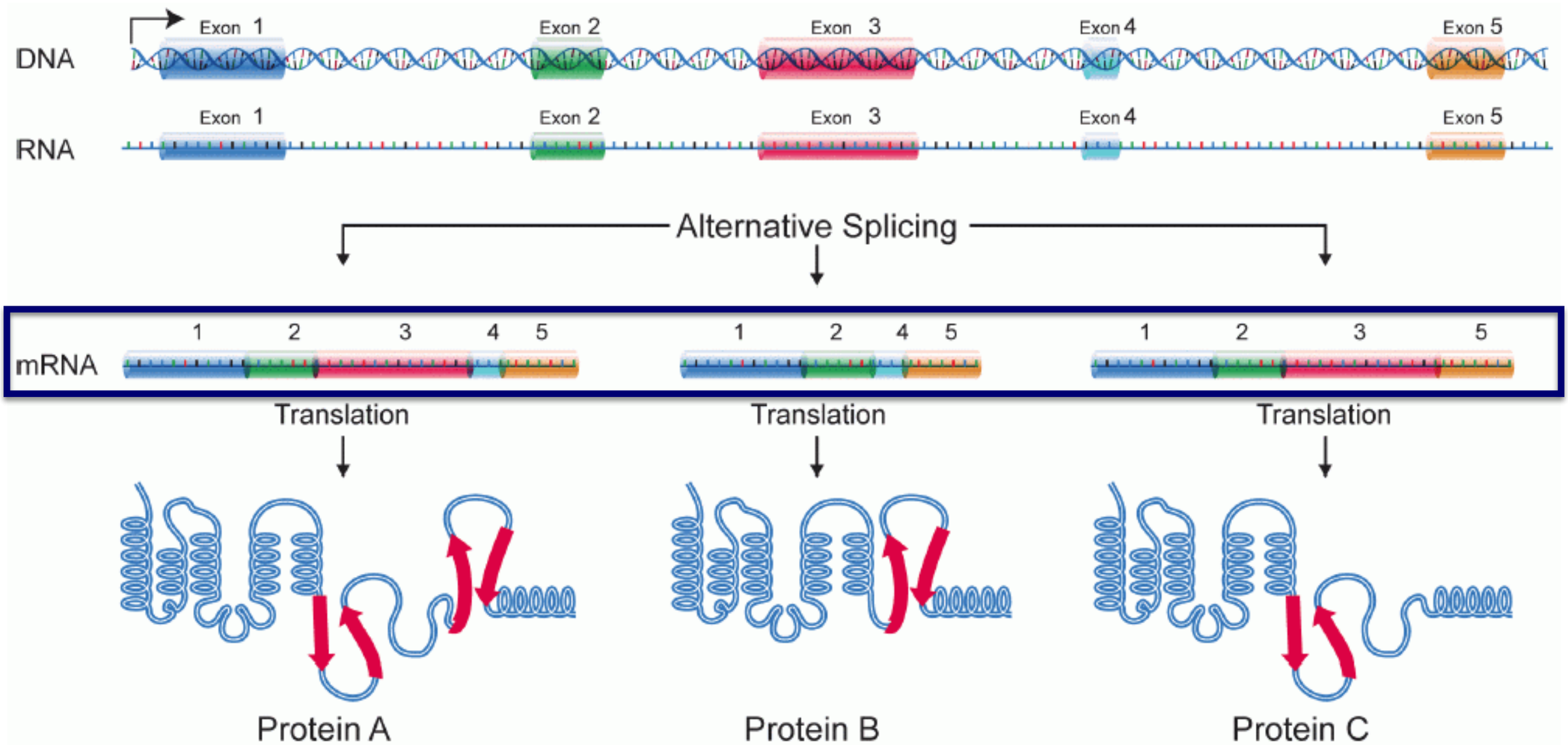
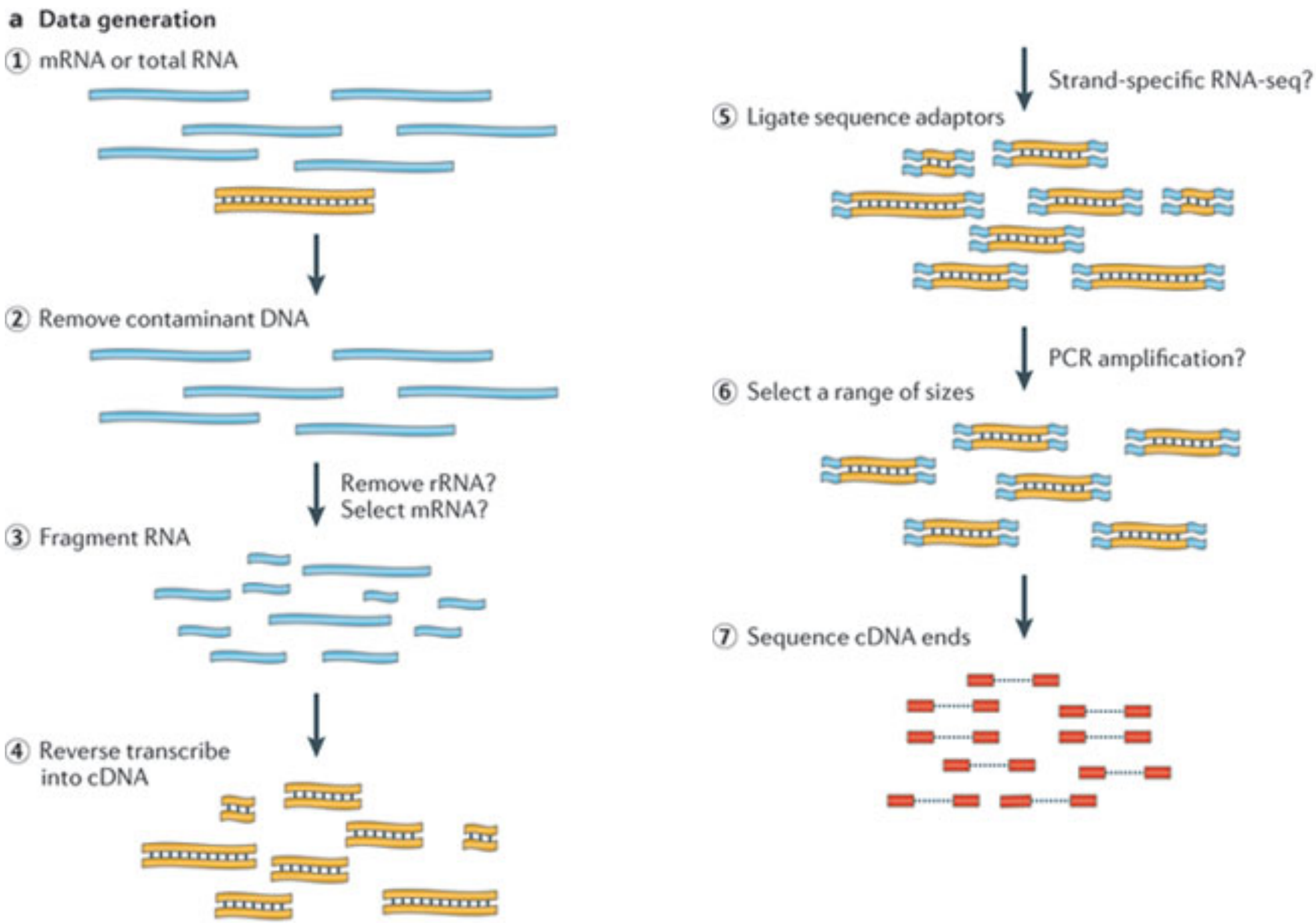
Alignment-based RNA-seq quantification

Charlotte Soneson
University of Zurich
Brixen 2017


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Wikipedia

Sequencing



Martin & Wang, Nat. Rev. Genetics (2011)

Single- vs paired-end sequencing

- Each fragment can be sequenced from one end only, or from both ends
- Single-end cheaper and faster
- Paired-end provide improved ability to localize the fragment in the genome and resolve mapping close to repeat regions - less multimapping reads

Experiment: SRX749151

Illumina HiSeq 2000 sequencing; E15 Cortex RNA-seq

View: XML **Download:** XML

Submitting Centre	Platform	Model		
Karolinska Institutet	ILLUMINA	Illumina HiSeq 2000		
Library Layout	**Library Strategy**	**Library Source**	**Library Selection**	**Library Name**
SINGLE	RNA-Seq	TRANSCRIPTOMIC	PolyA	E15 Cortex RNA-seq
Broker Name				
NCBI				

Experiment: SRX547157

Illumina HiSeq 2000 paired end sequencing; HKCI-3 RNA-Seq

View: XML **Download:** XML

Submitting Centre	Platform	Model		
The Chinese University of Hong Kong	ILLUMINA	Illumina HiSeq 2000		
Library Layout	**Library Strategy**	**Library Source**	**Library Selection**	**Library Name**
PAIRED	RNA-Seq	TRANSCRIPTOMIC	unspecified	HKCI-3 RNA-Seq
Broker Name				
NCBI				

Description

Genomic DNA was sequenced using Illumina HiSeq 2000 instruments following the manufacturer's standard protocols. Illumina sequencing libraries were constructed for paired-end sequencing (with an insert size of ~500 bp), paired-end sequencing was performed for the whole genome by Illumina HiSeq 2000 sequencing, at 2 × 100 bp runs.

Strand-specificity

- In “standard” protocols, we don’t know from which strand a read stems
- Various “strand-specific” protocols allow us to keep this information
- Strand-specificity leads to lower number of ambiguous reads (overlapping multiple genes)

RESEARCH ARTICLE | OPEN ACCESS

Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols

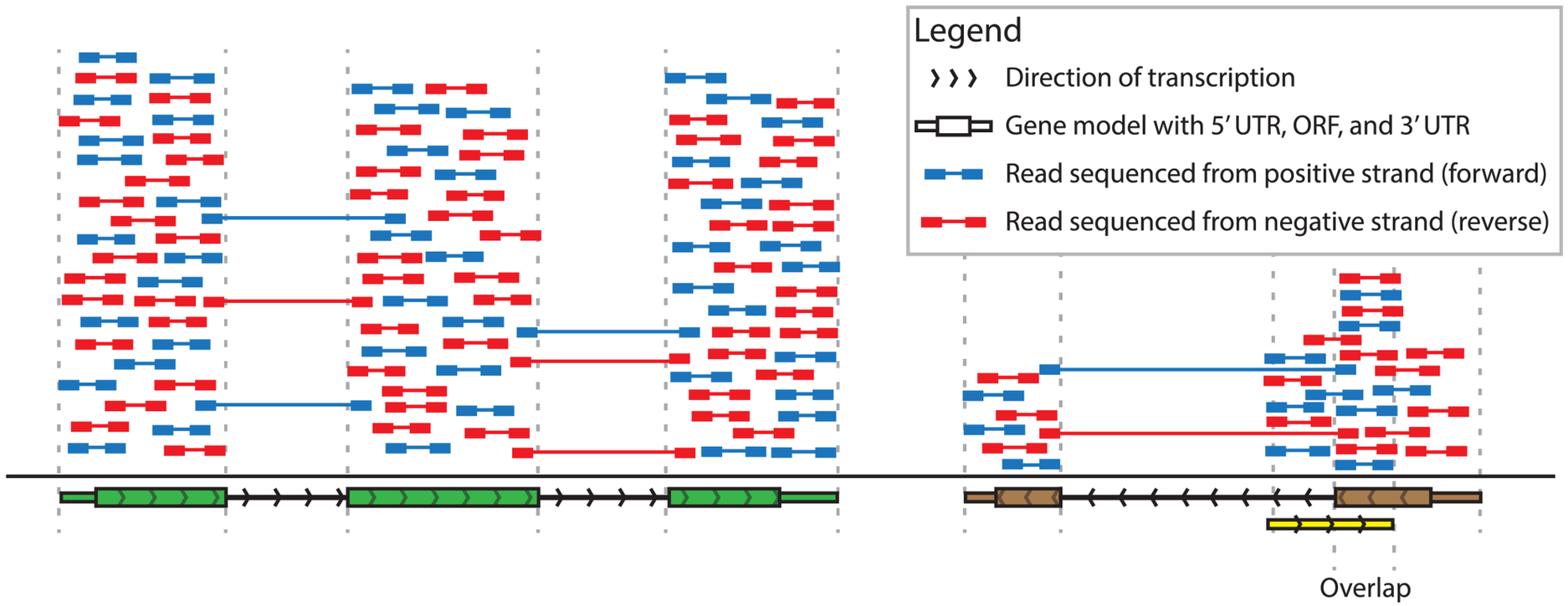
Susan M. Corley, Karen L. MacKenzie, Annemiek Beverdam, Louise F. Roddam and Marc R. Wilkins

BMC Genomics 2017 18:399 | DOI: 10.1186/s12864-017-3797-0 | © The Author(s). 2017 ReadCube

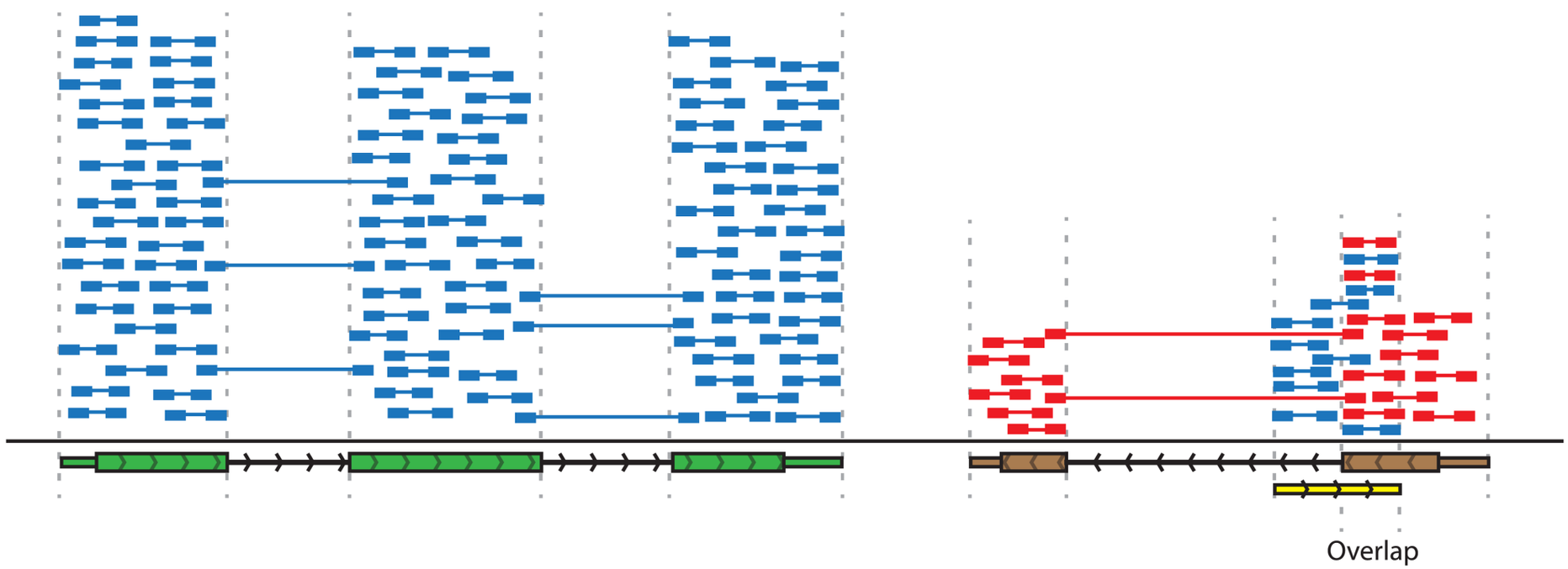
Received: 16 December 2016 | Accepted: 16 May 2017 | Published: 23 May 2017

Strand-specificity

A.



B.



https://doi.org/10.1371/journal.pcbi.1004393.g006

Raw reads - FASTQ format

- Combines sequence and base quality information
- Four lines per sequence (read)
  - ID line (starting with @)
  - sequence
  - another ID line (starting with +)
  - base qualities
- For paired-end sequencing: one file for “first” reads and one for “second” reads

FASTQ format - sequence ID line

```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

- D7MHBFNI - unique instrument name
- 202 - run ID
- D1BUDACXX - flowcell ID
- 4 - flowcell lane
- 1101 - tile number within lane
- 1340 - x-coordinate of cluster within tile
- 1967 - y-coordinate of cluster within tile
- 1 - member of pair (1 or 2). Older versions: /1 and /2
- Y/N - whether the read failed quality control (Y = bad)
- 0 - none of the control bits are on
- CATGCA - index sequence (barcode)

FASTQ format - base qualities

- For each letter, estimate the probability of being erroneous (p)
- Phred score $Q = -10 \cdot \log_{10}(p)$

Phred score	Probability of incorrect base call	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%

Quality format encoding

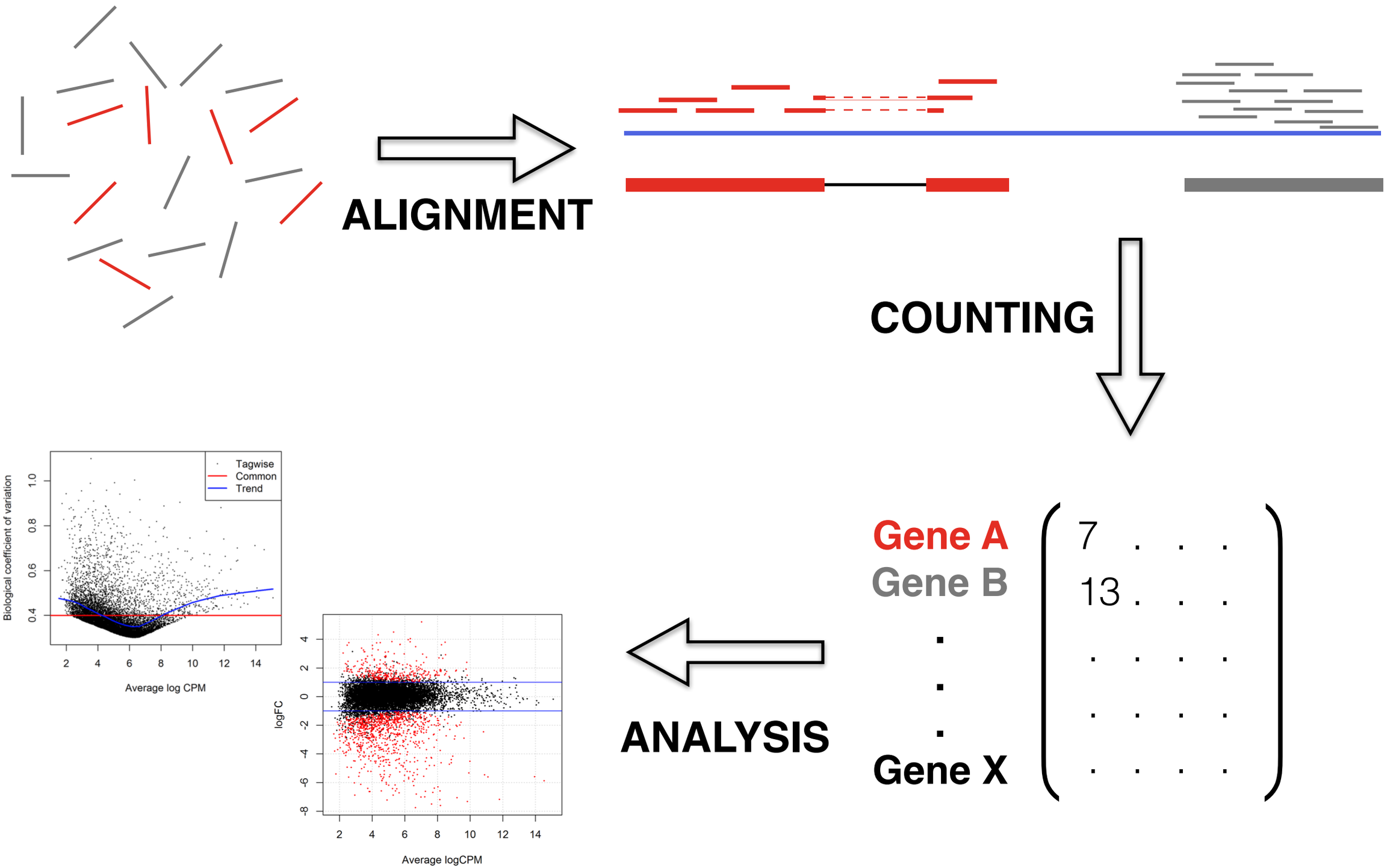
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                     3.....9.............................40
0........................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

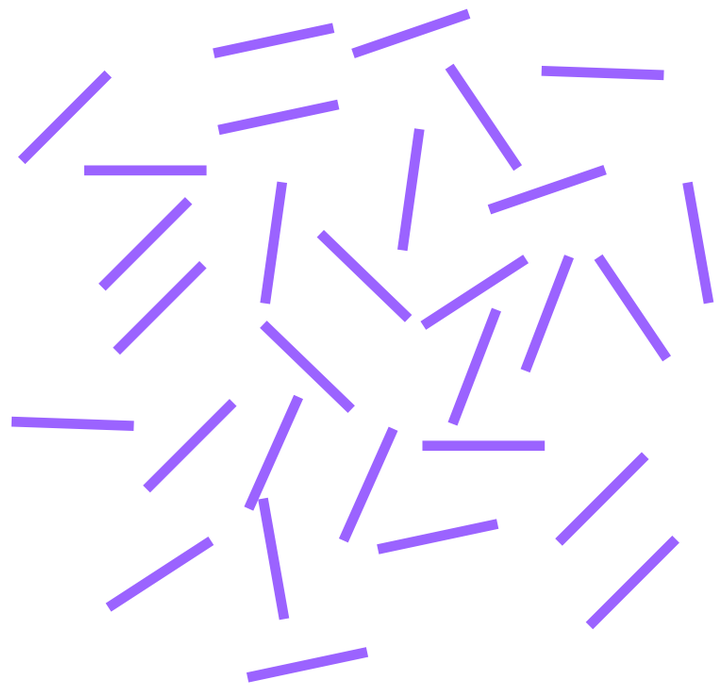
“Capital letters = good quality” (with Illumina 1.8+)

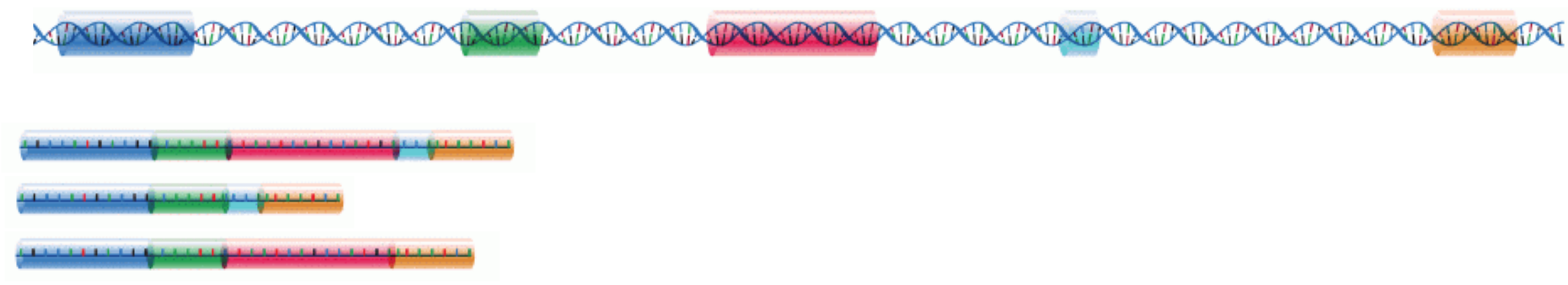
http://www.wikipedia.com/FASTQ_format

The alignment-based workflow



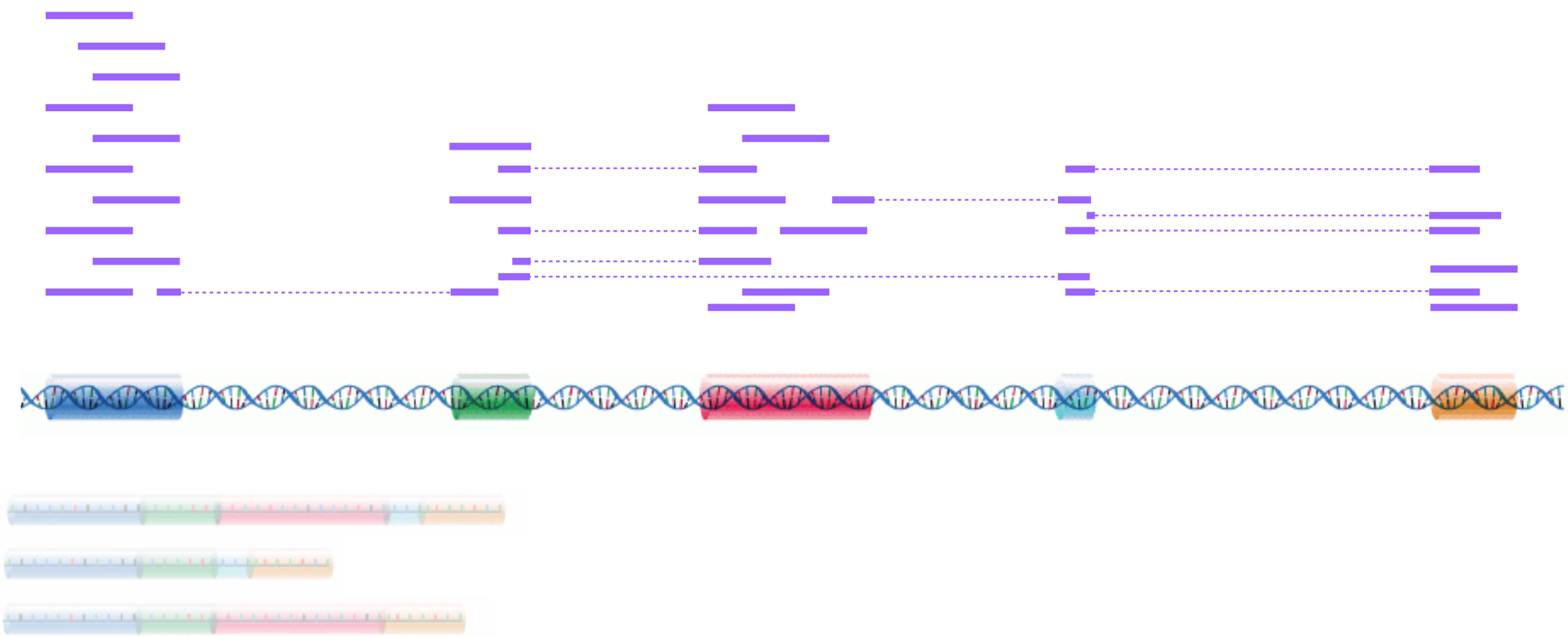
Abundance quantification





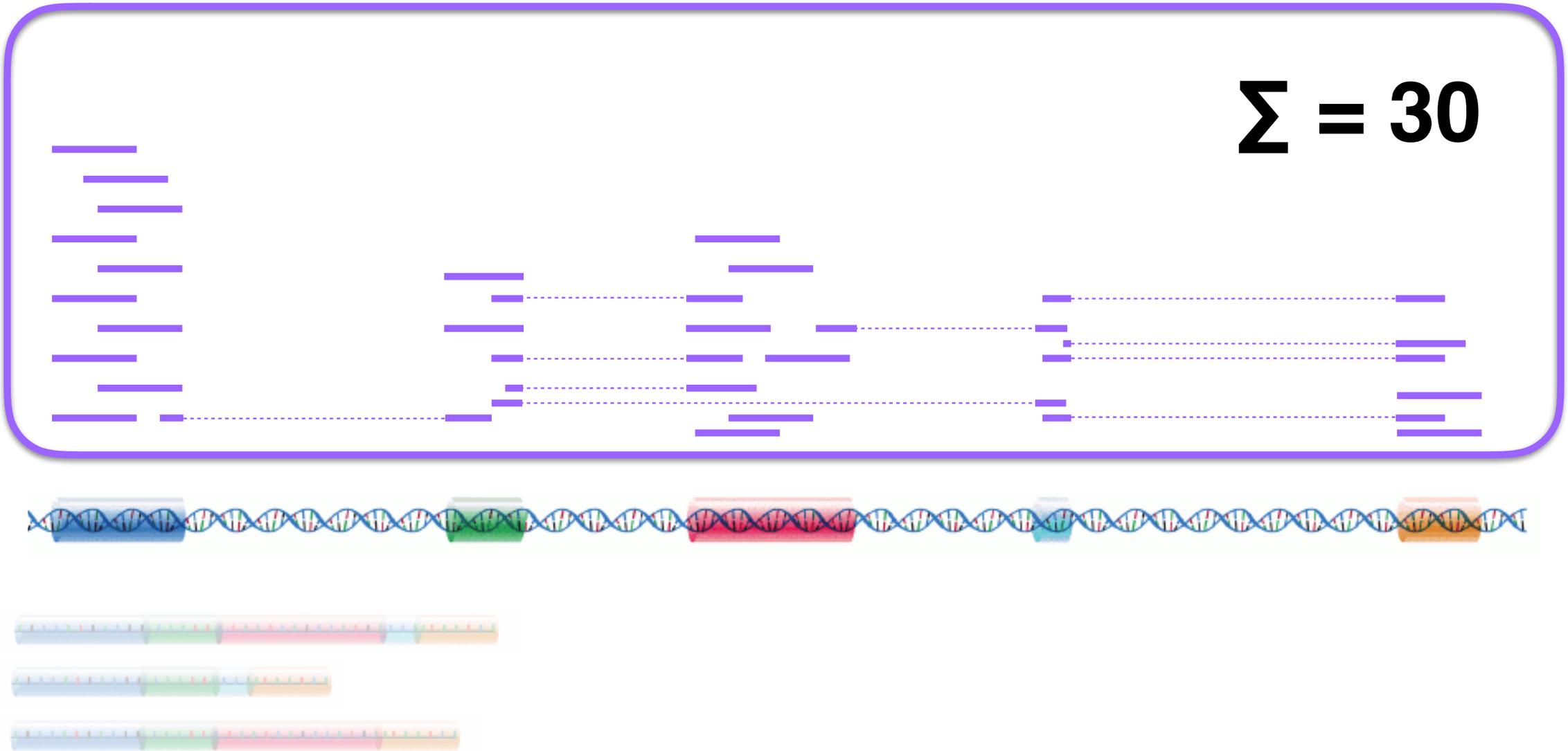
Abundance quantification

Gene-level counts, often obtained by genome alignment + overlap counting



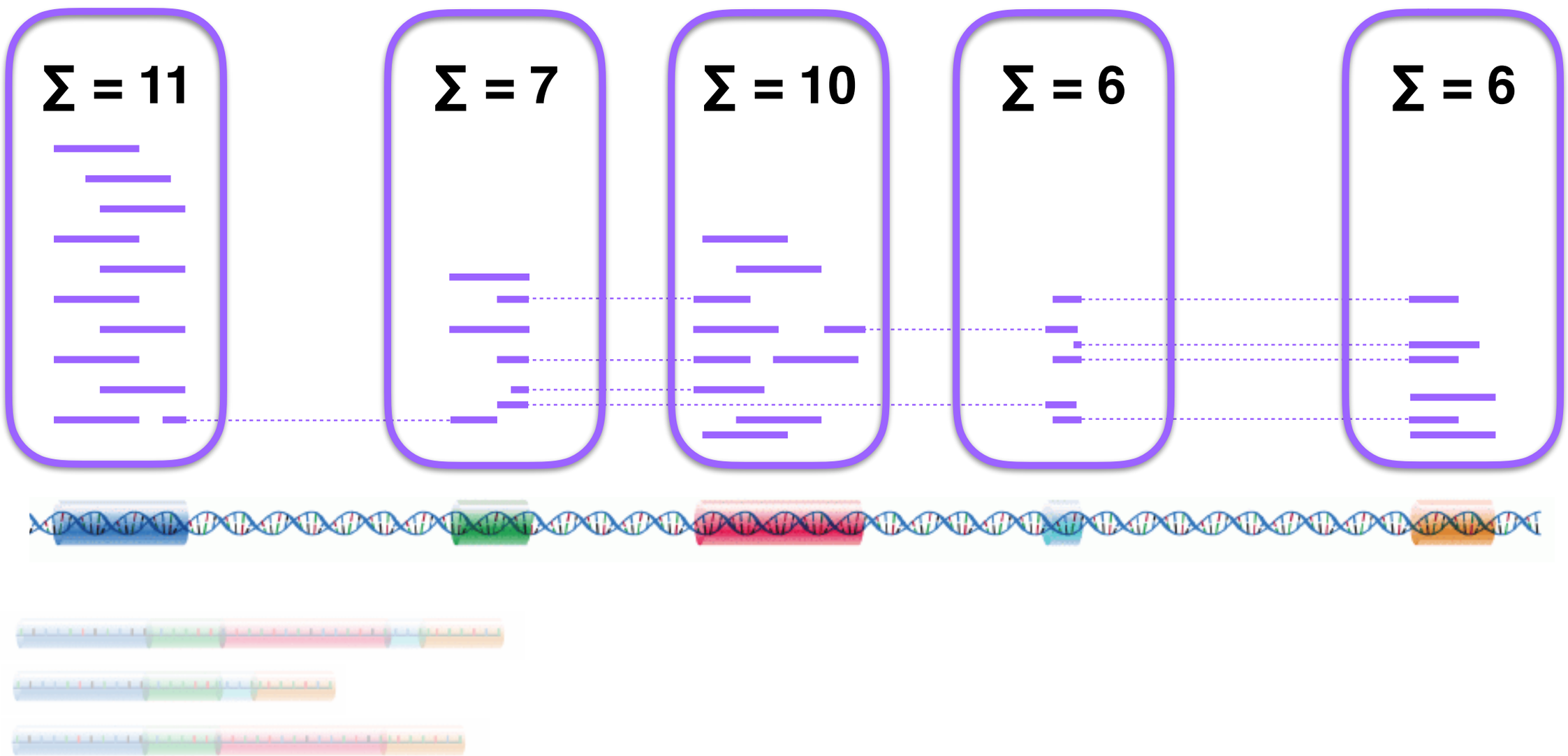
Abundance quantification

Gene-level counts, often obtained by genome alignment + overlap counting



Abundance quantification

Exon-level counts, often obtained by genome alignment + overlap counting



The (human) reference genome

- A “representative example” of the human genome sequence
- New versions are released periodically (the latest, GRCh38, in December 2013)
- Coordinates are not comparable across versions

The reference genome

- Typically provided as a **fasta** file - general sequence representation
- Two lines per sequence (e.g., chromosome)
  - Header line (starting with >)
  - Sequence

```
>chr1
......
GTTCTTGTTCTGTGTTCTTATAACCATACCAGAATTTTCTTCATCACAGA
CAGAGACTAAACTCTTTCTTCTCTTACCTTTCCTTTGATAATATTTTTGA
TCCAGGAATGGGGATAATTTTGCAGTTAAAATTTTCTTTTTATGATGGAA
GGTGAGGAGGAGAGAGAGGTTTACATTAGAAGTGACCCAACTCCATTTTC
TTCCAATGGTTTTTTTCAGTTTTATTTTTTTAAAGCGTGAACAGAGAATA
GTCACCTGATCAATTTAAATATGTCAAAAAGTGAAAGAAAAATCTCTCTT
TTAAAGGAAATGAGGGCAGTAACACAACCAAGGAATCAAAATTCAGGTTG
AGGCTGACCTTTGACCTGCAACTATGCTACTCCATGAACAGCAAGTAGGA
AATGGCTGATTTCATGAAGGTGGACTGGCATCAGAGGAGGCGAGGGATCC
AGGGTTCCTGATGAGTGGCAACATTCCTTGGTCTTTTGAGTTTGTTTGAT
TGGTGAATCAAATTTAGGTGACAGCCAGCTAAAGAGAGTGAGGGTGGCTG
TCTTGTGAATGGGAAGTGACCAAGCTTGAAAGCACAGACTgtggtggctc
......
```

The reference genome

www.ensembl.org/info/data/ftp/index.html

Single species data

Popular species are listed first. You can customise this list via our home page.

Show 10 entries Show/hide columns

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Varia (G
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G

The reference genome

Homo_sapiens.GRCh38.dna.chromosome.21.fa.gz	11.2 MB	06/03/2017, 20:21:00
Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz	10.9 MB	06/03/2017, 20:17:00
Homo_sapiens.GRCh38.dna.chromosome.3.fa.gz	57.0 MB	06/03/2017, 17:53:00
Homo_sapiens.GRCh38.dna.chromosome.4.fa.gz	54.7 MB	06/03/2017, 18:13:00
Homo_sapiens.GRCh38.dna.chromosome.5.fa.gz	52.0 MB	06/03/2017, 18:30:00
Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz	49.1 MB	06/03/2017, 18:40:00
Homo_sapiens.GRCh38.dna.chromosome.7.fa.gz	45.2 MB	06/03/2017, 18:47:00
Homo_sapiens.GRCh38.dna.chromosome.8.fa.gz	41.6 MB	06/03/2017, 19:00:00
Homo_sapiens.GRCh38.dna.chromosome.9.fa.gz	34.8 MB	06/03/2017, 19:05:00
Homo_sapiens.GRCh38.dna.chromosome.MT.fa.gz	5.4 kB	06/03/2017, 20:21:00
Homo_sapiens.GRCh38.dna.chromosome.X.fa.gz	44.0 MB	06/03/2017, 18:54:00
Homo_sapiens.GRCh38.dna.chromosome.Y.fa.gz	6.8 MB	06/03/2017, 20:14:00
Homo_sapiens.GRCh38.dna.nonchromosomal.fa.gz	2.9 MB	06/03/2017, 15:27:00
Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz	840 MB	06/03/2017, 20:23:00
Homo_sapiens.GRCh38.dna.toplevel.fa.gz	1013 MB	06/03/2017, 20:22:00
Homo_sapiens.GRCh38.dna_rm.alt.fa.gz	156 MB	06/03/2017, 16:25:00
Homo_sapiens.GRCh38.dna_rm.chromosome.1.fa.gz	36.5 MB	06/03/2017, 16:49:00
Homo_sapiens.GRCh38.dna_rm.chromosome.10.fa.gz	21.7 MB	06/03/2017, 19:19:00
Homo_sapiens.GRCh38.dna_rm.chromosome.11.fa.gz	20.9 MB	06/03/2017, 19:14:00
Homo_sapiens.GRCh38.dna_rm.chromosome.12.fa.gz	20.6 MB	06/03/2017, 19:27:00
Homo_sapiens.GRCh38.dna_rm.chromosome.13.fa.gz	16.1 MB	06/03/2017, 19:35:00
Homo_sapiens.GRCh38.dna_rm.chromosome.14.fa.gz	14.4 MB	06/03/2017, 19:42:00
Homo_sapiens.GRCh38.dna_rm.chromosome.15.fa.gz	13.5 MB	06/03/2017, 19:49:00
Homo_sapiens.GRCh38.dna_rm.chromosome.16.fa.gz	12.9 MB	06/03/2017, 19:54:00
Homo_sapiens.GRCh38.dna_rm.chromosome.17.fa.gz	13.0 MB	06/03/2017, 19:59:00
Homo_sapiens.GRCh38.dna_rm.chromosome.18.fa.gz	12.7 MB	06/03/2017, 20:05:00

Locations of genes on reference genome

- Typically provided in a **gtf** (gene transfer format) file
- Similar to **gff**, but more restrictive

```
seqname  source  feature  start  end  score  strand  frame  attribute
```

```
2R    protein_coding   exon  5139815    5141712    .    -    .    gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
exon_id "FBgn0020621:1";
2R    protein_coding   CDS   5141572    5141712    .    -    0    gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
protein_id "FBpp0111810";
2R    protein_coding   stop_codon 5141569    5141571    .    -    0    gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
```

http://www.ensembl.org/info/website/upload/gff.html

The gene coordinates

www.ensembl.org/info/data/ftp/index.html

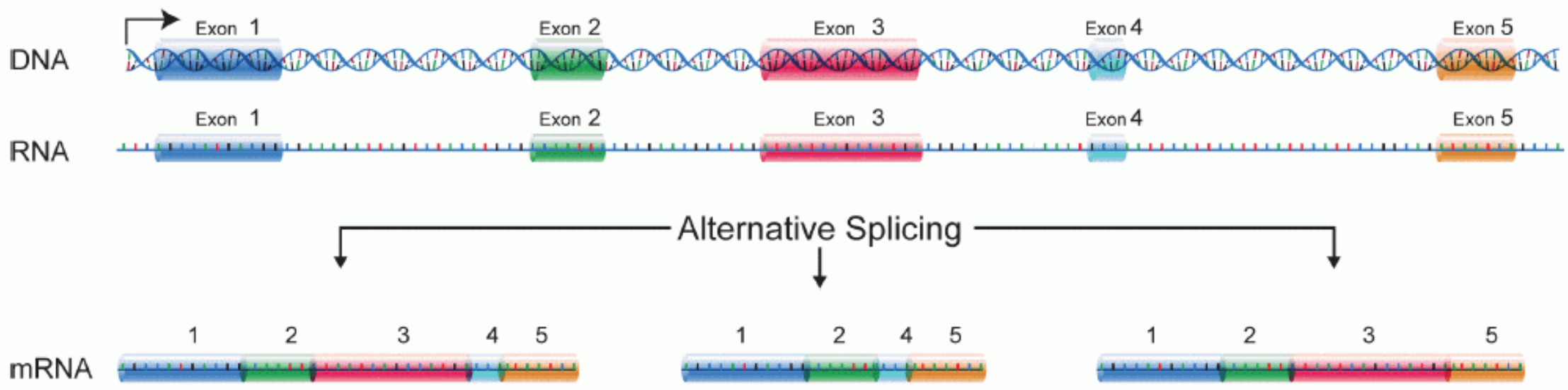
Single species data

Popular species are listed first. You can customise this list via our home page.

Show 10 entries Show/hide columns

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Varia (G
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
Y	**Mouse** *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G
Y	**Zebrafish** *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	G

Aligning RNA-seq reads



- Need a splice-aware aligner
- Common choices:
    - STAR
    - HiSAT2
    - TopHat2

STAR - step 1: indexing the genome

number of threads

output folder - name according to genome

```
$ STAR --runThreadN 24 \
       --runMode genomeGenerate \
       --genomeDir my_genome \
       --genomeFastaFiles my_genome.fa \
       --sjdbGTFfile my_genes.gtf \
       --sjdbOverhang 99
```

read length - 1

STAR - step 2: aligning the reads

created index

```
$ STAR --runThreadN 24 \
       --runMode alignReads \
       --genomeDir my_genome \
       --readFilesIn S1_read1.fq.gz \
         S1_read2.fq.gz \
       --readFilesCommand zcat \
       --outFileNamePrefix output/S1/ \
       --outSAMtype BAM SortedByCoordinate \
       --quantMode GeneCounts
```

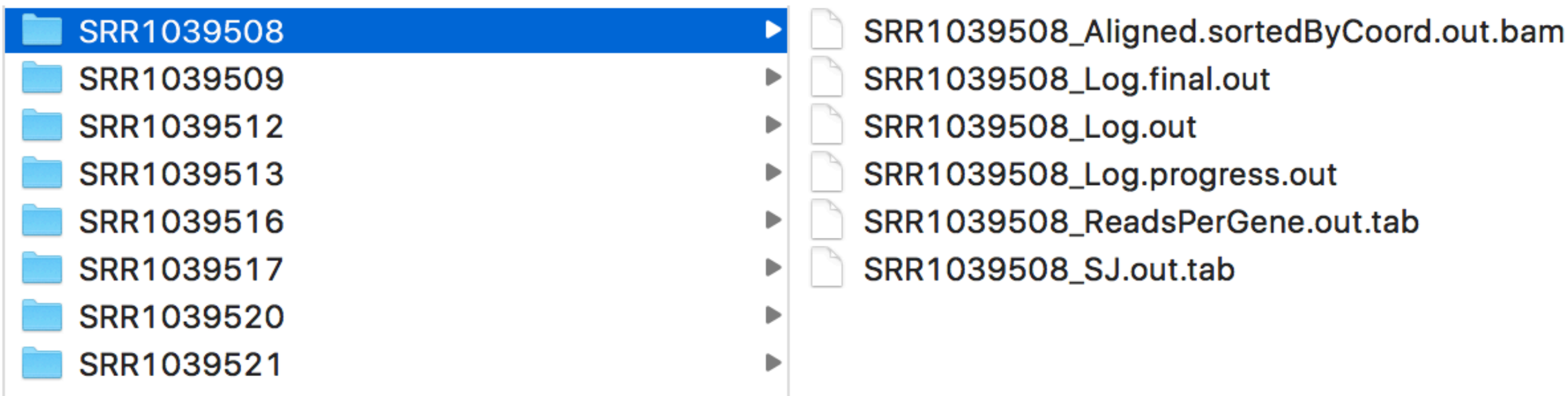
read file(s)

[include sample ID]

count reads

for compressed read files

STAR - output


SRR1039508
SRR1039509
SRR1039512
SRR1039513
SRR1039516
SRR1039517
SRR1039520
SRR1039521
SRR1039508_Aligned.sortedByCoord.out.bam
SRR1039508_Log.final.out
SRR1039508_Log.out
SRR1039508_Log.progress.out
SRR1039508_ReadsPerGene.out.tab
SRR1039508_SJ.out.tab

Representing alignments - SAM format

- Header

```
@SQ  SN:chr1  LN:249250621
@SQ  SN:chr2  LN:243199373
@SQ  SN:chr3  LN:198022430
@SQ  SN:chr4  LN:191154276
```

- Body

```
seq.13906018  0    chr10    101948233 255  101M *    0    0
GTCCACAGTCCTTTCTCTGAAACCCTTGGGNNAAGTTGTTTCAGAATTANGNAA  CBCFFFFFHHHHHJJJJJJJJJJJJJJJJJ##11?
DHIIIIJJHIJJJJ#0#07    0L:A:F    IH:i:1    HI:i:1
```

- Typically, one line per alignment
- BAM = binary SAM

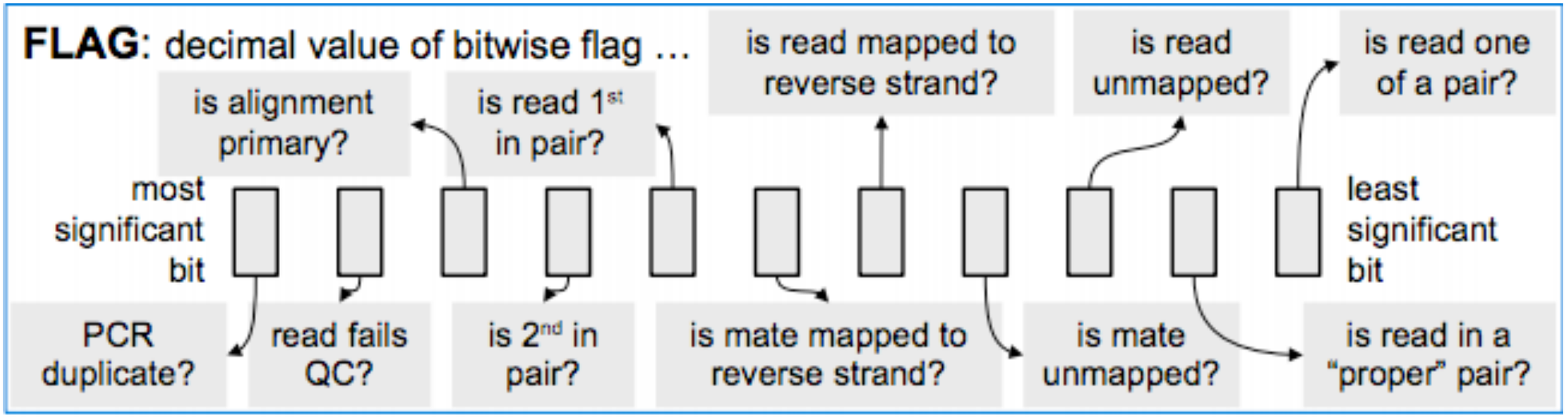
Representing alignments - SAM format

```
seq.13906018  0    chr10    101948233 255  101M *    0    0
GTCCACAGTCCTTTCTCTGAAACCCTTGGGNNAAGTTGTTTCAGAATTANGNAA  CBCFFFFFHHHHHJJJJJJJJJJJJJJJJJ##11?
DHIIIIJJHIJJJJ#0#07    0L:A:F    IH:i:1    HI:i:1
```

- Column 1 - sequence ID
- Column 2 - flag. Ex:
    - 0 - non-paired read, mapping to forward strand
    - 16 - non-paired read, mapping to reverse strand
    - 4 - unmapped read
- Column 3 - reference sequence name for the alignment
- Column 4 - position of alignment
- Column 5 - mapping quality
    - 255 - not available
    - 0 - multiple best hits
- Column 6 - CIGAR string
- Column 7-8 - reference name/position of mate/next segment
- Column 9 - observed template length
- Column 10 - sequence (represented as mapped on the reference (forward) strand!)
- Column 11 - base quality
- Remaining columns are optional, and are of the type TAG:TYPE:VALUE

http://samtools.sourceforge.net/SAM1.pdf

The SAM flag



- ex: 83 = 00001010011 = first in pair, read on reverse strand, part of properly mapped pair

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

The CIGAR string

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- Describes the mapping in more detail
- See also the MD tag

The CIGAR string - example

```
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: ACTAGAATGGCT
```

Aligning these two:

```
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
```

With the alignment above, you get:

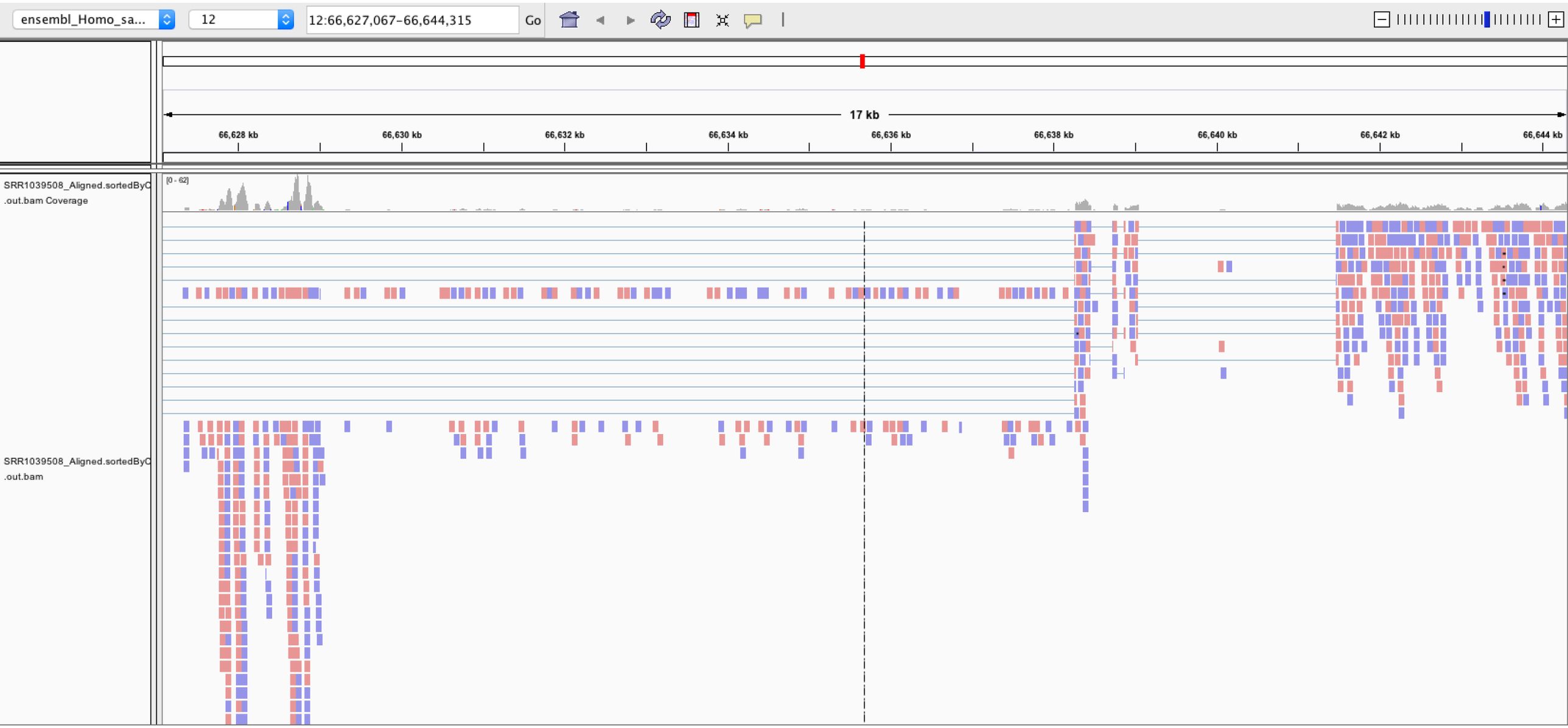
```
POS: 5
CIGAR: 3M1I3M1D5M
```

http://genome.sph.umich.edu/wiki/SAM

Working with SAM/BAM files

- SAMtools
  - convert between SAM/BAM
  - sort/index
  - view alignments
  - ...
  - R interface in the `Rsamtools` package

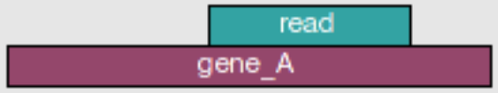
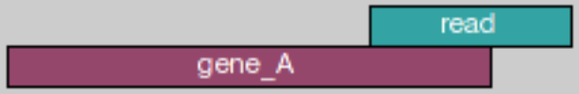
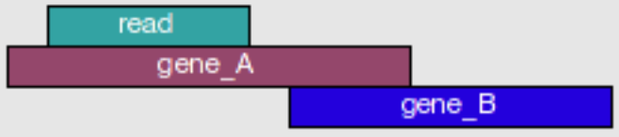
Visualizing alignments - IGV



Estimating abundances via overlap counting

- `STAR`
- `HTseq-count` (Python)
- `Rsubread::featureCounts` (R)
- `GenomicAlignments::summarizeOverlaps` (R)

Counting modes

	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

http://www-huber.embl.de/HTSeq/doc/count.html

HTSeq-count

linked to output sorting by STAR

default=yes

check your GTF file!

```
$ htseq-count --format=bam \
              --order=pos \
              --stranded=no \
              --type=exon \
              --idattr=gene_id \
              --mode=union \
              aligned.bam \
              my_genes.gtf
```

```
2R    protein_coding    exon    5139815    5141712    .    -    .    gene_id "FBgn0020621"; transcript_id
"FBtr0112897"; exon_number "10"; gene_name "Pkn"; gene_biotype "protein_coding"; transcript_name "Pkn-RG";
exon_id "FBgn0020621:1";
```

featureCounts

```
> featureCounts(files = bamfiles,
                annot.ext = “my_genes.gtf”,
                isGTFAnnotationFile = TRUE,
                GTF.featureType = “exon”,
                GTF.attrType = “gene_id”,
                useMetaFeatures = TRUE,
                isPairedEnd = TRUE,
                strandSpecific = 0)
```

check your GTF file!

The alignment-based workflow