H	88	X	104	h	120	x
41	)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[	107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	|
45	-	61	=	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	[backspace]

NGS ANALYSIS WORKFLOW



QC RAW READS

Goals:

- ✓Detect sequencing problems
- ✓Detect adaptors
- ✓Compare Libraries for posterior analysis

QC RAW READS

Questions to have in mind when assessing the quality of the experiment and sequencing

- How is the experiment design?
- Which are the DNA/RNA features I want to interrogate?
- How were this features enriched in our samples?
- Which kind of library has been done?

Different experiments require different interpretation of the QC analysis.

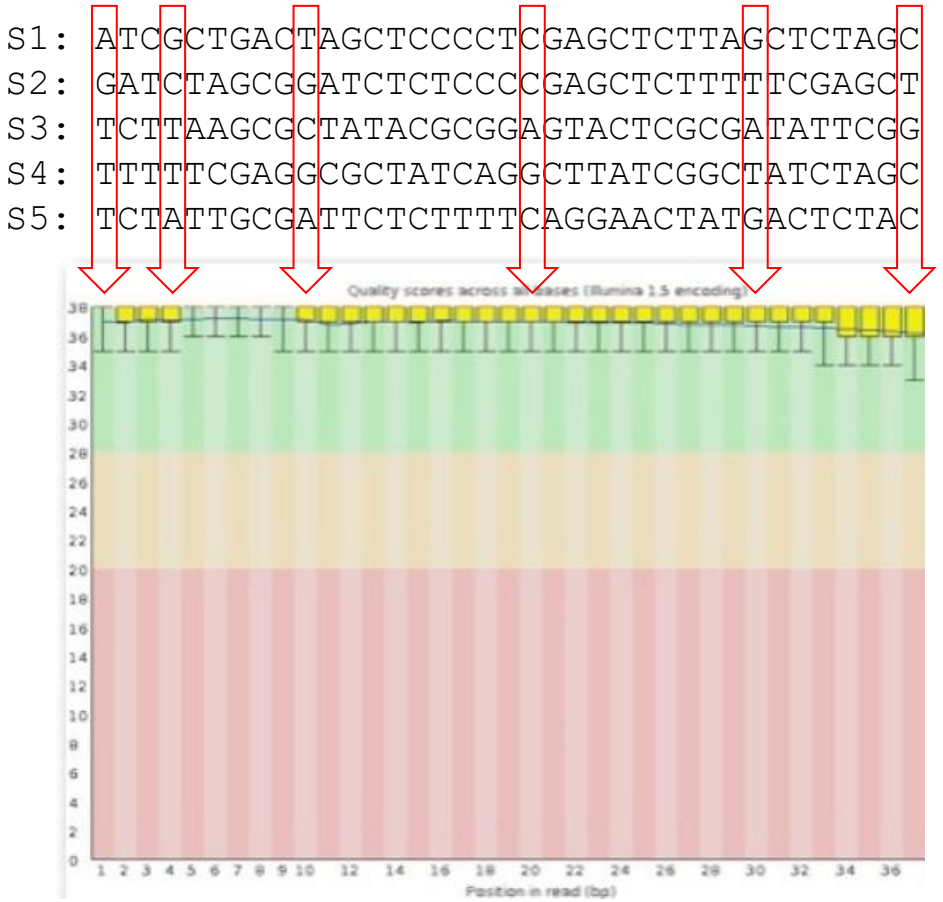
Tools used:

✓ FastQC: For evaluating quality of experiment.

QC RAW READS

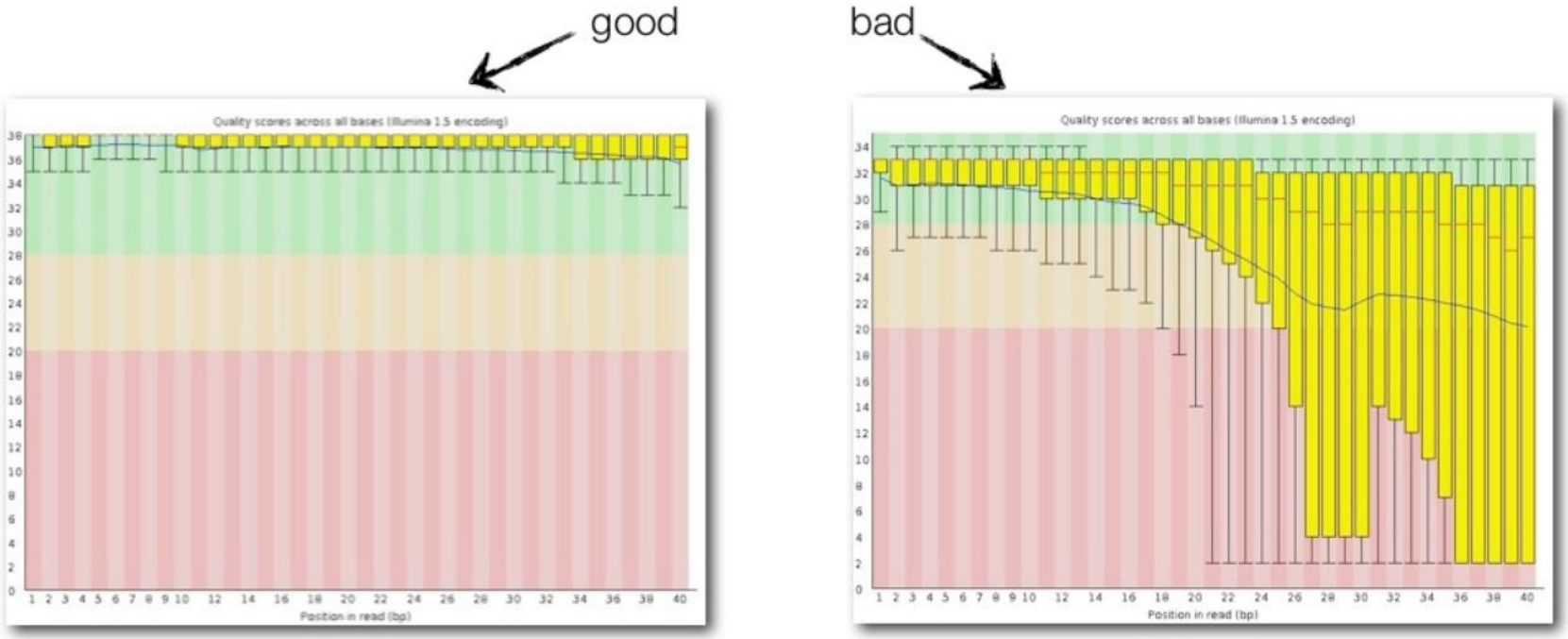
Per Base Quality Score

- Represents the distribution of quality scores across all reads at each position on the read.
- The distribution is represented as a boxplot.
- Background color indicates high quality values as green, good quality as pink and low quality as red



QC RAW READS ON RNA-SEQ

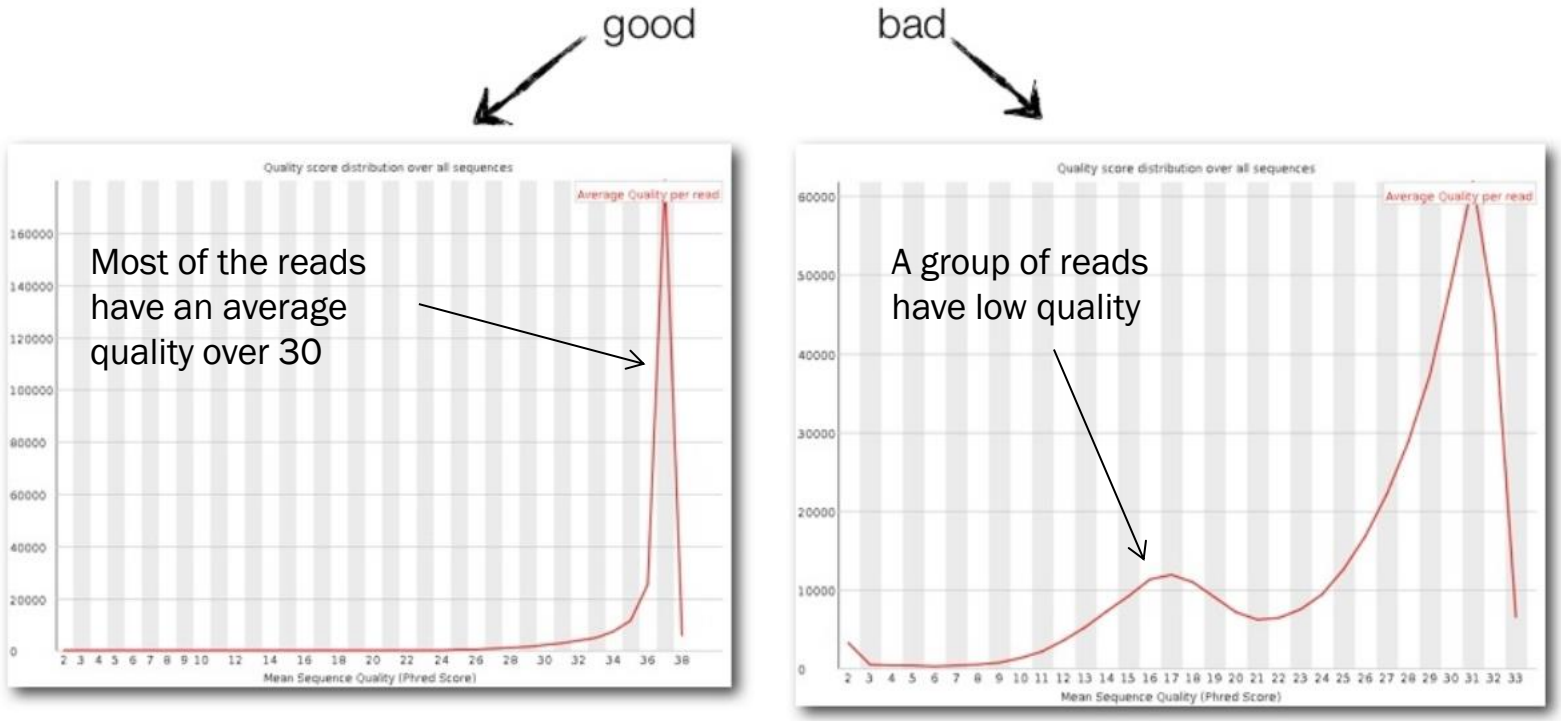
Per Base Quality Score



Depending on the experiment, low quality values at the end of the reads may be acceptable. In example for SNP calling low Q values are unacceptable. For RNA Seq they will be acceptable as long as they do not reduce the abitlity to map the read to the features.

QC RAW READS ON RNA-SEQ

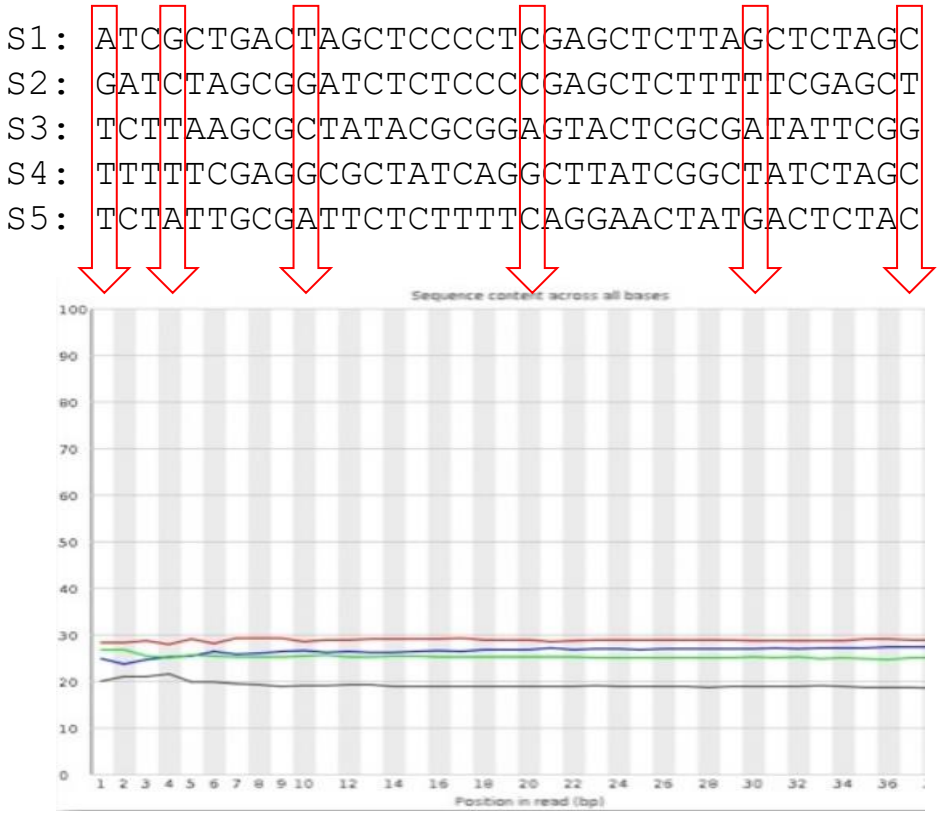
Per Sequence Mean Quality Score



QC RAW READS

Per Position Base Proportions

- Represents the distribution of frequencies of the four nucleotides across all reads at each position on the read.
- The frequency is represented as a line plot.
- Red: T, Blue: C, Green: A, Black: G

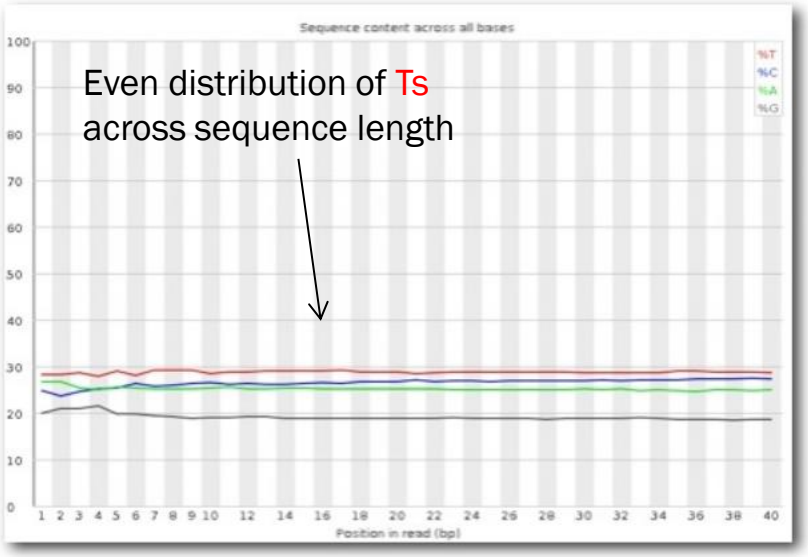


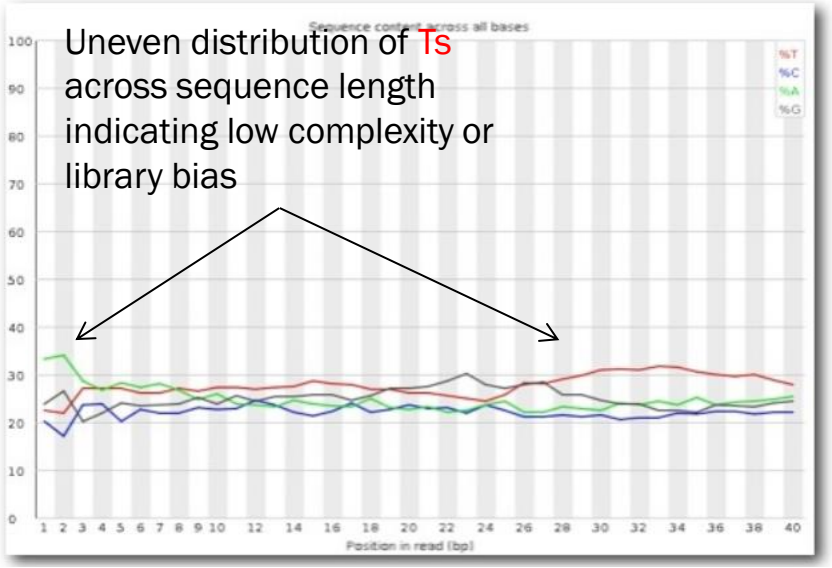
QC RAW READS ON RNA-SEQ

Per Position Base Content

good

bad



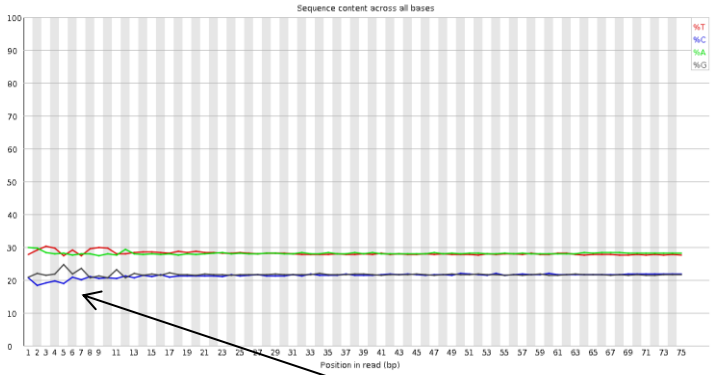


QC RAW READS

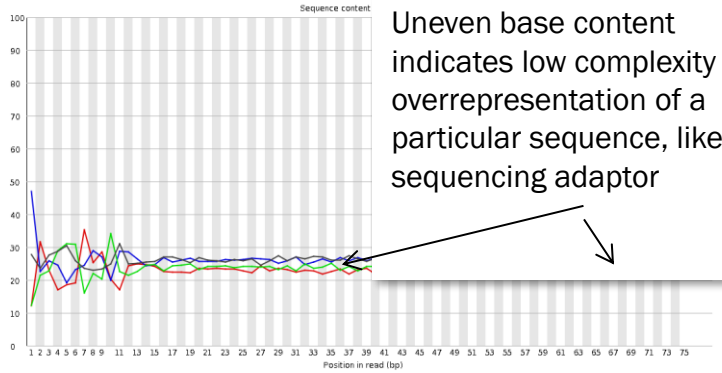
Per Position Base Content

Normal values in different experiments

RNASeq high quality base sequence content

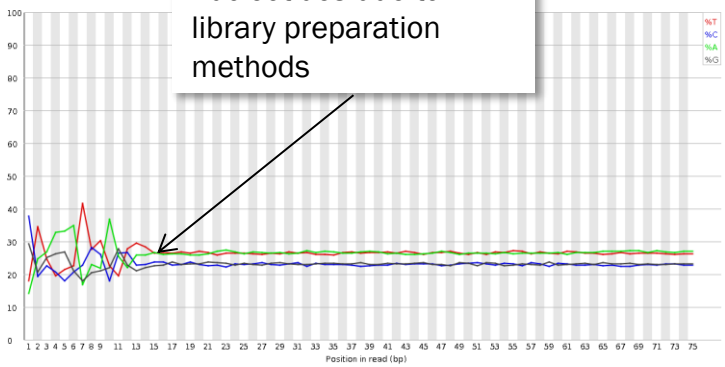


RNASeq low quality base sequence content

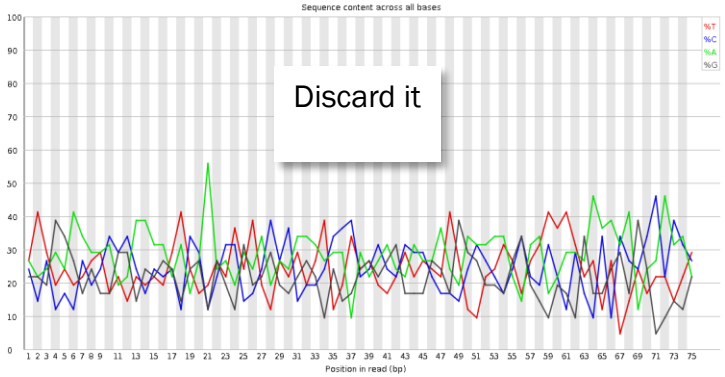


Uneven base content indicates low complexity or overrepresentation of a particular sequence, like the sequencing adaptor

Normal bias at first nucleotides due to library preparation methods



RNASeq very poor quality base sequence content



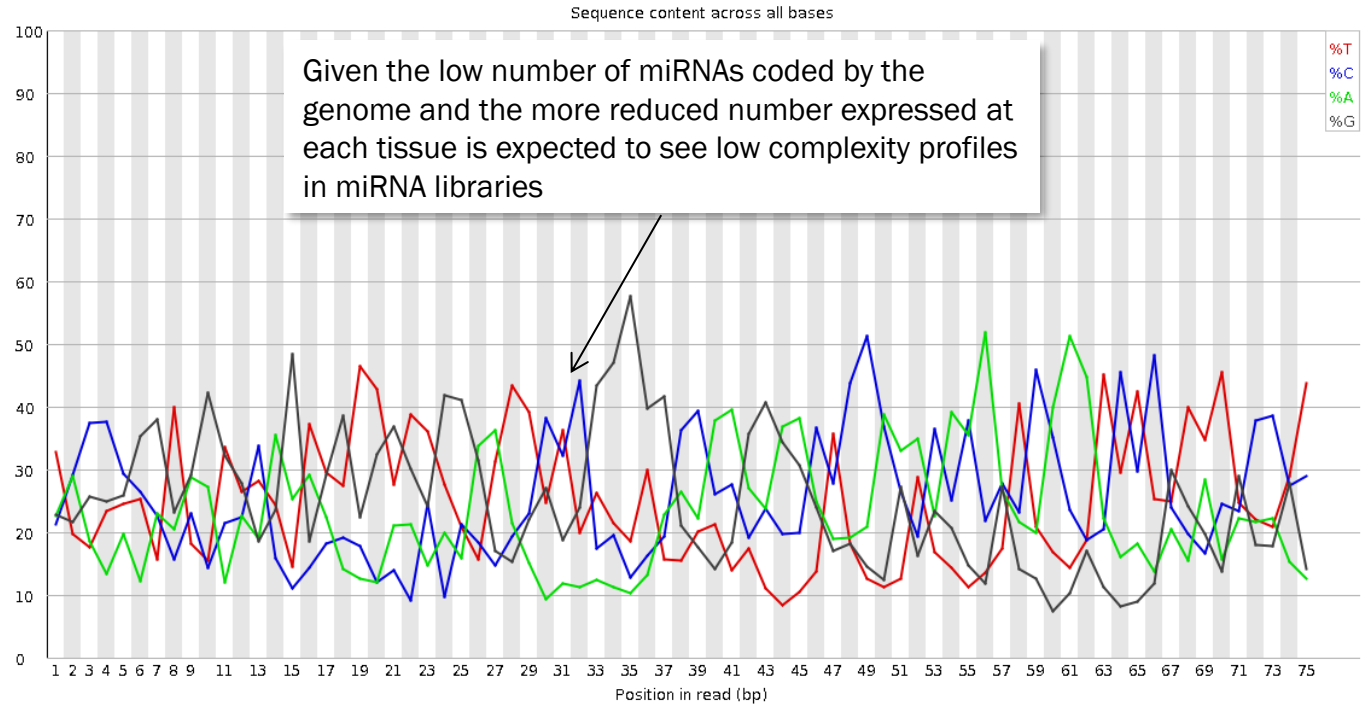
Discard it

QC RAW READS ON RNA-SEQ

Per Position Base Content

Normal values in different experiments

Good quality base content for a miRNA library



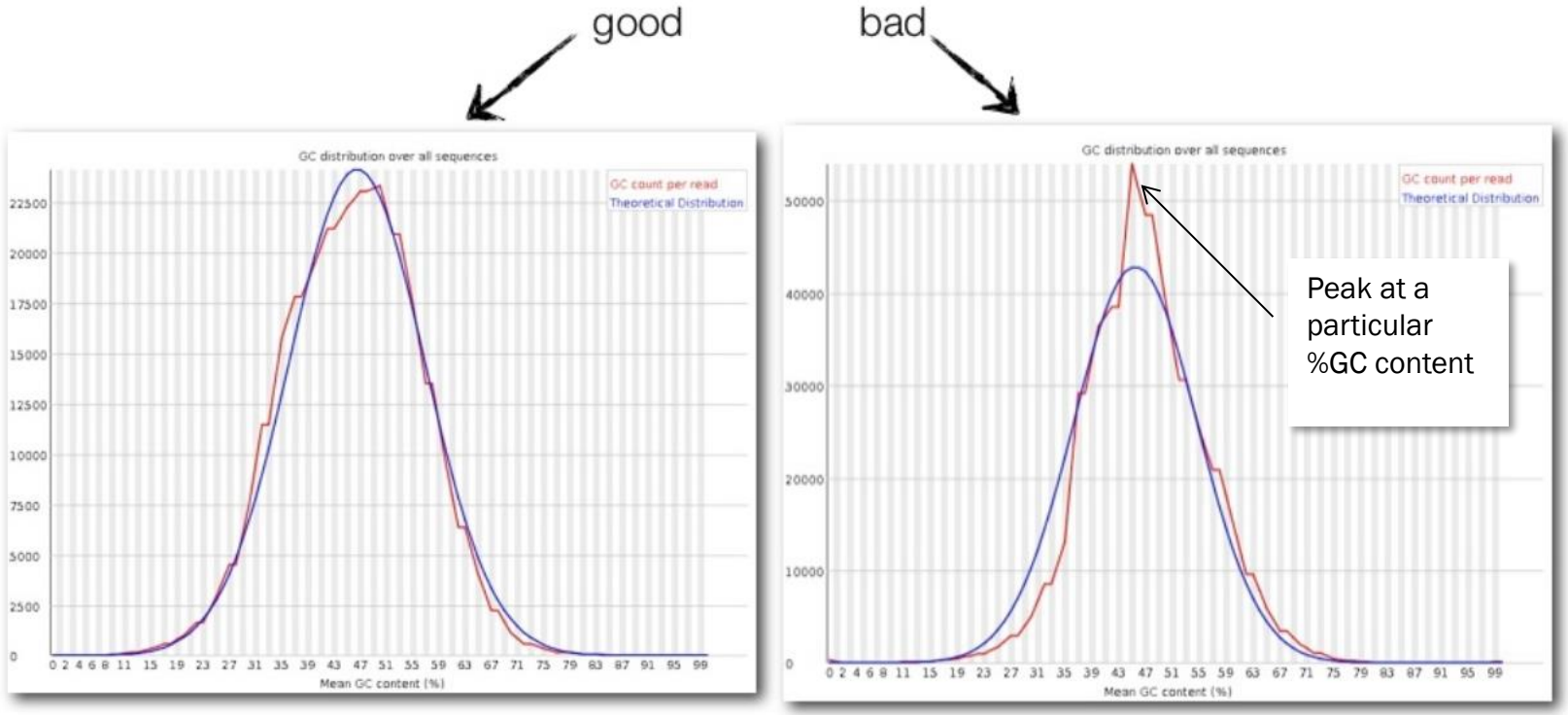
QC RAW READS

Per Sequence GC Content

S1: ATCGCTGACTAGCTCCCCTCGAGCTCTTAGCTCTAGC 47% Gs or Cs

S2: GATCTAGCGGATCTCTCCCCGAGCTCTTTTTCGAGCT 56% Gs or Cs

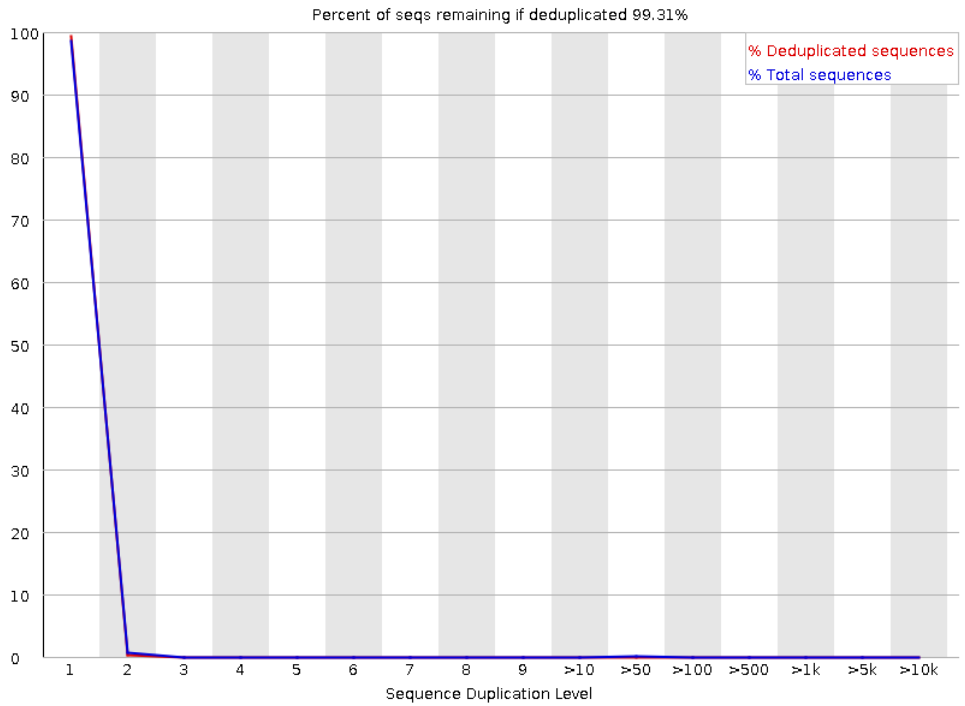
Peaks at certain %GCs may indicate the presence of a particular overrepresented sequence, like sequencing adaptor or a portion of a highly expressed transcript.



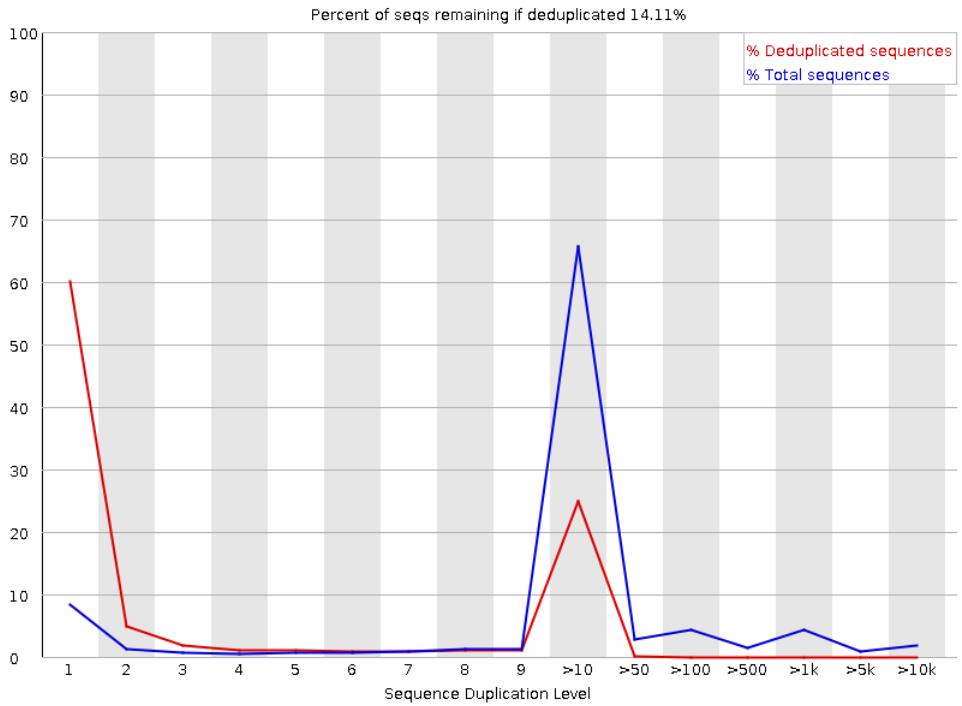
QC RAW READS ON RNA-SEQ

Sequence Duplication Levels

Low duplicated library



Highly duplicated library



QC RAW READS

Overrepresented Sequences:

- List of sequences that are present more times than expected by chance.
- The overrepresented list is annotated with the type of sequence if a list is provided to compare to.
- Typically, adaptor sequences may be found overrepresented if there are high levels of primer dimer ligations in the library prep step. Often due to an unbalance between levels of adaptors and levels of sample fragments.
- Some RNA-Seq from particular tissues may give also overrepresented sequences. For example blood samples contain high amounts of Haemoglobin transcripts which always is reported as overrepresented reads.
- miRNA samples allways report overrepresented sequences.
- Total RNA libraries report overrepresented sequences from ribosomal RNAs.

QC RAW READS

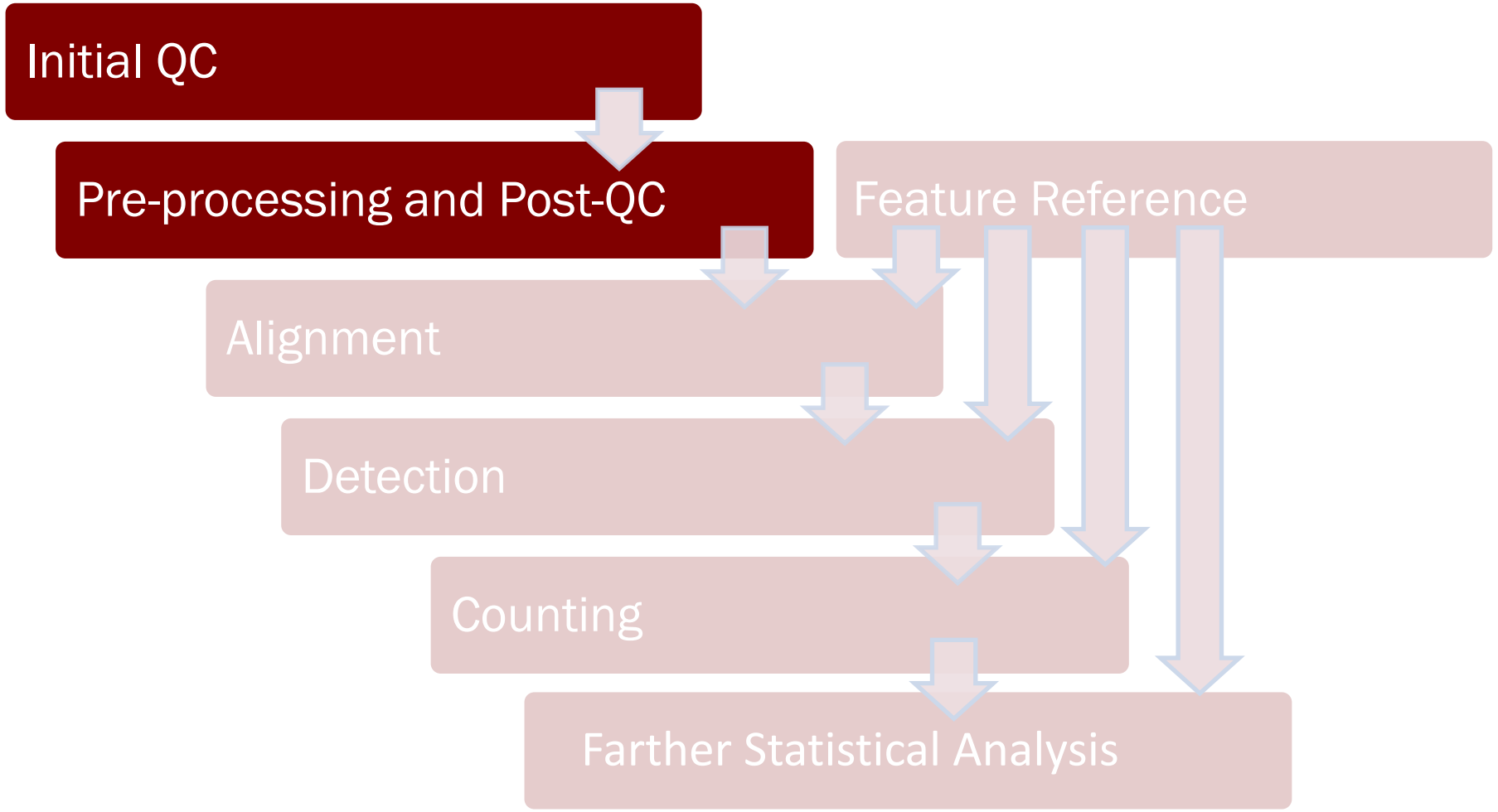
Overrepresented Sequences:

List of overrepresented sequences from a blood sample.

- The twelve first reads are coming from Haemoglobin transcripts.
- The last record is the Illumina True Seq adaptor.

Sequence	Count	Percentage	Possible Source
CAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGC	50648	0.3864742261873831	No Hit
GTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG	46849	0.357485606986509	No Hit
CACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGG	45022	0.34354451530975283	No Hit
CAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTC	43595	0.33265566045330447	No Hit
GCAGAATCCAGATGCTCAAGGCCCTTCATAATATCCCCCAGTTTAGTAGTTGGACTTAGGGAACAAAGGAACCTT	38630	0.29476977092123297	No Hit
AAGAAAGCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACCAGCCACCACTTTCTGATAGGCAGCCTG	37009	0.28240058120693534	No Hit
CACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACT	36309	0.2770591667713965	No Hit
CTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCG	34391	0.2624236912180203	No Hit
CTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTC	33986	0.25933330143745853	No Hit
CTTTAATAGAAATTGGACAGCAAGAAAGCGAGCTTAGTGATACTTGTGGGCCAGGGCATTAGCCACACCAGCCAC	29007	0.2213405836166763	No Hit
AGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCAC	28831	0.21999759941574082	No Hit
CCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG	25167	0.1920391101417207	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA	24532	0.18719368418948196	TruSeq Adapter, Index 1 (100% over 63bp)

NGS ANALYSIS WORKFLOW



PREPROCESSING READS

Goals:

- ✓Improve quality
- ✓Improve mapability
- ✓Remove contaminants
- ✓Remove biases
- ✓Remove non-informative segments

PREPROCESSING READS

Processing steps of fastq reads depends on the type of experiment and QCs obtained on raw data

Questions to have in mind when deciding which processing steps apply to the reads

- How is the experiment design?
- Which are the DNA/RNA features I want to interrogate?
- How were this features enriched in our samples?
- Which kind of library has been done?

There are many tools in the community developed to this end.

Most common tools used:

- ✓ fastx_trimmer
- ✓ fastq_quality_trimmer
- ✓ fastx_clipper
- ✓ Cutadapt

PREPROCESSING READS

What to do when quality of base calls is low

Considerations:

- Base quality can affect variant calling analysis and if severe, also mapping to features.
- Base quality usually drops at the end of the read and sometimes at the beginning.
- Also, low quality sequencing may produce group of reads that are low quality along it's length.
- It is essential to remove low quality bases for Variant Calling analysis.
- For other analysis only remove it if it affects mapping performance.

Actions to improve analysis:

- Trim those segments of the reads with low quality.
- Discard reads with low mean quality value.

Possible tools:

- fastq_quality_trimmer
- Cutadapt
- …

PREPROCESSING READS

What to do when Adaptors are overrepresented

Considerations:

- When adaptors are overrrepresented means the fragment size of the library is small on average.
- For miRNAs that's normal as we are enriching in sequences of 22nts but for RNASeq of DNASeq is not ideal.
- The presence of adaptor in the sequence will avoid the alignment of the read as no sequence in the reference will have adaptor.
- It is essential to remove the adaptor sequences to increase mappability.

Actions to improve analysis:

- Trim those segments with adaptor sequence.
- Discard reads smaller than 30nts except in the case of small-RNASeq.

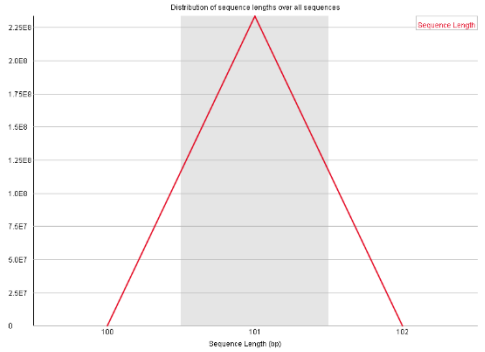
Possible tools:

- clip
- Cutadapt or TrimGalore
- fastx_clipper

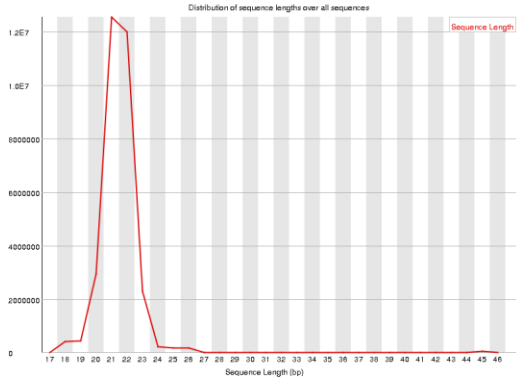
QC RAW READS ON RNA-SEQ

Sequence Length

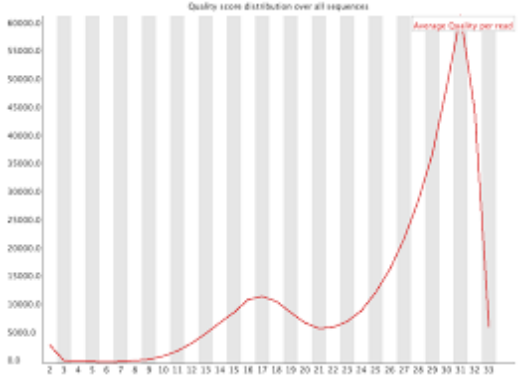
Before Processing



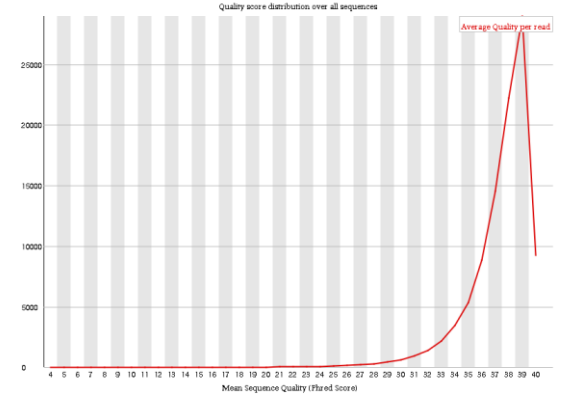
Good Quality miRNASeq library



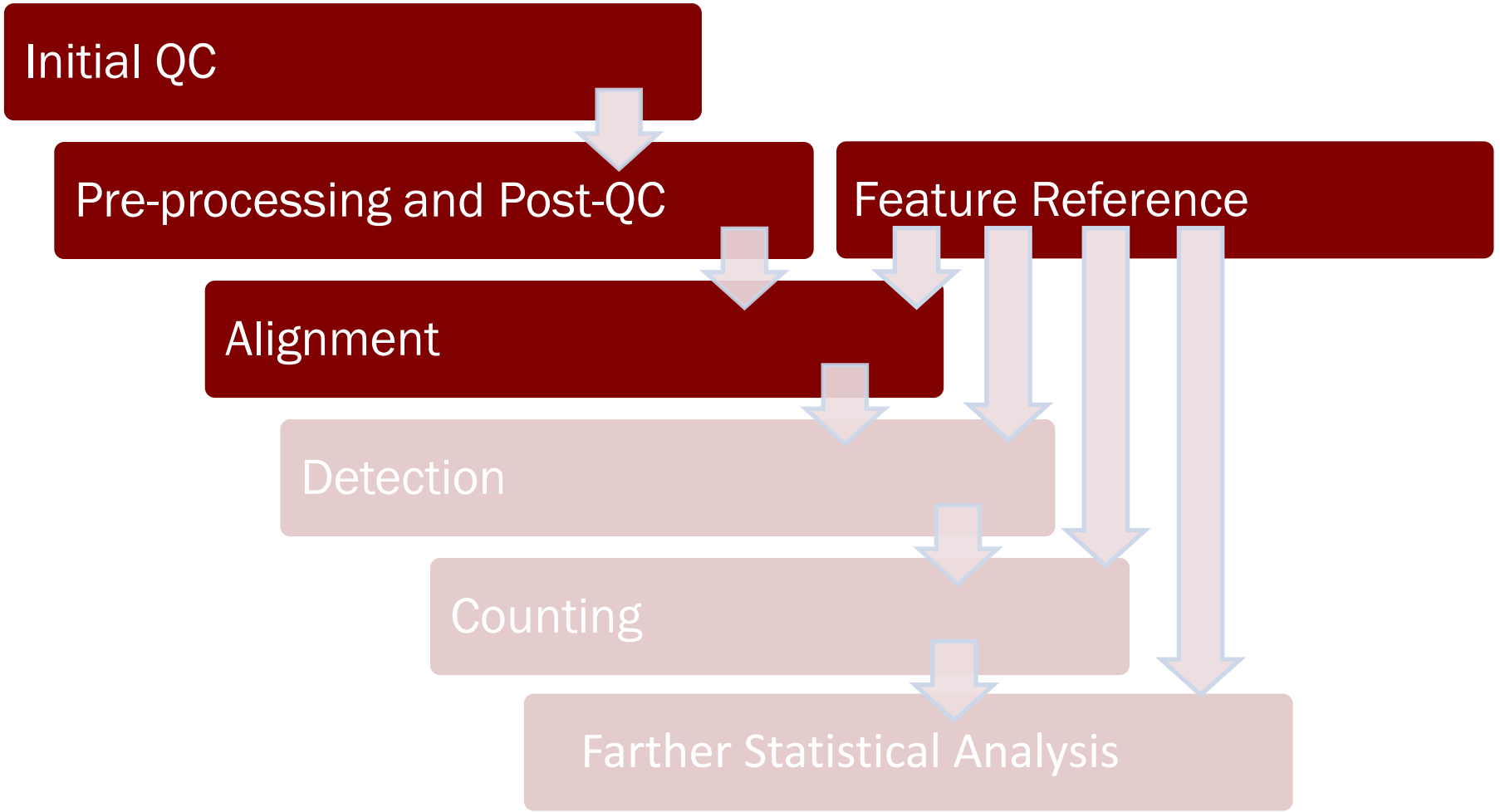
Medium quality



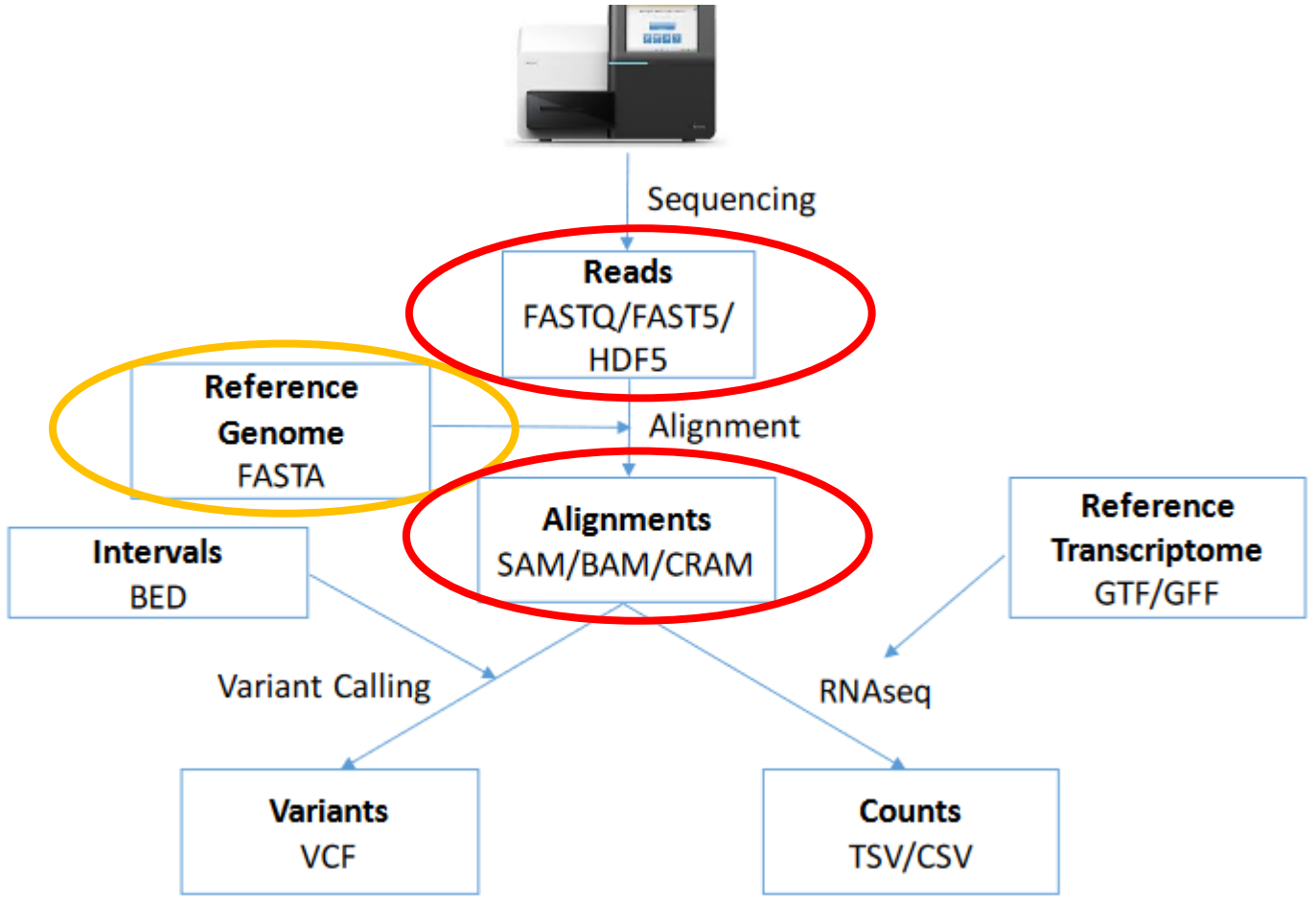
Good Quality RNASeq library



NGS ANALYSIS WORKFLOW



DATA FORMATS



MAPPING VS ALIGNMENT

- **Alignment**:
  - for each locus to which a sequence can be mapped, determine the optimal base by base alignment of the query sequence to the reference sequence

- **Mapping**:
  - (quickly) find the best possible loci to which a sequence could be aligned

TRADITIONAL ALIGNERS

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
```

Query Sequence

```
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

COVERAGE

Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

Definition: Coverage **C = N L / G**

How much coverage is enough?

Lander-Waterman model: **Prob[not covered bp] = e^{-C}**
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

MAPPING READS

Requirements for aligners:

- ✓ A reference in fasta format
- ✓ Reads in fastq/fasta
- ✓ An indexed reference

Goals:

- ✓ Map reads **to features** by alignment
- ✓ Identify unambiguously which feature produced it.
- ✓ Work efficiently

MAPPING READS

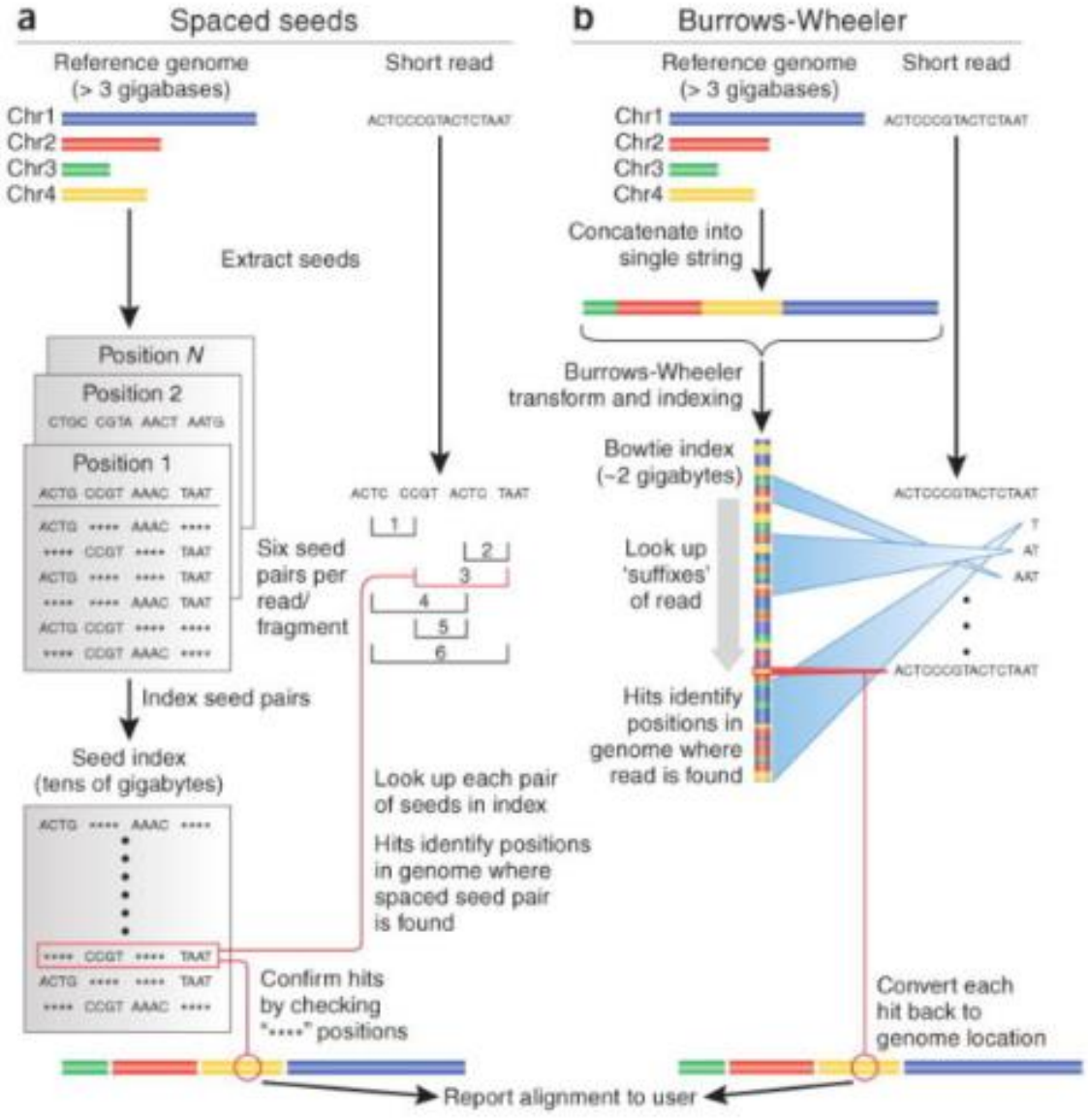
The Indexed Reference:

- ✓ The indexed reference allows the new aligners to map efficient and accurately millions of reads.
- ✓ BWA and Bowtie aligners use the Borrows Wheeler Transformation method or versions of it to index and compress the reference.
- ✓ All aligners provide tools to index the reference specifically for the aligner from a FASTA file.

MAPPING/ALIGNMENT ALGORITHMS FOR NGS

- Indexing
  - Organize genome information to be more accessible for algorithms
  - One time per genome
- Hash tables
  - Space seeds based: Maq
  - NeedIman-Wunsch: Novoalign
- Burrows-Wheeler transformation based
  - Bowtie2
  - BWA

http://web.stanford.edu/class/cs262/presentations/lecture4.pdf



BWA

ANA

X = BANANA

BANANA
ANANA
NANA
ANA
NA
A

suffixes of BANANA

ANA

X = BANANA$

BANANA$
ANANA$
NANA$
ANA$
NA$
A$
$

ANA

X = BANANA$

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

ROTACIONES

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

SORT ALPHABETICALLY

ANA

X = BANANA$

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

ANA

X = BANANA$

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

ANA

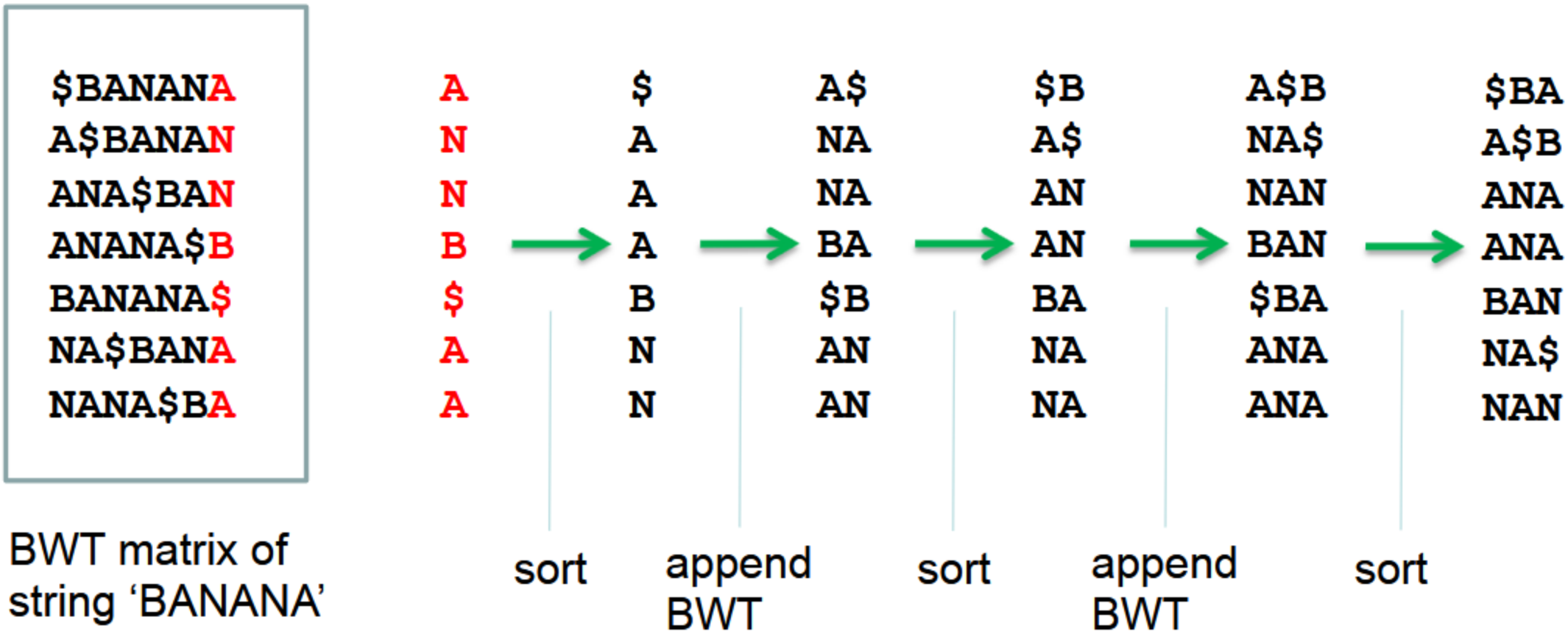
X = BANANA$

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

BWT matrix of string 'BANANA'

BWT(BANANA) = ANNB$AA



BWT matrix of string 'BANANA'

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

BWT matrix of string ‘BANANA’

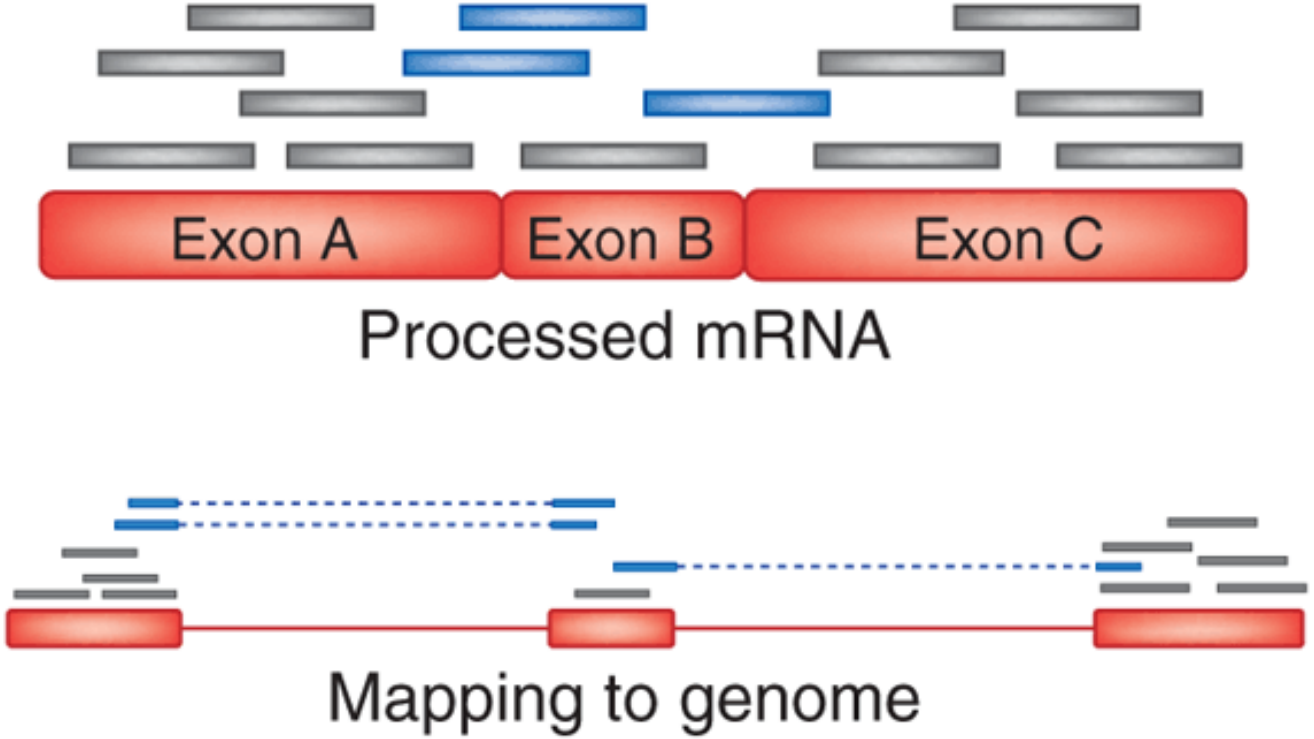
Lemma. The i-th occurrence of character c in last column is the same text character as the i-th occurrence of c in the first column

A$BANAN
NA$BANA
NANA$BA
BANANA$
$BANANA
ANA$BAN
ANANA$B

A$BANAN
ANA$BAN
ANANA$B

Same words, same sorted order

MAPPING READS: THE RNA-SEQ ALIGNMENT PROBLEM



MAPPING READS: THE RNA-SEQ ALIGNMENT PROBLEM

- Alignment against **TRANSCRIPTOME** (RSEM,BWA,Bowtie):
  - The reference is a list of ~100K transcripts of average length around 2-3kb.

- Problems:
  - Transcriptome is not always available
  - Transcript variance is not fully known even for humans.

MAPPING READS: THE RNA-SEQ ALIGNMENT PROBLEM

- Alignment against **GENOME** and **TRANSCRIPTOME**:
  - The reference is the genome. 24 chromosomes of ~100Mb each on average.
  - Transcriptome can be used if available.
  - Try to infer transcript structure from alignments

- Problems:
  - Assembly of transcriptome from reads is very complex due to the use of shared exons.
  - Detection of small exons is difficult.

MAPPING READS

Main tools for Deep Sequecing Read Aligments:

- ✓ For DNA
  - BWA/BWA-MEM (http://bio-bwa.sourceforge.net/)
  - Bowtie/Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/)
- ✓ For RNA
  - HISAT2(based on Bowtie2)
  - RSEM (based on Bowtie1, 2 or STAR)
  - STAR
  - KALISTO
  - SALMON

MAPPING REPORT

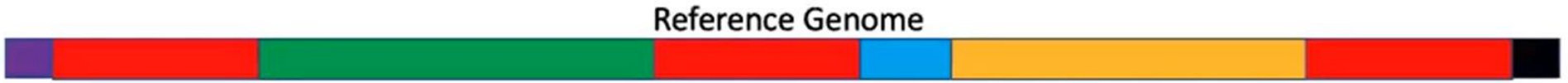
- A read can may be mapped to 0, 1 or more locations.
  - Unmapped
  - Uniquely mapped
  - Multi mapped reads

- When several locations are found, they can be subdivided in primary (best scoring) and secondary (lower and still good scores) alignment

- We can choose to report
  - All alignments
  - All best alignments
  - Random alignment of the best ones
  - All above a threshold

MAPPING REDUNDANCIES

The human genome is large and contains many repetitive sequences.



Read

chr10:1,020 ATGAGAGAGTATCTCGACTCTAGGCCGATACCA

read AGTAACTCGACTCTA mismatches: 1 MQ = 10

Mapping Quality

chr2:2,139 CGATCGAAGTAACTCGACGTTAGGCCGATACCA

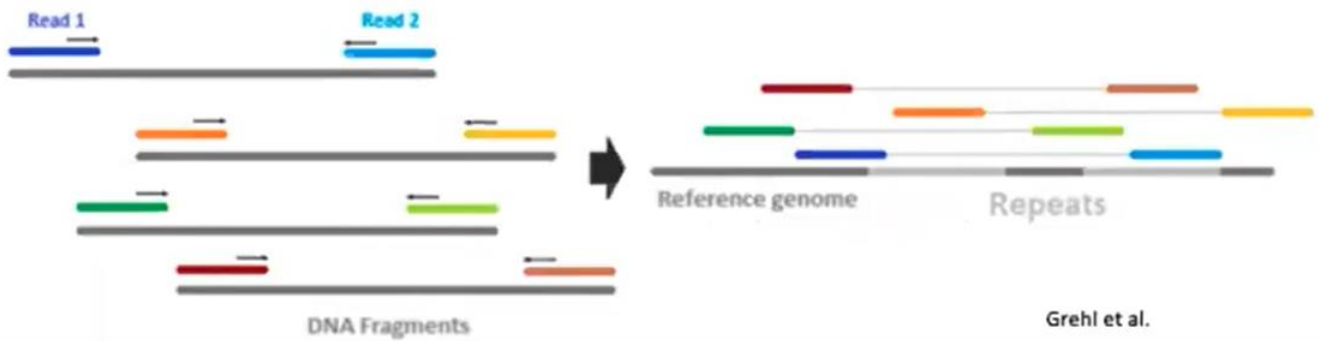
read AGTAACTCGACTCTA mismatches: 2 MQ = 1

MQ=0 if the read maps with the same probability to several places in the genome ->
the aligner assigns the read to one of them randomly

Reducing Ambiguous Mapping

Single-end sequencing → The DNA library is read only from one side.

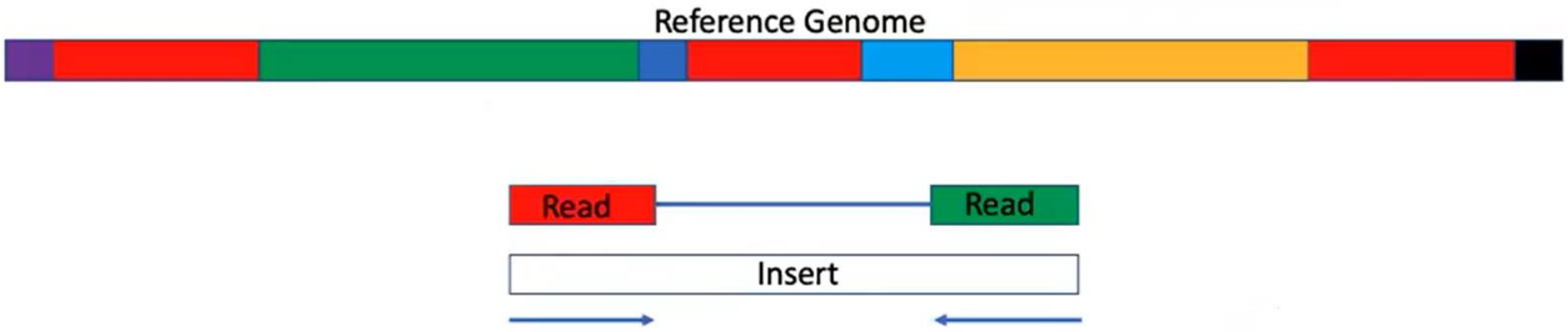
Paired-end sequencing → The DNA library is read from both sides.



Grehl et al.

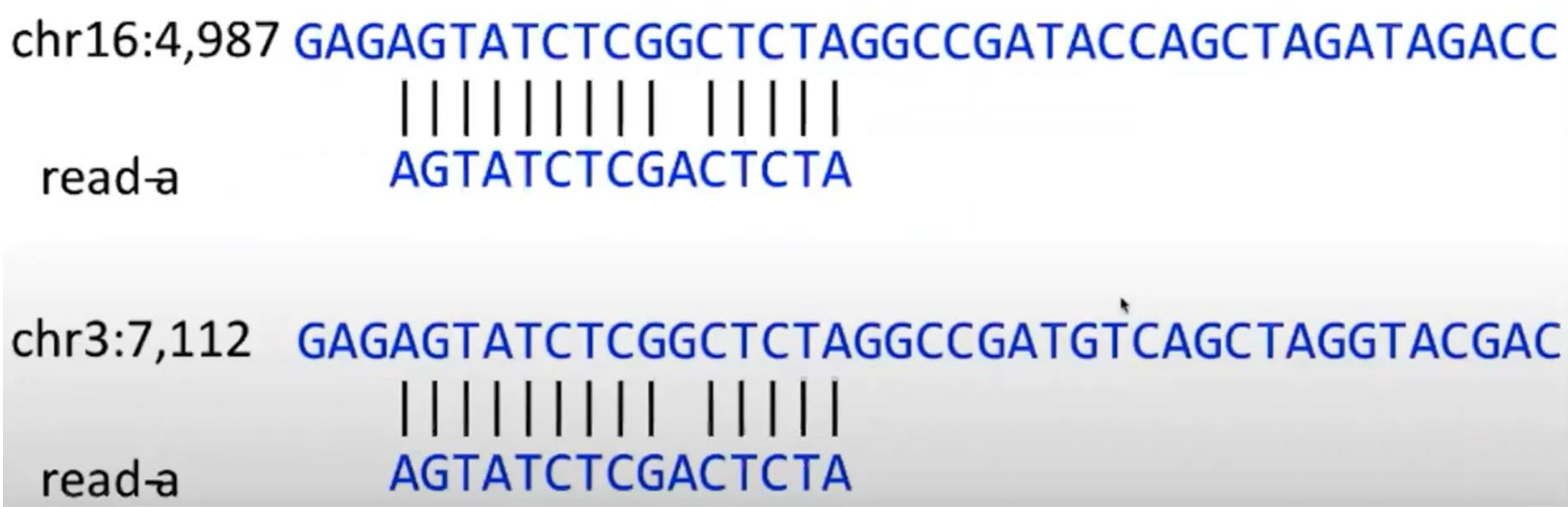
Reducing Ambiguous Mapping

Paired-end sequencing → The DNA library is read from both sides.



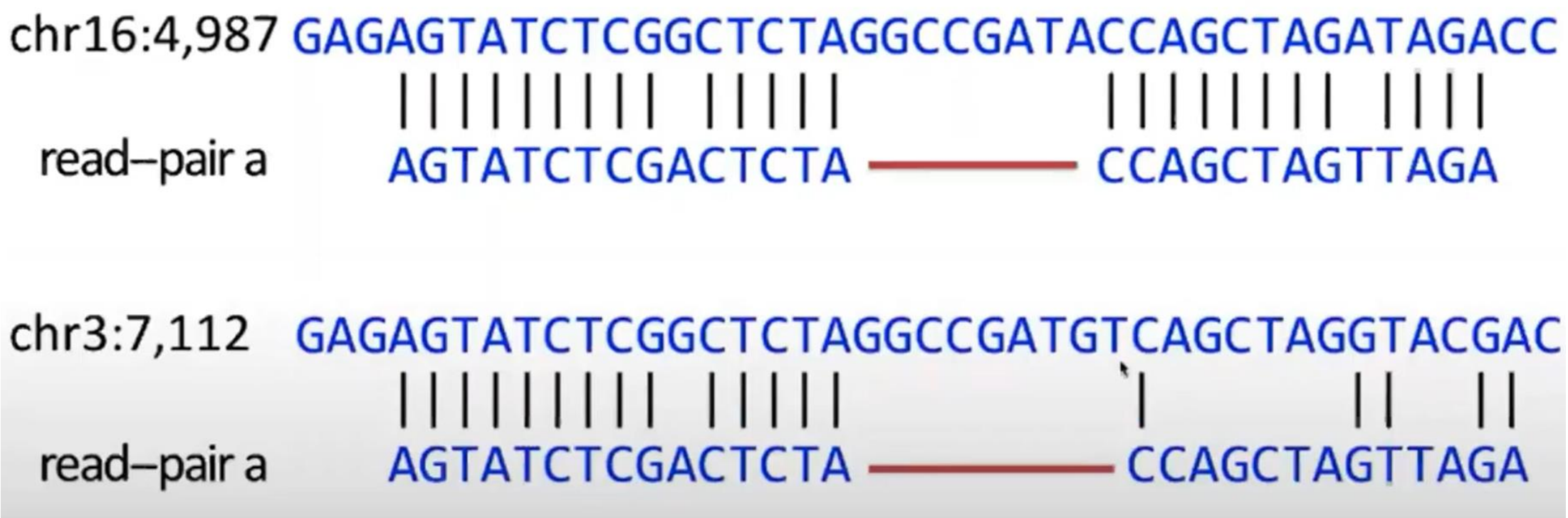
Reducing Ambiguous Mapping

Paired-end sequencing → The DNA library is read from both sides.



Reducing Ambiguous Mapping

Paired-end sequencing → The DNA library is read from both sides.



SAM/BAM FORMAT

Format designed to store information about alignments of NGS sequencing data

The file starts with a header followed by alignment records.

Header

Provides information about the reference and the tool used to map the reads

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
@PG     ID:GATK TableRecalibration      VN:1.0.3471     CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCova
```

http://samtools.github.io/hts-specs/SAMv1.pdf

SAM/BAM FORMAT

Alignment Record

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	LEN	SEQ	
1:497:R:-272+13M17D24M	113	1	497	37	37M	15	100338662		0	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
19:20389:F:275+18M2D19M	99	1	17644	0	37M	=	17919	314	TATGACTGCTAATAATACCTACACATGTTAGAACCAT	>>>>>>>>
19:20389:F:275+18M2D19M	147	1	17919	0	18M2D19M	=	17644	-314	GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT	;44999;4
9:21597+10M2I25M:R:-209	83	1	21678	0	8M2I27M	=	21469	-244	CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT	<;9<<5><

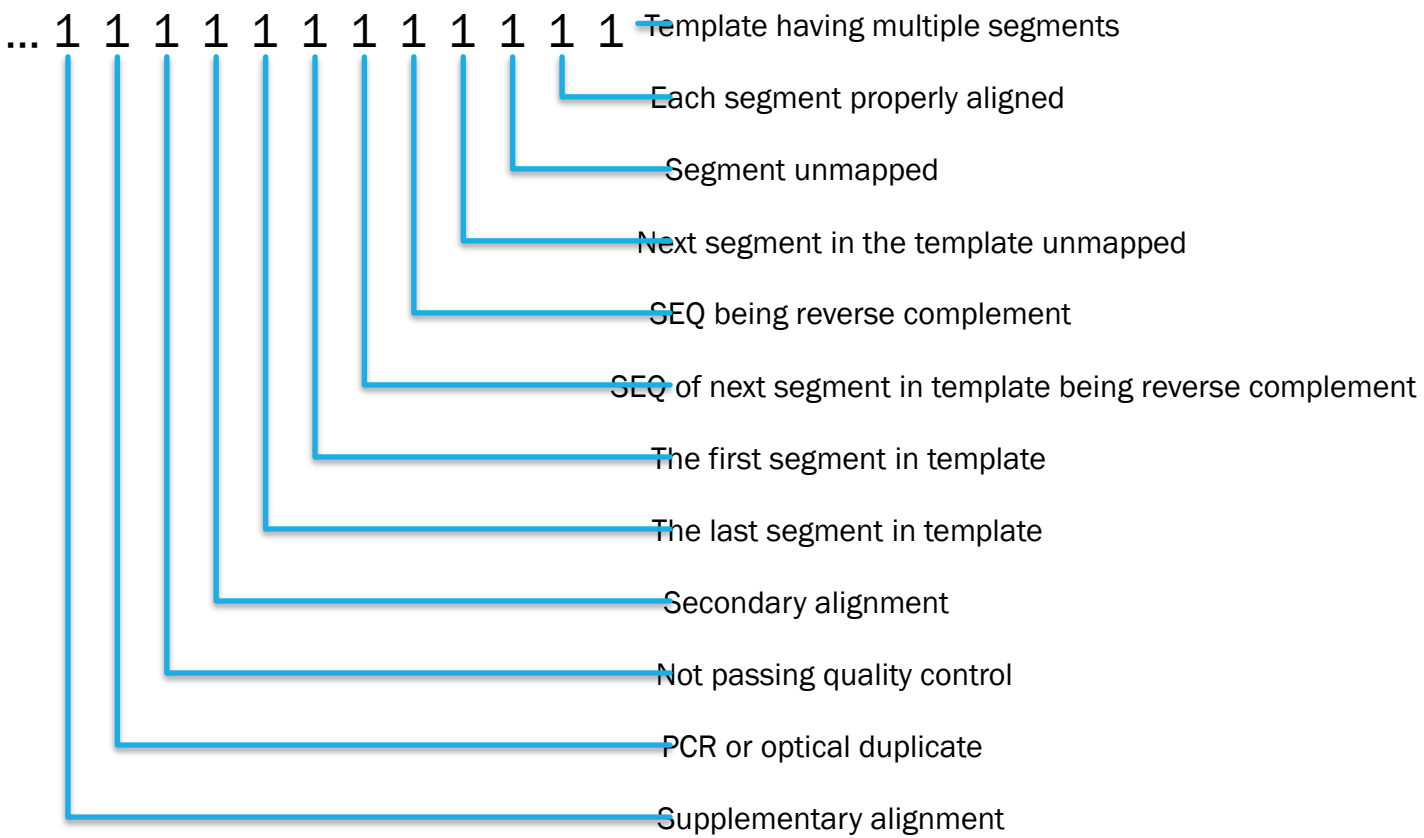
Columns:

QNAME: Query template/pair NAME
FLAG: bitwise FLAG
RNAME: Reference sequence NAME
POS: 1-based leftmost POSition/coordinate of clipped sequence
MAPQ: MAPping Quality (Phred-scaled)
CIGAR: extended CIGAR string
MRNM: Mate Reference sequence NaMe ('=' if same as RNAME)
MPOS: 1-based Mate POSistion
LEN: inferred Template LENgth (insert size)
SEQ: query SEQuence on the same strand as the reference
QUAL: query QUALity (ASCII-33 gives the Phred base quality)
OPT: variable OPTional fields in the format TAG:VTYPE:VALUE

SAM/BAM FORMAT

FLAG Column

If you think of a standard integer as being composed of 32 bits (0 or 1) then it would look like



SAM/BAM FORMAT

FLAG Column

Converting to a number	Information
Bit 0 - false - add nothing	Single End
Bit 1 - true - add 2**1 = 2	Properly Aligned
Bit 2 - false - add nothing	Is Mapped
Bit 3 - true - add 2**3 = 8	NA
Bit 4 - true - add 2**4 = 16	Reverse strand

Bit pattern = 11010 = 16+8+2 = 26

So the flag value would be 26.

https://broadinstitute.github.io/picard/explain-flags.html

SAM/BAM FORMAT

CIGAR Line Column

The CIGAR Line explains how is the alignment between segment and template

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

POS: 5

CIGAR: 3M1I3M1D5M

M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch

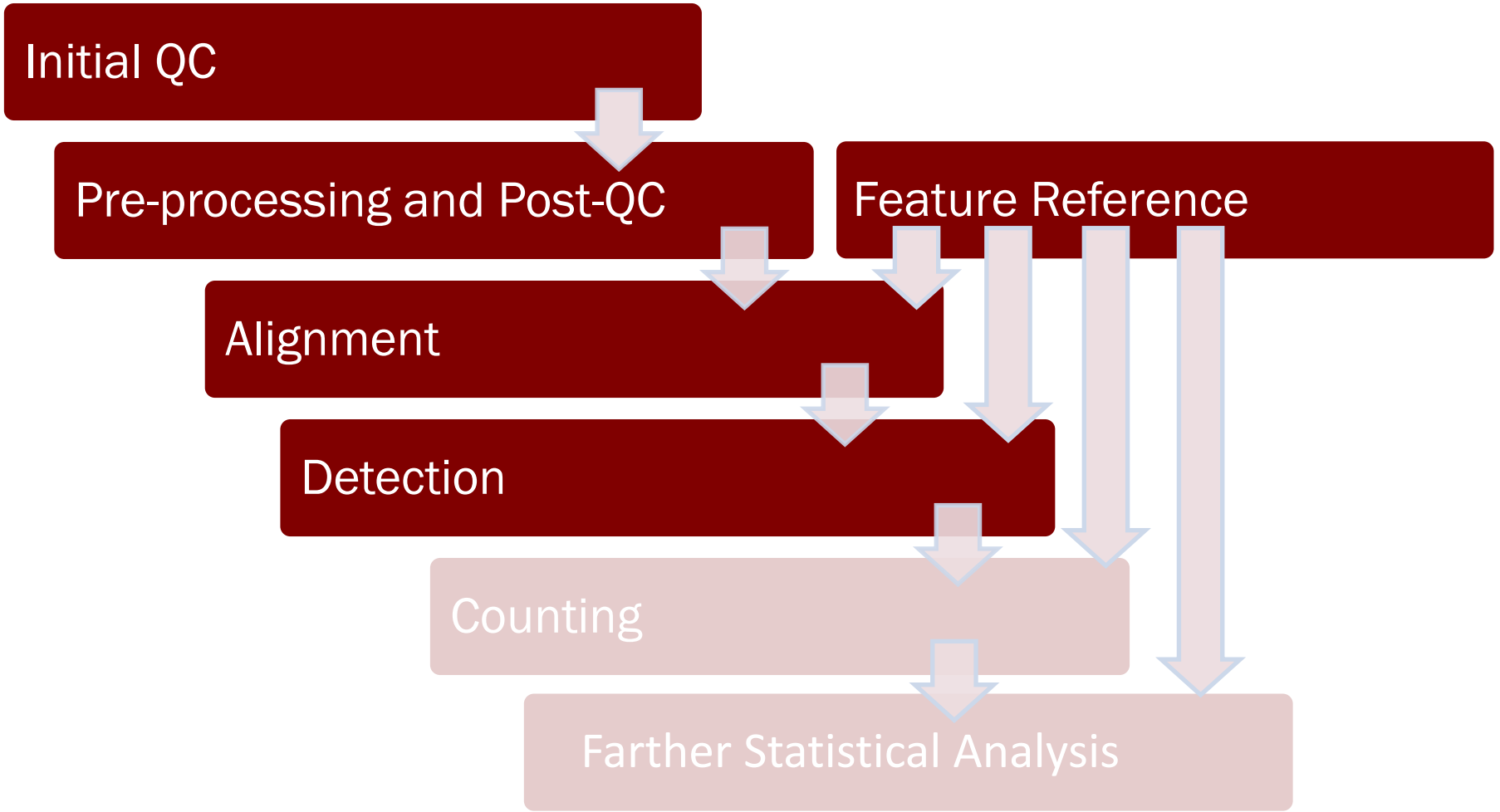
SAM/BAM FORMAT

TAGs Columns

Tag	Type	Description
AM	i	The smallest template-independent mapping quality in the template
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence identifying the sample
BQ	Z	Offset to base alignment quality (BAQ)
BZ	Z	Phred quality of the unique molecular barcode bases in the OX tag
CB	Z	Cell identifier
CC	Z	Reference name of the next hit
CG	B,I	BAM only: CIGAR in BAM's binary encoding if (and only if) it consists of >65535 operators
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read base qualities
CR	Z	Cellular barcode sequence bases (uncorrected)
CS	Z	Color read sequence
CT	Z	Complete read annotation tag, used for consensus annotation dummy features
CY	Z	Phred quality of the cellular barcode sequence in the CR tag
E2	Z	The 2nd most likely base calls
FI	i	The index of segment in the template
FS	Z	Segment suffix
FZ	B,S	Flow signal intensities
GC	?	Reserved for backwards compatibility reasons
GQ	?	Reserved for backwards compatibility reasons
GS	?	Reserved for backwards compatibility reasons
HO	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index
IH	i	Query hit total count
LB	Z	Library
MC	Z	CIGAR string for mate/next segment
MD	Z	String encoding mismatched and deleted reference bases
MF	?	Reserved for backwards compatibility reasons
MI	Z	Molecular identifier; a string that uniquely identifies the molecule from which the record was derived
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contain the query in the current record
NM	i	Edit distance to the reference
OA	Z	Original alignment
OC	Z	Original CIGAR (deprecated; use OA instead)
OP	i	Original mapping position (deprecated; use OA instead)
OQ	Z	Original base quality
OX	Z	Original unique molecular barcode bases
PG	Z	Program
PQ	i	Phred likelihood of the template
PT	Z	Read annotations for parts of the padded read sequence
PU	Z	Platform unit
Q2	Z	Phred quality of the mate/next segment sequence in the R2 tag
QT	Z	Phred quality of the sample barcode sequence in the BC tag
QX	Z	Quality score of the unique molecular identifier in the RX tag
R2	Z	Sequence of the mate/next segment in the template
RG	Z	Read group
RT	?	Reserved for backwards compatibility reasons
RX	Z	Sequence bases of the (possibly corrected) unique molecular identifier
S2	?	Reserved for backwards compatibility reasons
SA	Z	Other canonical alignments in a chimeric alignment
SM	i	Template-independent mapping quality
SQ	?	Reserved for backwards compatibility reasons
TC	i	The number of segments in the template
TS	A	Transcript strand
U2	Z	Phred probability of the 2nd call being wrong conditional on the best being wrong
UQ	i	Phred likelihood of the segment, conditional on the mapping being correct

https://samtools.github.io/hts-specs/SAMtags.pdf

NGS ANALYSIS WORKFLOW



DETECTION BY ENRICHMENT

GOAL: Identify new unknown features

HOW?:

Each type of feature requires a specific algorithm to detect it.

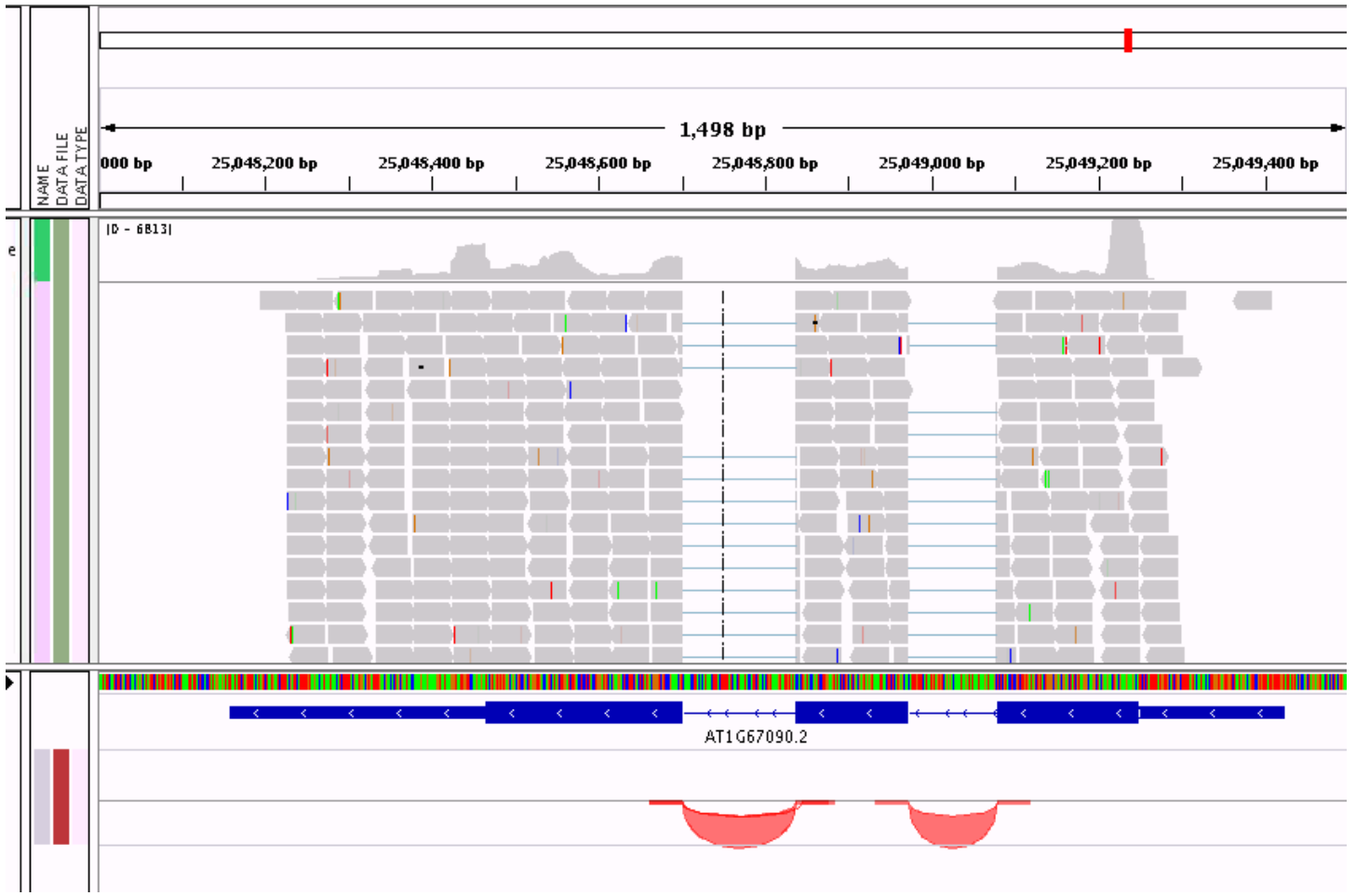
The algorithm may depend on the protocol used to extract your sample

All algorithms are based on:

- Analysis of the level of enrichment of the feature
- Analysis of the compatibility of the region identified with the typical characteristics of the feature we want to detect: Is it compatible with a miRNA or mRNA or TF binding site structure?

DETECTION BY ENRICHMENT

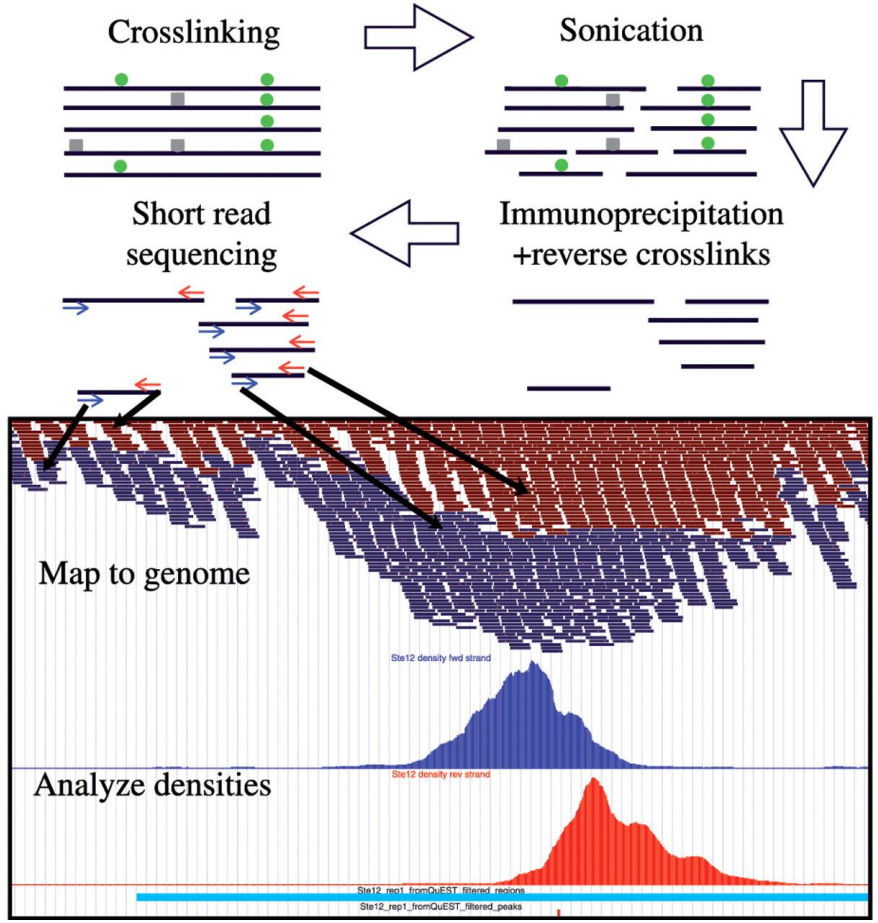
RNA-Seq



Tools: TopHat2 + Cufflinks, ...

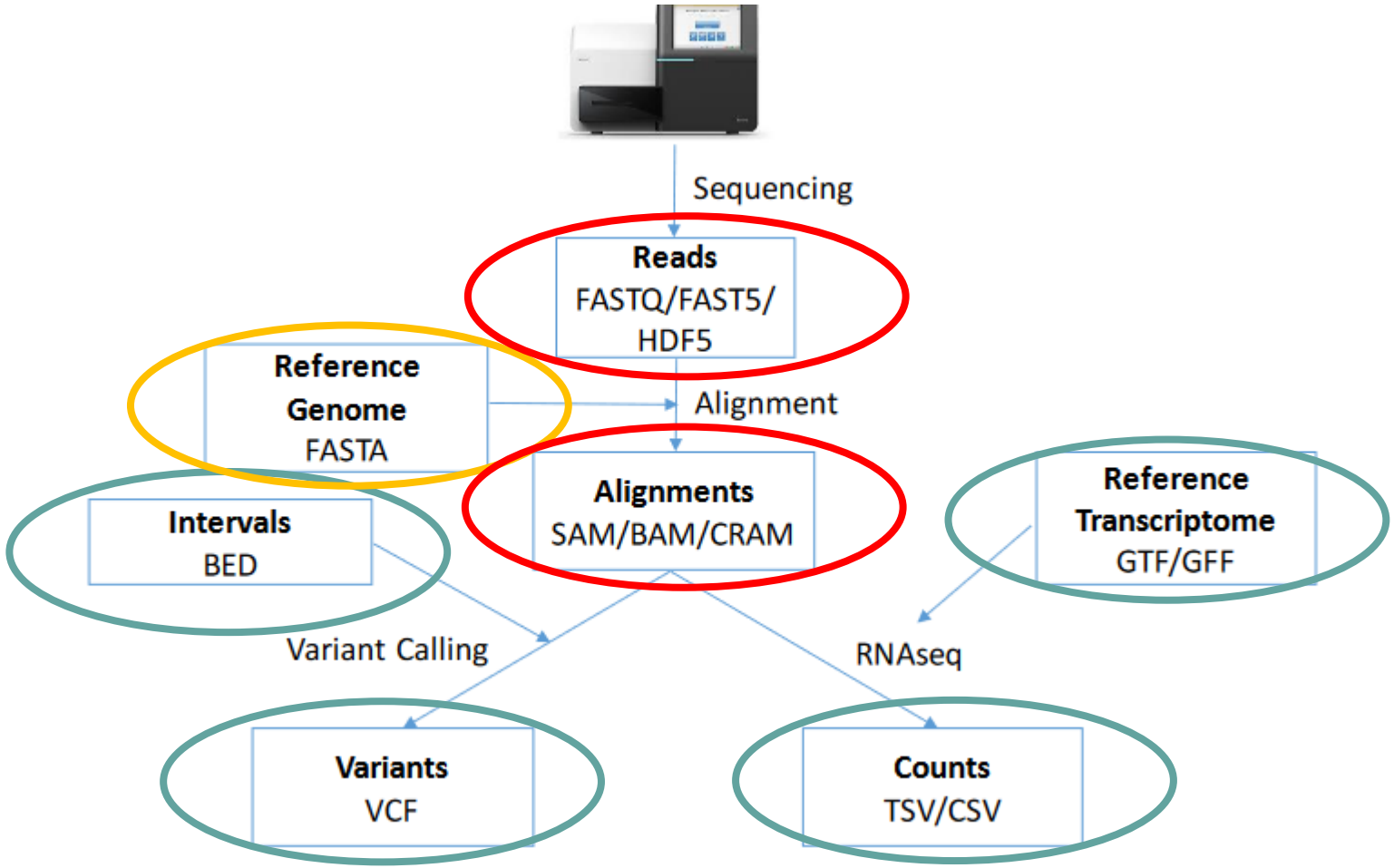
DETECTION BY ENRICHMENT

Chip-Seq



Tools: MACS, PICS, ...

DATA FORMATS



INFORMATION FORMATS

- BED: For annotation of genome features
- GTF: For annotation of genes
- GFF3: For annotation of genes
- VCF: For annotation of variants
- Wig: For annotation of running scores
- ....

BED/BIGBED(BB)

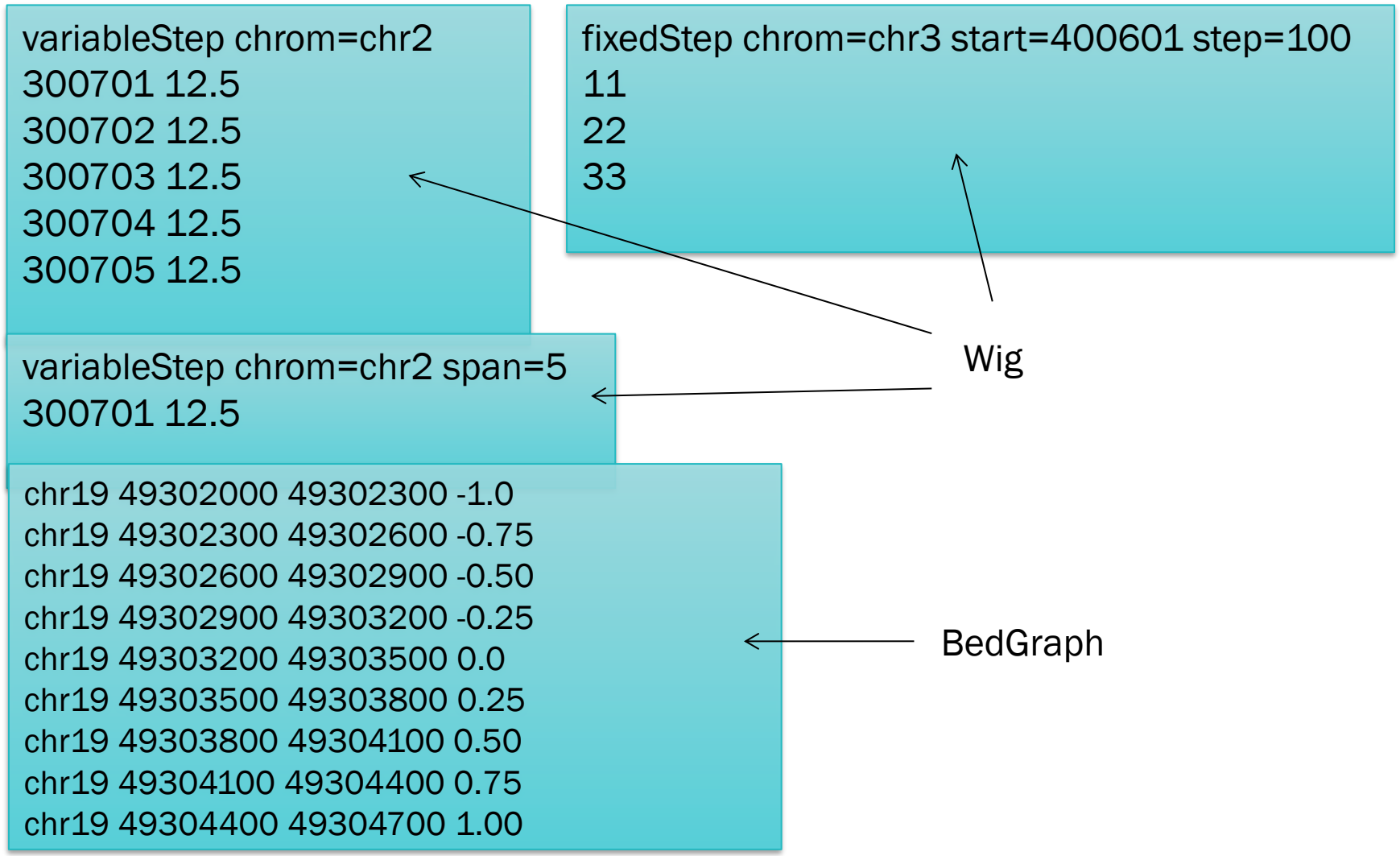
ANALYSIS FEATURES

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

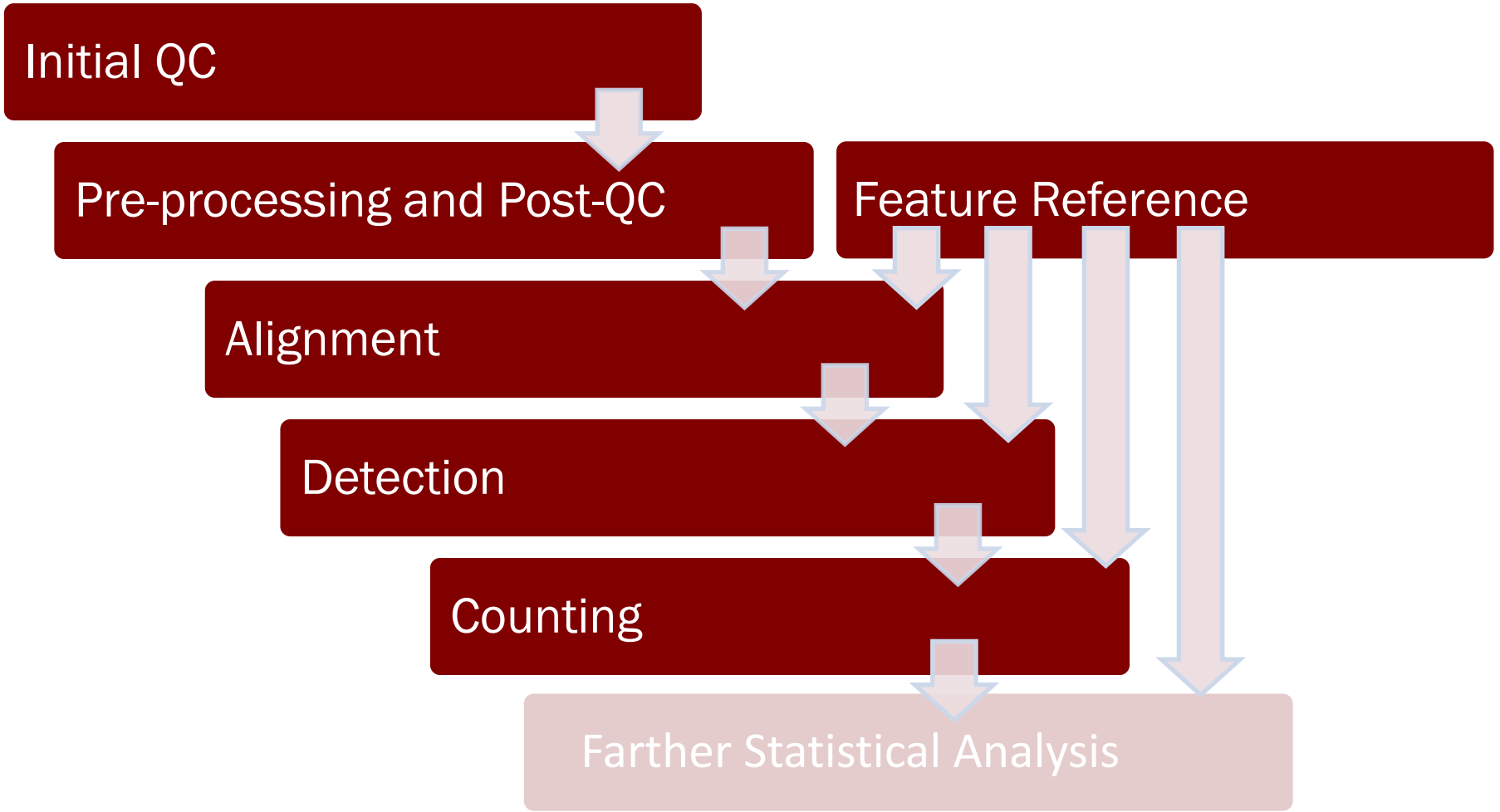
SeqID
Start
End
Feature Name
Score
Strand
thickStart
thickEnd
RGB
blockCount
blockSizes
blockStarts

WIG/BIGWIG OR BEDGRAPH

RUNNING SCORES



NGS ANALYSIS WORKFLOW



COUNT OCCURRENCES

GOAL: How much feature do I have in my sample?

HOW?

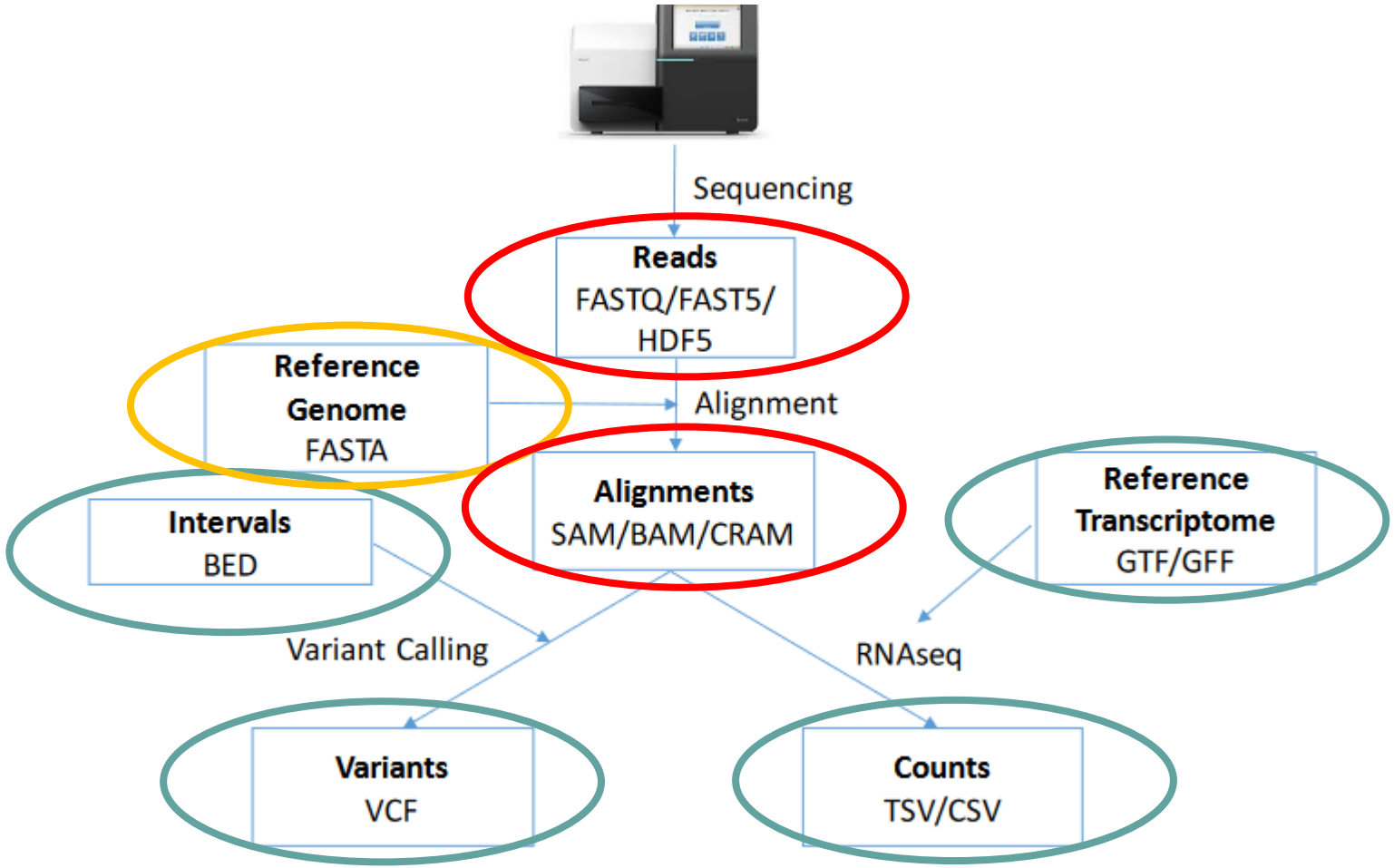
- ✓ Count how many fragments do I obtain from that feature

  In singe-end sequencing a fragment is defined by one read.

  In paired-end sequencing a fragment is defined by two reads.

- ✓ Fragment should be compatible with the feature

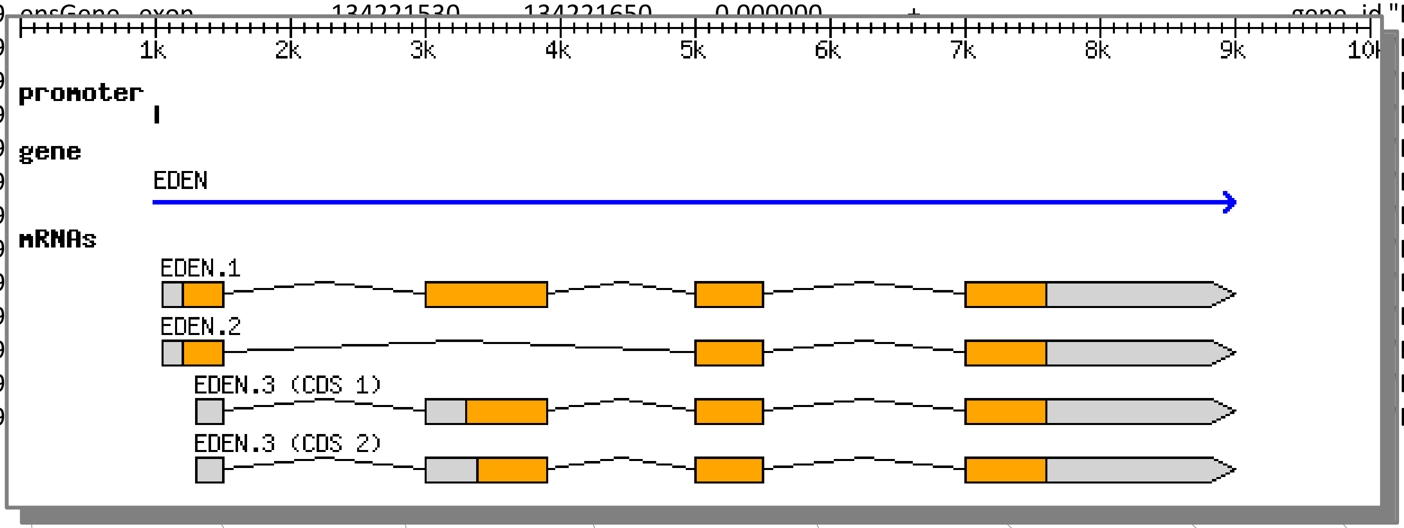
DATA FORMATS



GTF/TABIX

GENE FEATURES

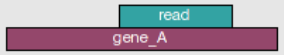
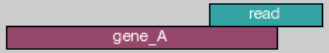
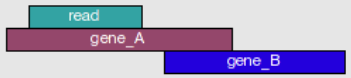
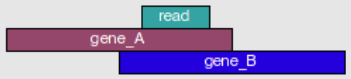
chr1	mm9_ensGene	start_codon	134212807	134212809	0.000000	+	.	gene_id "ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9_ensGene	CDS	134212807	134213049	0.000000	+	0	gene_id "ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9_ensGene	exon	134212703	134213049	0.000000	+	.	gene_id "ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9_ensGene	CDS	134221530	134221650	0.000000	+	0	gene_id "ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9_ensGene	exon	134221530	134221650	0.000000	+		gene_id "ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM
chr1	mm9							ENSMUSG00000009772"; transcript_id "ENSM



SeqID
Source
Type
Start
End
Score
Strand
Phase
Attributes

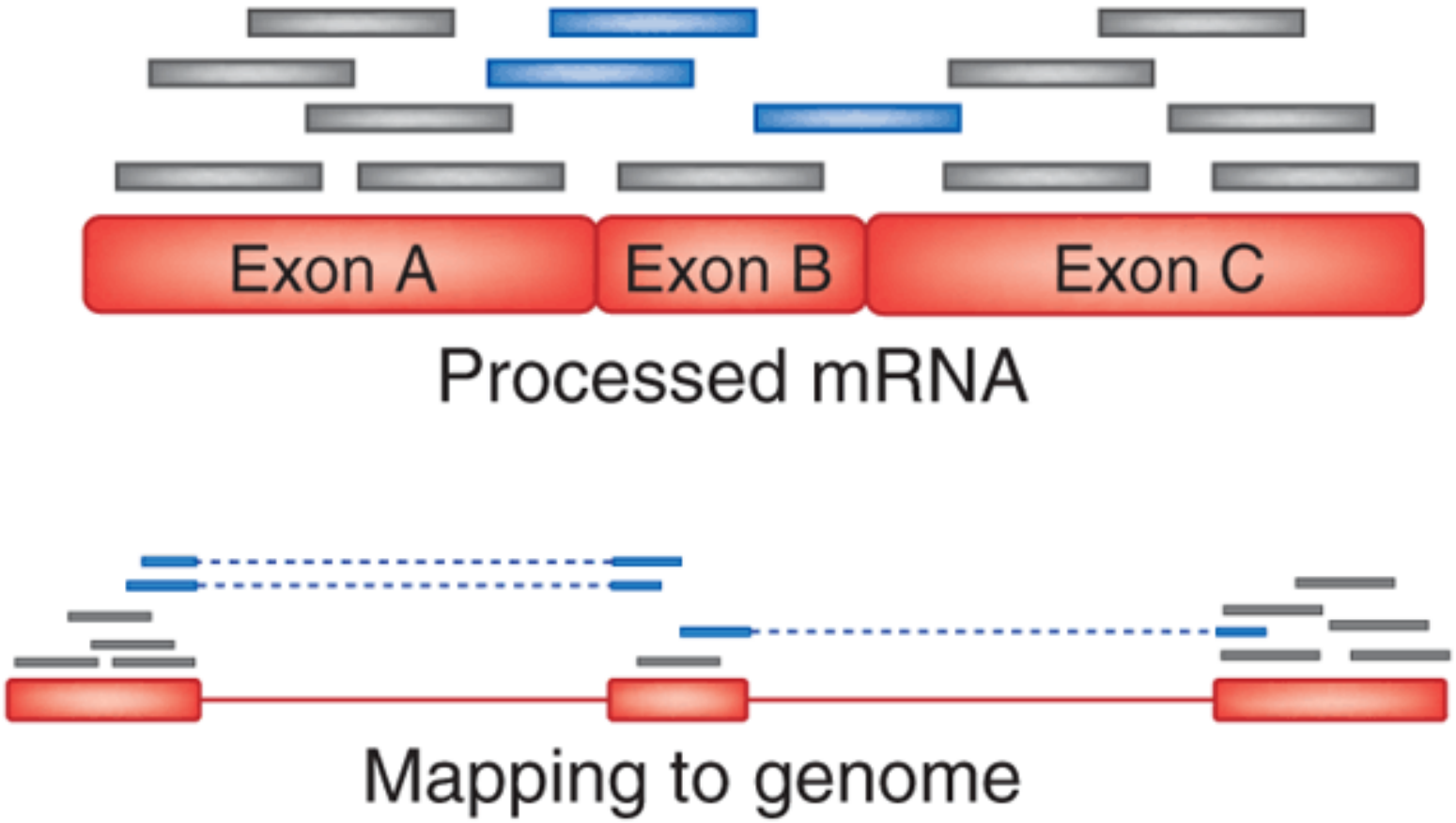
COUNT OCCURRENCES

Htseq count method

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Tools: HTSeq, RSEM, Cufflinks,...

MAPPING READS: THE RNA-SEQ ALIGNMENT PROBLEM



Mapping Reads: The RNA-SEQ alignment problem

Alignment against **TRANSCRIPTOME:**

The reference is a list of ~100.000 transcripts of average length around 2-3kb.

Problems:

Transcriptome is not always available

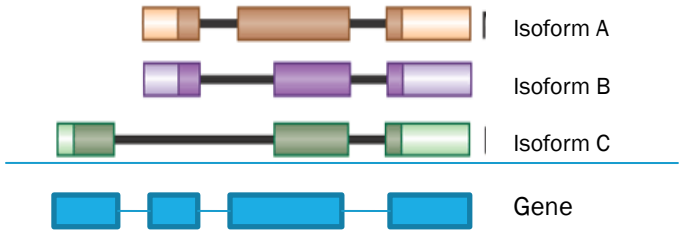
Transcript variance is not fully known even for humans

Multimapping reads

GENE EXPRESSION QUANTIFICATION FROM RNA-SEQ

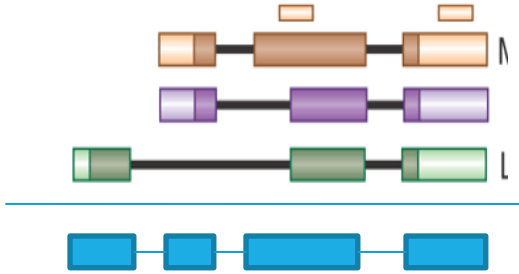
.fastq

Reference Genome/Transcriptome (gtf file)



1. Align reads to reference (Bowtie)

.bam



2. Quantify

.tsv/csv

$C_A = 20$

$C_B = 16$

$C_C = 12$

$C_G = 48$

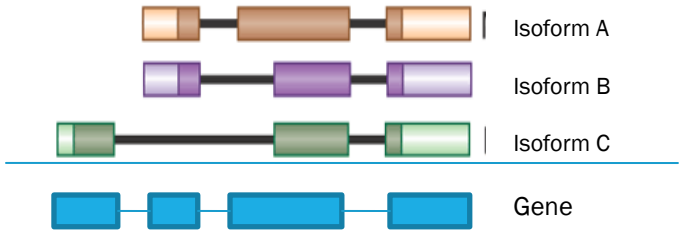
3. Generate estimators

3a. For absolute gene/isoform quantification

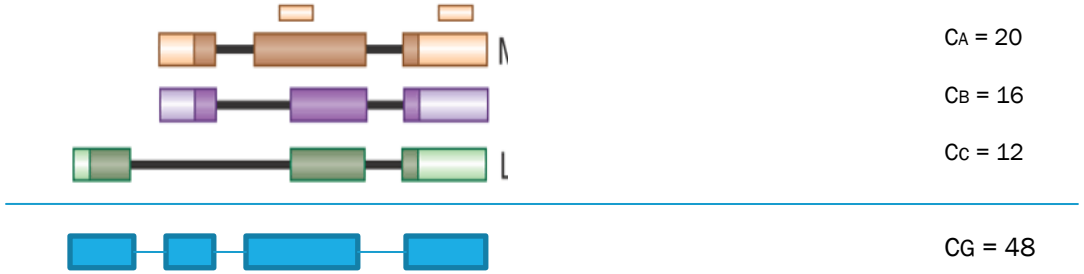
3b. For gene/isoform Differential expression

GENE EXPRESSION QUANTIFICATION FROM RNA-SEQ

Reference Genome/Transcriptome (gtf file)



1. Align reads to reference (Bowtie)
.bam



2. Quantify
.tsv/csv

$C_A = 20$

$C_B = 16$

$C_C = 12$

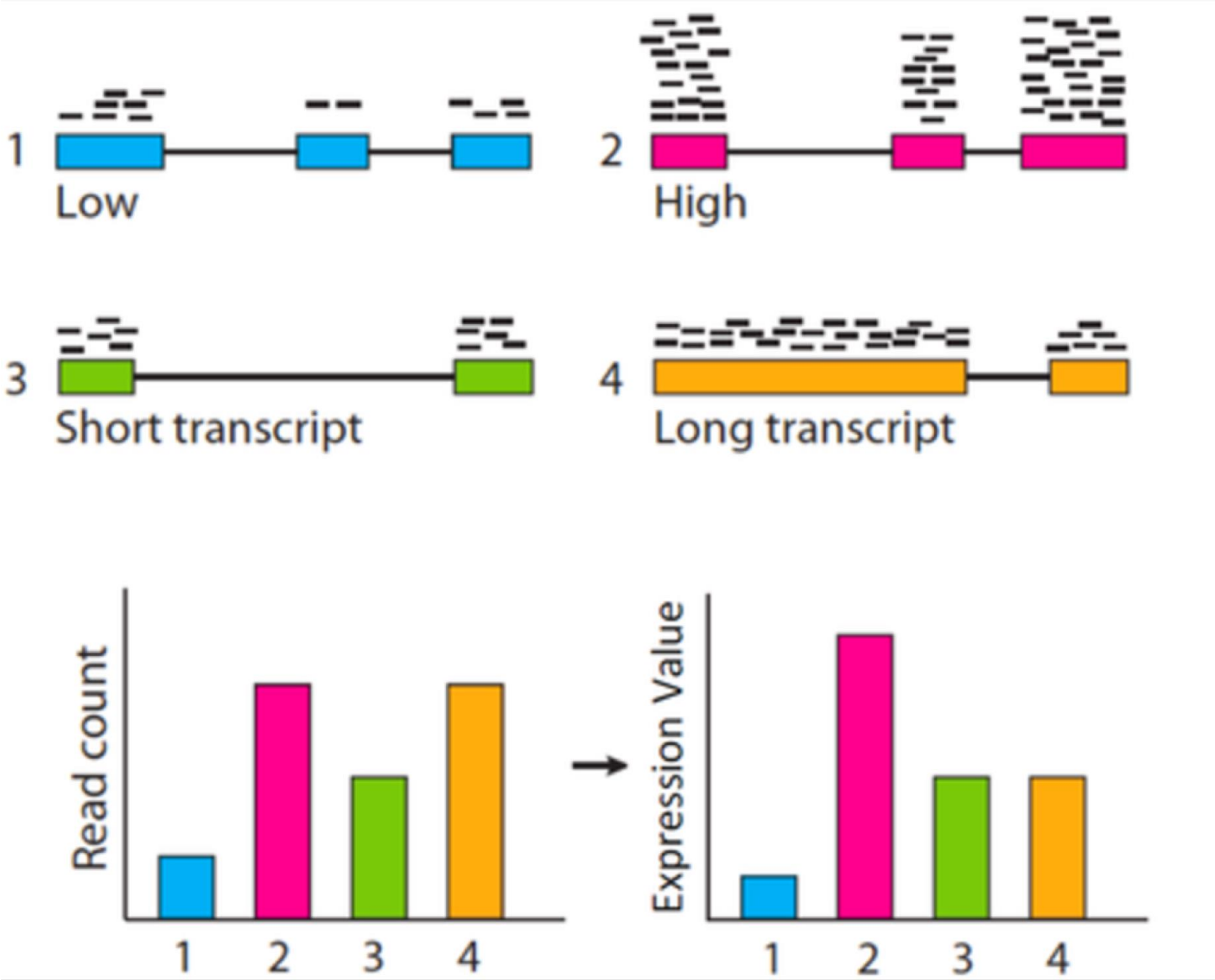
$C_G = 48$

3. Generate estimators

3a. For absolute gene/isoform quantification

3b. For gene/isoform Differential expression

NEED TO NORMALIZE BY LENGTH AND LIBRARY SIZE



QUANTIFICATION METRICS FROM RAW READS: RPKMS

RPKM (FPKM)

- Reads (fragments) per Kilobase Per Million

$$\text{RPKM} = \frac{\text{raw number of reads}}{\text{exon length}} \times \frac{1{,}000{,}000}{\text{Number of reads mapped in the sample}}$$

- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - The number of fragments is biased towards larger genes
  - Total number of fragments is related to total library depth
- Use of FPKM/RPKM normalizes for gene size and library depth

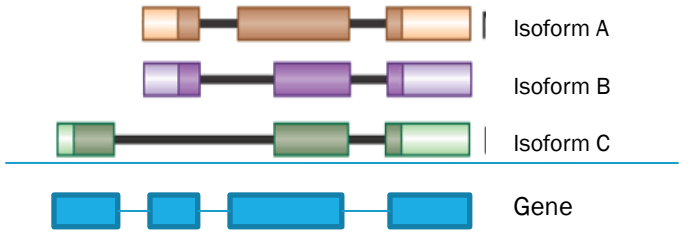
QUANTIFICATION METRICS FROM RAW READS: TPMS

$$\text{RPKM} = 10^9 * \frac{Reads\ mapped\ to\ the\ transcript}{Total\ reads * Transcript\ length}$$

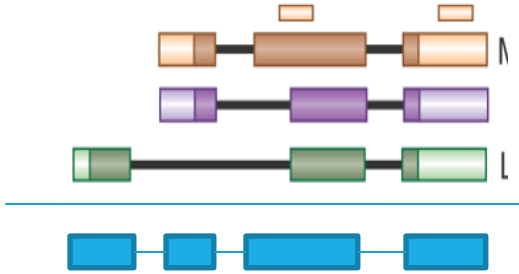
$$\text{TPM} = 10^6 * \frac{reads\ mapped\ to\ transcript\ /\ transcript\ length}{Sum\ (reads\ mapped\ to\ transcript\ /\ transcript\ length)}$$

MULTIMAPPING READS?

Reference Genome/Transcriptome (gtf file)



1. Align reads to reference (Bowtie)
.bam



2. Quantify
.tsv/csv

$C_A = 20$

$C_B = 16$

$C_C = 12$

$C_G = 48$

3. Generate estimators

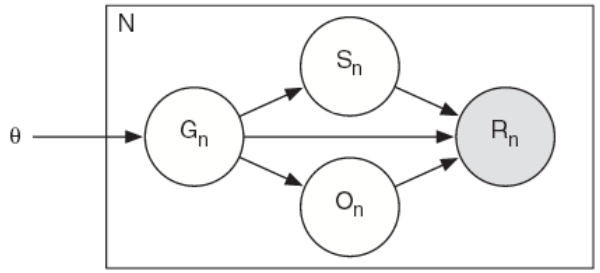
3a. For absolute gene/isoform quantification

3b. For gene/isoform Differential expression

RSEM/SALMON/SAILFISH TO DEAL WITH MULTIMAPPING READS

We have the data (N reads of length L): $R_1,...R_N$

We want to infer the expression of each isoform $[\theta_0... \theta_M]$ assuming the data structure:



Complete Data Likelihood:

$$P(g,s,o,r|\theta)=\prod_{n=1}^{N}P(g_n|\theta)P(s_n|g_n)P(o_n|g_n)P(r_n|g_n,s_n,o_n).$$

Observed Likelihood:

$$P(r|\theta)=\prod_{n=1}^{N}\sum_{i=0}^{M}\theta_i P(r_n|G_n=i).$$

→ Iteratively estimate $[\theta_0... \theta_M]$ maximizing the likelihood

For each read, there is a certain probability to belong to each isoform, what depends on the isoform expression level (θ), read alignment(s) start position (S) and orientation (O)

RSEM OUTPUT

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM	pme_expected_count	pme_TPM	pme_FPKM	TPM_ci_lower_bound	TPM_ci_upper_bound	FPKM_ci_lower_bound	FPKM_ci_upper_bound
	ENSMUSG00000000001	ENSMUST00000000001	3262	3207.7	2529	74.29	41.82	2529	72.54	41.62	69.7951	75.4632	40.0213	43.2715
	ENSMUSG00000000003	ENSMUST00000000003,ENSMUST00000114041	799.5	745.2	0	0	0	0	0.25	0.14	0.00537937	0.604143	0.00260875	0.346176
	ENSMUSG00000000028	ENSMUST00000000028,ENSMUST00000096990,ENSM	1987.22	1932.91	207	10.09	5.68	207	10.15	5.83	8.67045	11.6744	4.97811	6.70237
	ENSMUSG00000000031	ENSMUST00000132294,ENSMUST00000136359,ENSM	2204.58	2150.27	915	40.1	22.57	915	40.64	23.32	37.3474	44.0638	21.4331	25.2876
	ENSMUSG00000000037	ENSMUST00000019101,ENSMUST00000074802,ENSM	4681.6	4627.3	41	0.83	0.47	41	1.55	0.89	0.940912	2.31145	0.539678	1.32629
	ENSMUSG00000000049	ENSMUST00000000049,ENSMUST00000133383,ENSM	1190	1135.7	7	0.58	0.33	7	1.37	0.78	0.534927	2.3483	0.306986	1.3477
	ENSMUSG00000000056	ENSMUST00000103015,ENSMUST00000151088,ENSM	4395	4340.7	1881	40.83	22.98	1881	41.62	23.88	38.7643	44.9376	22.2269	25.7701
	ENSMUSG00000000058	ENSMUST00000000058,ENSMUST00000115459,ENSM	2714.32	2660.02	799	28.31	15.93	799	28.24	16.2	26.1399	30.3115	14.9942	17.39
	ENSMUSG00000000078	ENSMUST00000000080	4217	4162.7	4183	94.69	53.3	4183	92.44	53.04	89.7126	95.313	51.45	54.6621
	ENSMUSG00000000085	ENSMUST00000000087,ENSMUST00000064991,ENSM	2504.94	2450.65	600	23.07	12.99	600	25.6	14.69	22.2064	29.1364	12.7437	16.7199
	ENSMUSG00000000088	ENSMUST00000000090	690	635.7	11105.97	1647.83	927.52	11105.97	1608.44	922.88	1578.71	1638.18	905.846	939.984
	ENSMUSG00000000093	ENSMUST00000000095	3626	3571.7	152	4.01	2.26	152	3.94	2.26	3.34723	4.59043	1.90237	2.61598
	ENSMUSG00000000094	ENSMUST00000000096,ENSMUST00000108045,ENSM	2959	2904.7	1	0.03	0.02	1	0.12	0.07	0.022024	0.244917	0.0126373	0.140546
	ENSMUSG00000000103	ENSMUST00000115891	2816	2761.7	0	0	0	0	0.03	0.02	1.04E-07	0.100797	5.94E-08	0.0578369
	ENSMUSG00000000120	ENSMUST00000000122	3446	3391.7	19	0.53	0.3	19	0.54	0.31	0.316652	0.780288	0.181512	0.447536
	ENSMUSG00000000125	ENSMUST00000000127	3076	3021.7	0	0	0	0	0.03	0.02	6.44E-07	0.0919652	3.69E-07	0.0527683
	ENSMUSG00000000126	ENSMUST00000000128,ENSMUST00000108783	3318	3263.7	123	3.55	2	123	3.52	2.02	2.91472	4.13889	1.67756	2.37999
	ENSMUSG00000000127	ENSMUST00000000129,ENSMUST00000038080,ENSM	2999.93	2945.62	754	24.12	13.58	754	23.78	13.64	22.1399	25.5253	12.699	14.6404
	ENSMUSG00000000131	ENSMUST00000009344,ENSMUST00000163959,ENSM	3438.5	3384.19	893	24.87	14	893	28.94	16.6	25.4074	32.6361	14.5771	18.7249
	ENSMUSG00000000134	ENSMUST00000077680,ENSMUST00000079542,ENSM	2985.47	2931.17	454	14.6	8.22	454	15.94	9.15	13.8352	18.2613	7.9338	10.4744
	ENSMUSG00000000142	ENSMUST00000052915,ENSMUST00000106711,ENSM	3959.6	3905.29	31	0.75	0.42	31	1.66	0.95	0.943915	2.43437	0.541943	1.39705
	ENSMUSG00000000148	ENSMUST00000041588,ENSMUST00000064662,ENSM	3307.16	3252.85	199	5.76	3.24	199	5.91	3.39	5.07522	6.69813	2.91754	3.84894
	ENSMUSG00000000149	ENSMUST00000000153,ENSMUST00000166751	3416	3361.7	2495	69.94	39.37	2495	68.77	39.46	65.9836	71.6377	37.8634	41.108
	ENSMUSG00000000154	ENSMUST00000052348,ENSMUST00000105917,ENSM	1379	1324.7	7	0.5	0.28	7	1.39	0.8	0.597083	2.3248	0.342716	1.33398
	ENSMUSG00000000157	ENSMUST00000000161,ENSMUST00000113772,ENSM	2036.4	1982.1	0	0	0	0	0.33	0.19	0.0584069	0.710492	0.0335427	0.407687
	ENSMUSG00000000159	ENSMUST00000000163,ENSMUST00000081093,ENSM	1554	1499.7	1	0.06	0.04	1.07	0.88	0.51	0.219588	1.72664	0.125919	0.990486
	ENSMUSG00000000167	ENSMUST00000000171,ENSMUST00000124029,ENSM	1296.71	1242.41	14	1.06	0.6	14	1.93	1.11	1.08944	2.86604	0.619167	1.63877
	ENSMUSG00000000168	ENSMUST00000034567,ENSMUST00000125919,ENSM	2648.88	2594.57	6250	227.01	127.78	6250	222.81	127.84	211.778	233.756	121.669	134.278
	ENSMUSG00000000171	ENSMUST00000000175	1287	1232.7	7620	582.73	328	7620	568.82	326.38	556.186	581.612	319.1	333.688
	ENSMUSG00000000182	ENSMUST00000000186	1814	1759.7	0	0	0	0	0.05	0.03	1.23E-06	0.157191	7.08E-07	0.0902044
	ENSMUSG00000000183	ENSMUST00000000187	1232	1177.7	0	0	0	0	0.08	0.04	2.18E-07	0.232767	1.25E-07	0.133561
	ENSMUSG00000000184	ENSMUST00000000188	5701	5646.7	5584	93.18	52.45	5584	90.96	52.19	88.5351	93.3252	50.8045	53.5517
	ENSMUSG00000000194	ENSMUST00000056433,ENSMUST00000127872,ENSM	3723.93	3669.63	1428	36.67	20.64	1427.71	38.7	22.2	35.1242	42.4497	20.1561	24.3589
	ENSMUSG00000000197	ENSMUST00000000201,ENSMUST00000072254	5458.5	5404.2	0	0	0	0	0.04	0.02	0.000634084	0.091341	0.000397396	0.0524487
	ENSMUSG00000000202	ENSMUST00000000206,ENSMUST00000137965,ENSM	1423.86	1369.57	0	0	0	0	0.89	0.51	0.232421	1.69159	0.13232	0.969688
	ENSMUSG00000000204	ENSMUST00000000208,ENSMUST00000164672,ENSM	2485.25	2430.96	0	0	0	0	0.33	0.19	0.0258611	0.817995	0.0138819	0.468392
	ENSMUSG00000000214	ENSMUST00000000219,ENSMUST00000105929,ENSM	1265	1210.7	0	0	0	0	0.58	0.33	0.146407	1.13867	0.084008	0.653359
	ENSMUSG00000000215	ENSMUST00000000220,ENSMUST00000105930,ENSM	447	392.89	0	0	0	0	2.08	1.19	0.720816	3.59725	0.415513	2.06612
	ENSMUSG00000000216	ENSMUST00000000221,ENSMUST00000169565	2991	2936.7	0	0	0	0	0.06	0.04	0.00106143	0.149017	0.000609494	0.085524
	ENSMUSG00000000223	ENSMUST00000000228,ENSMUST00000113224,ENSM	6012.06	5957.76	15	0.24	0.13	15	0.48	0.28	0.257377	0.747123	0.147654	0.428752
	ENSMUSG00000000244	ENSMUST00000009396,ENSMUST00000075172,ENSM	1584.37	1530.07	39	2.4	1.35	39	3.7	2.12	2.50382	4.96612	1.44506	2.85803
	ENSMUSG00000000247	ENSMUST00000000253,ENSMUST00000133661,ENSM	1532.54	1478.24	7	0.45	0.25	7	1.24	0.71	0.439268	2.17085	0.250048	1.24382
	ENSMUSG00000000248	ENSMUST00000000254,ENSMUST00000075789,ENSM	2404	2349.7	1	0.04	0.02	1	0.58	0.33	0.0776316	1.27809	0.0444415	0.733319
	ENSMUSG00000000253	ENSMUST00000000260,ENSMUST00000128873,ENSM	1548.74	1494.44	1304	82.25	46.3	1304	81.18	46.58	76.4817	85.9259	43.8756	49.2957

RSEM TO DEAL WITH MULTIMAPPING READS

A	B	C	D	E	F	G	H	I	J	K	L	M	N
gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM	pme_expected_count	pme_TPM	pme_FPKM	TPM_ci_lower_bound	TPM_ci_upper_bound	FPKM_ci_lower_bound	FPKM_ci_upper_bound
ENSMUSG00000000001	ENSMUST00000000001	3262	3207.7	2529	74.29	41.82	2529	72.54	41.62	69.7951	75.4632	40.0213	43.2715

According to Reference

Part of the gene really expressed

C

Transcripts per million

$$\tau_i = \frac{v_i}{\ell_i}\left(\sum_j \frac{v_j}{\ell_j}\right)^{-1},$$

(x 10^6)

Fragments per Kb
per Million mapped Reads

$$FPKM_I = \frac{C_I / l_I}{N_m / 10^9} = \frac{\tau_I}{\sum \tau_j l_j} \times 10^9$$

Counts and TPM: Comparison analysis
FPKMs: Absolute quantification

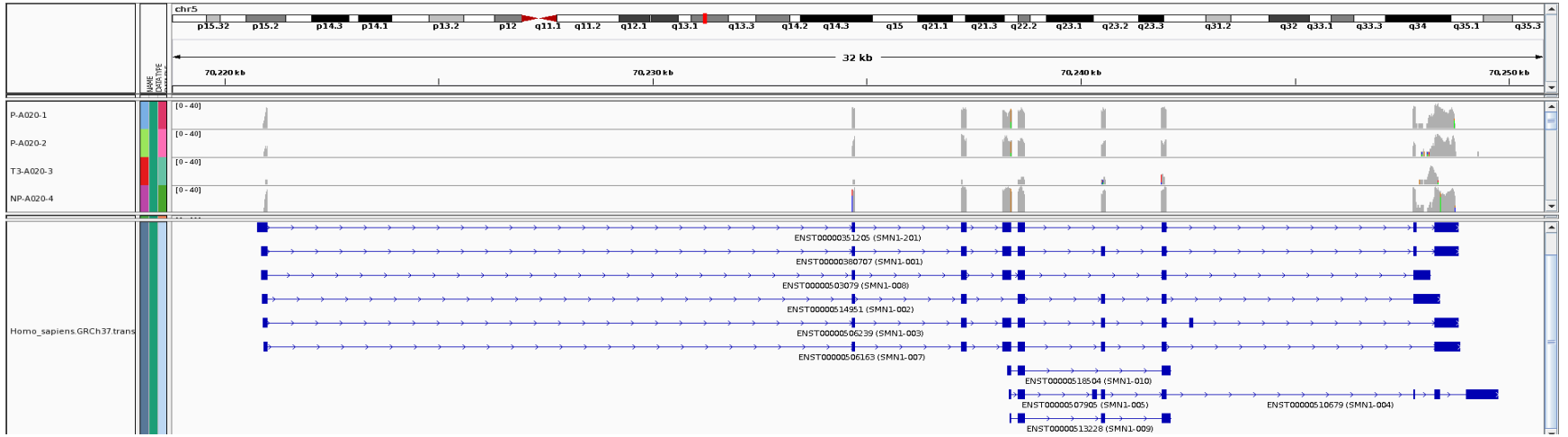
RSEM TO DEAL WITH MULTIMAPPING READS

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct	pme_expected_count	pme_TPM	pme_FPKM	IsoPct_from_pme_TPM	TPM_ci_lower_bound	TPM_ci_upper_bound	FPKM_ci_lower_bound	FPKM_ci_upper_bound
ENSMUST00000000001	ENSMUSG00000000001	3262	3207.7	2529	74.29	41.82	100	2529	72.54	41.62	100	69.7951	75.4632	40.0213	43.2715
ENSMUST00000000003	ENSMUSG00000000003	902	847.7	0	0	0	0	0	0.11	0.06	43.12	3.02E-06	0.326719	1.73E-06	0.187454
ENSMUST00000114041	ENSMUSG00000000003	697	642.7	0	0	0	0	0	0.14	0.08	56.88	1.59E-06	0.431759	9.14E-07	0.247738
ENSMUST00000000028	ENSMUSG00000000028	2143	2088.7	172.88	7.8	4.39	77.28	166.74	7.39	4.24	72.75	5.47529	9.33949	3.13871	5.35619
ENSMUST00000096990	ENSMUSG00000000028	1747	1692.7	28.14	1.57	0.88	15.52	33.9	1.9	1.09	18.68	0.00183	3.72657	0.000386208	2.13777
ENSMUST00000115585	ENSMUSG00000000028	832	777.7	5.99	0.73	0.41	7.19	6.36	0.87	0.5	8.58	0.211385	1.63265	0.121137	0.936587
ENSMUST00000132294	ENSMUSG00000000031	452	397.74	0	0	0	0	1.13	0.49	0.28	1.21	6.25E-06	1.46732	3.59E-06	0.842268
ENSMUST00000136359	ENSMUSG00000000031	2286	2231.7	897.02	37.88	21.32	94.46	849.69	35.06	20.12	86.27	31.0887	38.5542	17.8444	22.1292
ENSMUST00000140716	ENSMUSG00000000031	817	762.7	17.98	2.22	1.25	5.54	20.47	2.59	1.49	6.37	0.204421	5.27501	0.116679	3.02605
ENSMUST00000149974	ENSMUSG00000000031	935	880.7	0	0	0	0	2	0.31	0.18	0.77	6.05E-06	0.920688	3.47E-06	0.528281
ENSMUST00000152754	ENSMUSG00000000031	1853	1798.7	0	0	0	0	41.72	2.18	1.25	5.38	6.03E-05	5.95843	3.46E-05	3.41928
ENSMUST00000019101	ENSMUSG00000000037	4842	4787.7	37.15	0.73	0.41	87.59	17.21	0.35	0.2	22.53	0.0295913	0.634787	0.0170179	0.364368
ENSMUST00000074802	ENSMUSG00000000037	3440	3385.7	0	0	0	0	2.27	0.09	0.05	5.73	4.60E-06	0.255803	2.64E-06	0.14675
ENSMUST00000077375	ENSMUSG00000000037	3550	3495.7	3.85	0.1	0.06	12.41	2.37	0.09	0.05	5.71	3.62E-07	0.234194	2.08E-07	0.134406
ENSMUST00000087090	ENSMUSG00000000037	2953	2898.7	0	0	0	0	2.25	0.1	0.06	6.64	3.97E-08	0.291516	2.28E-08	0.167275
ENSMUST00000087091	ENSMUSG00000000037	2959	2904.7	0	0	0	0	2.17	0.1	0.06	6.46	1.14E-06	0.286562	6.56E-07	0.164439
ENSMUST00000101113	ENSMUSG00000000037	2931	2876.7	0	0	0	0	2.03	0.1	0.06	6.24	9.38E-07	0.272205	5.38E-07	0.156184
ENSMUST00000112342	ENSMUSG00000000037	3440	3385.7	0	0	0	0	2.22	0.09	0.05	5.63	1.20E-06	0.248412	6.87E-07	0.142577
ENSMUST00000112345	ENSMUSG00000000037	4847	4792.7	0	0	0	0	7.46	0.16	0.09	10.46	1.38E-06	0.410731	7.90E-07	0.235675
ENSMUST00000112348	ENSMUSG00000000037	3597	3542.7	0	0	0	0	2.04	0.08	0.05	5.08	7.09E-06	0.213131	4.07E-06	0.122292
ENSMUST00000124775	ENSMUSG00000000037	517	462.71	0	0	0	0	0.99	0.4	0.23	25.52	1.07E-05	1.06668	6.15E-06	0.611952
ENSMUST00000000049	ENSMUSG00000000049	1190	1135.7	7	0.58	0.33	100	7	0.65	0.37	47.38	0.253343	1.11528	0.144859	0.639592
ENSMUST00000133383	ENSMUSG00000000049	367	312.85	0	0	0	0	0	0.3	0.17	21.58	6.51E-06	0.886131	3.74E-06	0.508576
ENSMUST00000146050	ENSMUSG00000000049	374	319.84	0	0	0	0	0	0.29	0.17	21.1	2.47E-06	0.863103	1.42E-06	0.495325
ENSMUST00000152958	ENSMUSG00000000049	731	676.7	0	0	0	0	0	0.14	0.08	9.94	8.08E-07	0.405393	4.64E-07	0.232586
ENSMUST00000103015	ENSMUSG00000000056	4395	4340.7	1881	40.83	22.98	100	1875.06	39.75	22.81	95.5	37.9729	41.5992	21.8059	23.8865
ENSMUST00000151088	ENSMUSG00000000056	525	470.71	0	0	0	0	0.04	0.2	0.12	0.49	6.61E-06	0.603681	3.79E-06	0.346428
ENSMUST00000154047	ENSMUSG00000000056	435	380.75	0	0	0	0	5.9	1.67	0.96	4.01	5.34E-05	4.46644	3.06E-05	2.56342
ENSMUST00000000058	ENSMUSG00000000058	2733	2678.7	796.07	28.01	15.76	98.94	792.86	27.26	15.64	96.54	25.3767	29.1731	14.5706	16.7493
ENSMUST00000115459	ENSMUSG00000000058	974	919.7	2.93	0.3	0.17	1.06	3.42	0.44	0.25	1.57	0.046991	0.92865	0.0290171	0.534952
ENSMUST00000115462	ENSMUSG00000000058	694	639.7	0	0	0	0	2.73	0.54	0.31	1.9	1.72E-05	1.54996	9.85E-06	0.889469
ENSMUST00000000080	ENSMUSG00000000078	4217	4162.7	4183	94.69	53.3	100	4183	92.44	53.04	100	89.7126	95.313	51.45	54.6621
ENSMUST00000000087	ENSMUSG00000000085	3353	3298.7	327.03	9.34	5.26	40.49	320.37	8.96	5.14	34.99	5.12269	12.3325	2.9422	7.07969
ENSMUST00000064991	ENSMUSG00000000085	2548	2493.7	0	0	0	0	31.9	1.21	0.7	4.74	6.19E-05	3.26465	3.55E-05	1.87347
ENSMUST00000106298	ENSMUSG00000000085	3090	3035.7	218.64	6.79	3.82	29.42	144.98	4.42	2.54	17.27	0.401461	8.44625	0.226024	4.84184
ENSMUST00000106301	ENSMUSG00000000085	2584	2529.7	0	0	0	0	31.88	1.2	0.69	4.67	1.33E-05	3.04592	7.63E-06	1.74777
ENSMUST00000122860	ENSMUSG00000000085	530	475.71	0	0	0	0	0.74	0.34	0.19	1.32	5.40E-06	1.00058	3.10E-06	0.574122
ENSMUST00000132116	ENSMUSG00000000085	623	568.7	19.11	3.17	1.78	13.74	19.31	3.29	1.89	12.84	1.768	4.84039	1.01403	2.77668
ENSMUST00000133169	ENSMUSG00000000085	443	388.74	1.54	0.38	0.21	1.63	1.77	0.66	0.38	2.56	0.0166942	1.52986	0.00960296	0.877742
ENSMUST00000133290	ENSMUSG00000000085	557	502.7	0	0	0	0	0.15	0.21	0.12	0.82	4.33E-06	0.632159	2.49E-06	0.362798
ENSMUST00000134375	ENSMUSG00000000085	1132	1077.7	2.47	0.22	0.12	0.94	2.66	0.31	0.18	1.22	0.0335413	0.668715	0.0194505	0.383914
ENSMUST00000136801	ENSMUSG00000000085	1536	1481.7	26.48	1.68	0.95	7.3	38.18	2.43	1.4	9.5	0.732508	4.23487	0.421119	2.43096
ENSMUST00000144555	ENSMUSG00000000085	3544	3489.7	0	0	0	0	0.9	0.05	0.03	0.2	1.19E-06	0.150944	6.84E-07	0.086615
ENSMUST00000144862	ENSMUSG00000000085	595	540.7	0	0	0	0	1.62	0.45	0.26	1.74	6.20E-06	1.31677	3.56E-06	0.755492

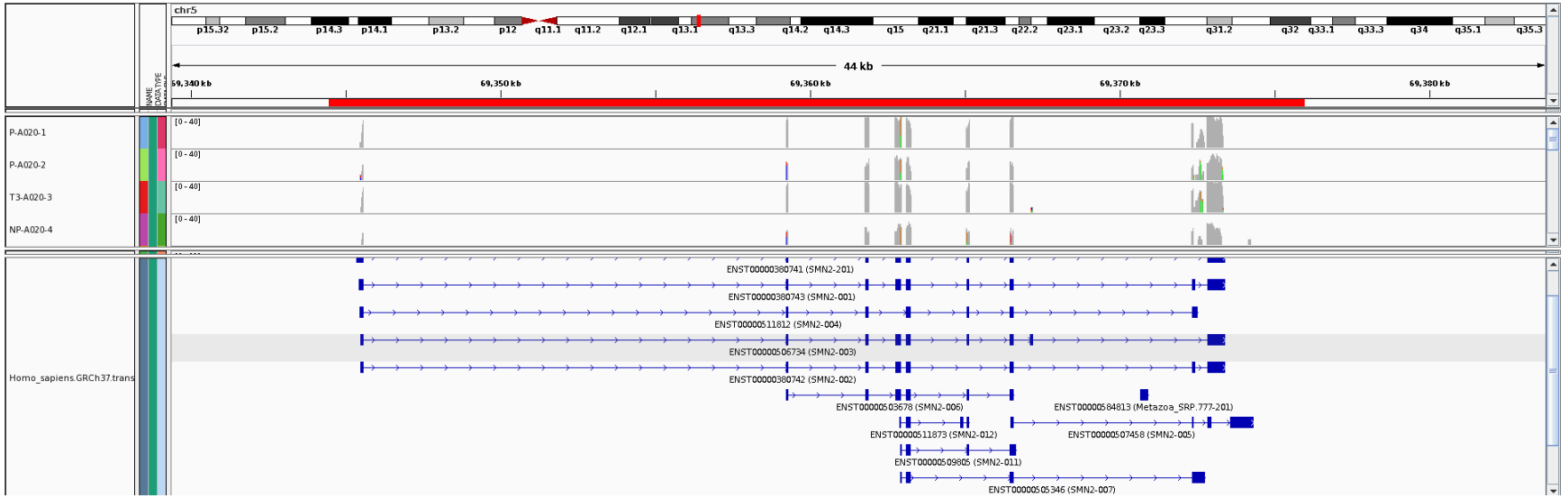
% of expression of this isoform over the whole gene exp

RSEM TO DEAL WITH MULTIMAPPING READS

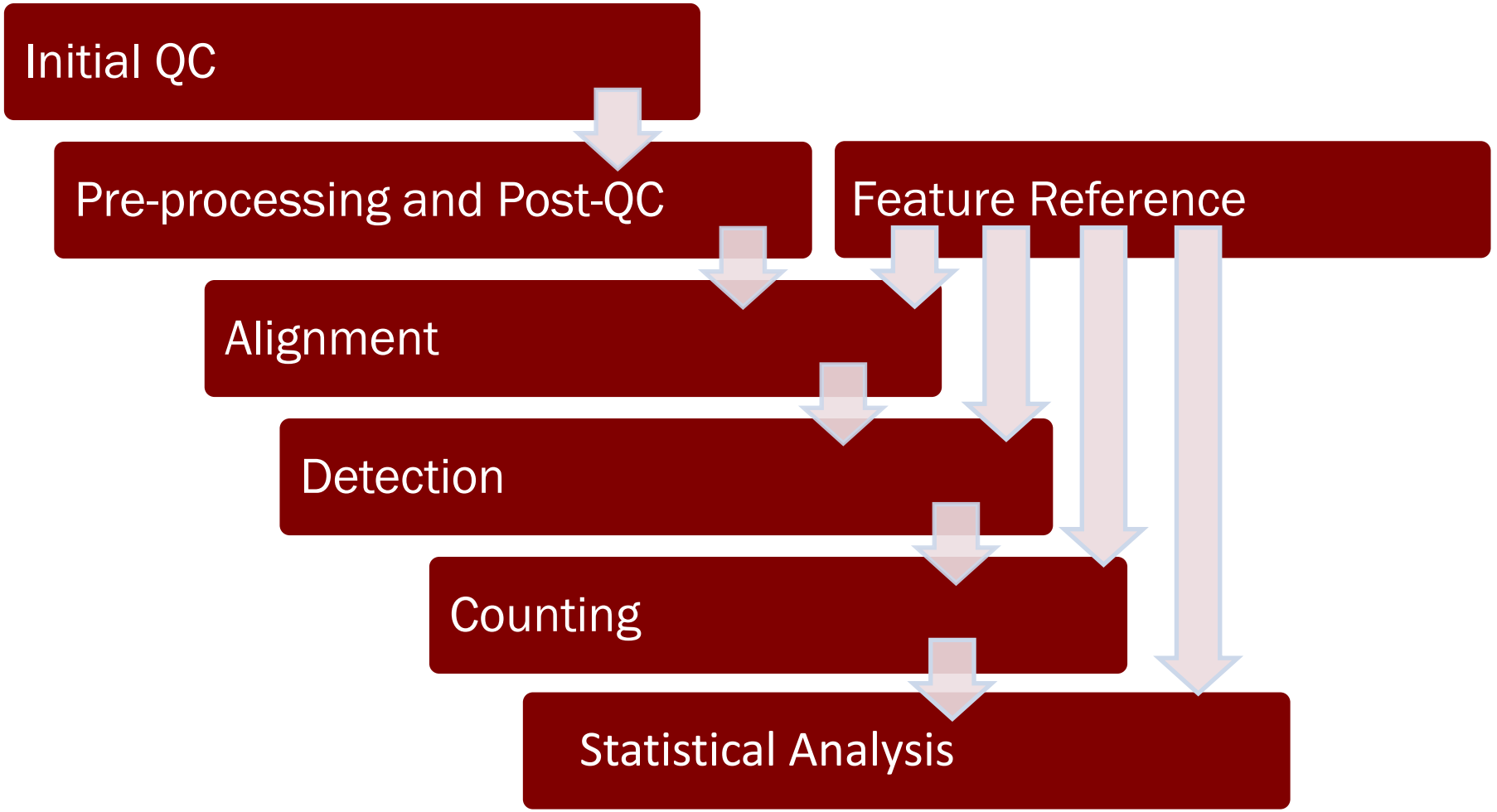
SMN1



SMN2



NGS ANALYSIS WORKFLOW



EVALUATE DIFFERENCES

- ✓ Gene Expression: Estimate significance of fold changes between conditions.
- ✓ Chip-Seq: Estimate significance of fold enrichments above a background or negative control.

EVALUATE DIFFERENCES

Tools to evaluate differences in RNASeq:

- edgeR
- DESeq2
- Cuffdiff
- Sleuth
- EBSeq

Issues they solve:

- Overestimation of Dispersion due to low “n”
- Not in the range of Central Limit to consider Normal Distribution.
- Correct for Multiple testing

ANALYSIS PIPELINE FOR DIFFERENTIAL GENE/TRANSCRIPT EXPRESSION USING RNA-SEQ