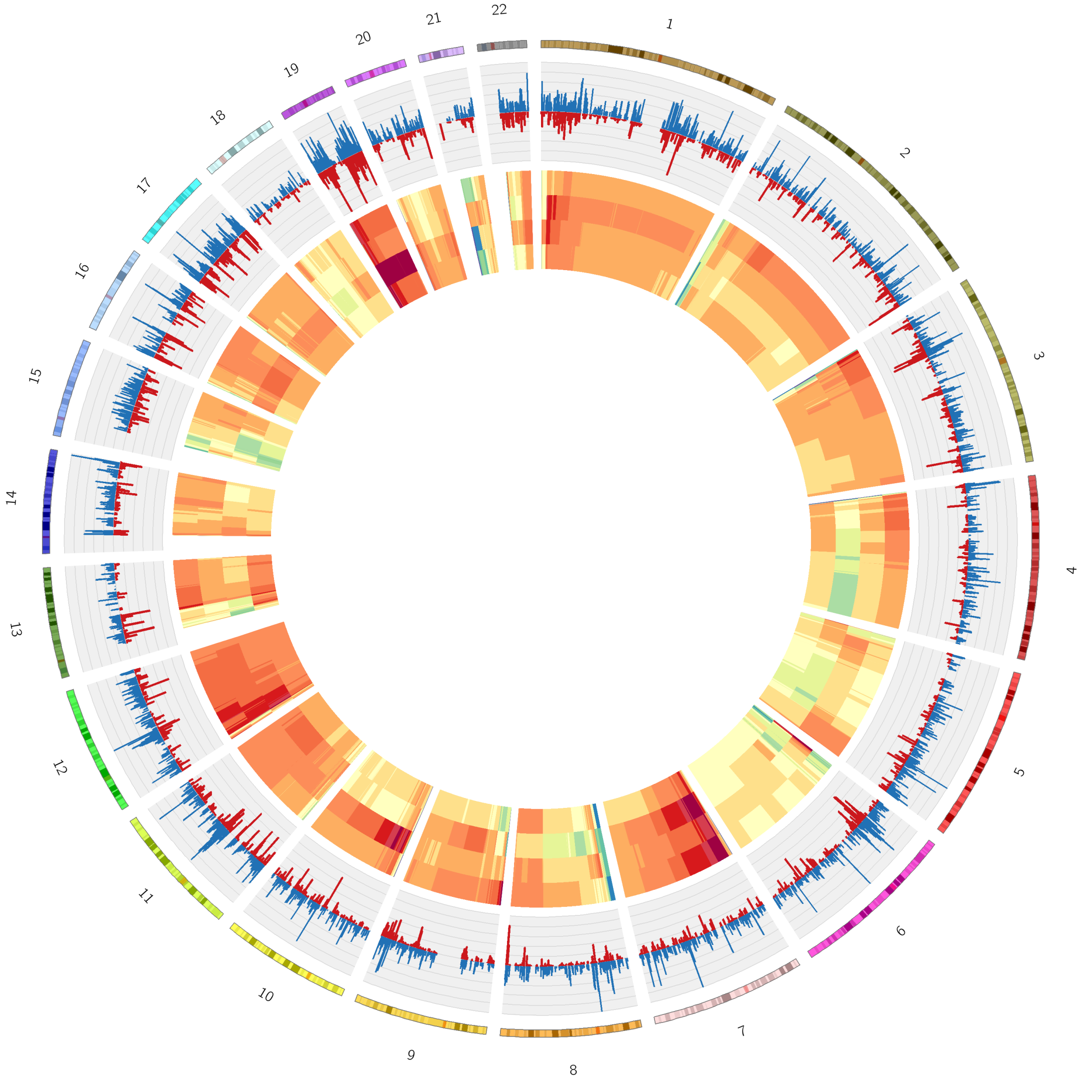


Short Reads Alignment to a Reference Genome

Joanna Krupka

CRUK Summer School in Bioinformatics

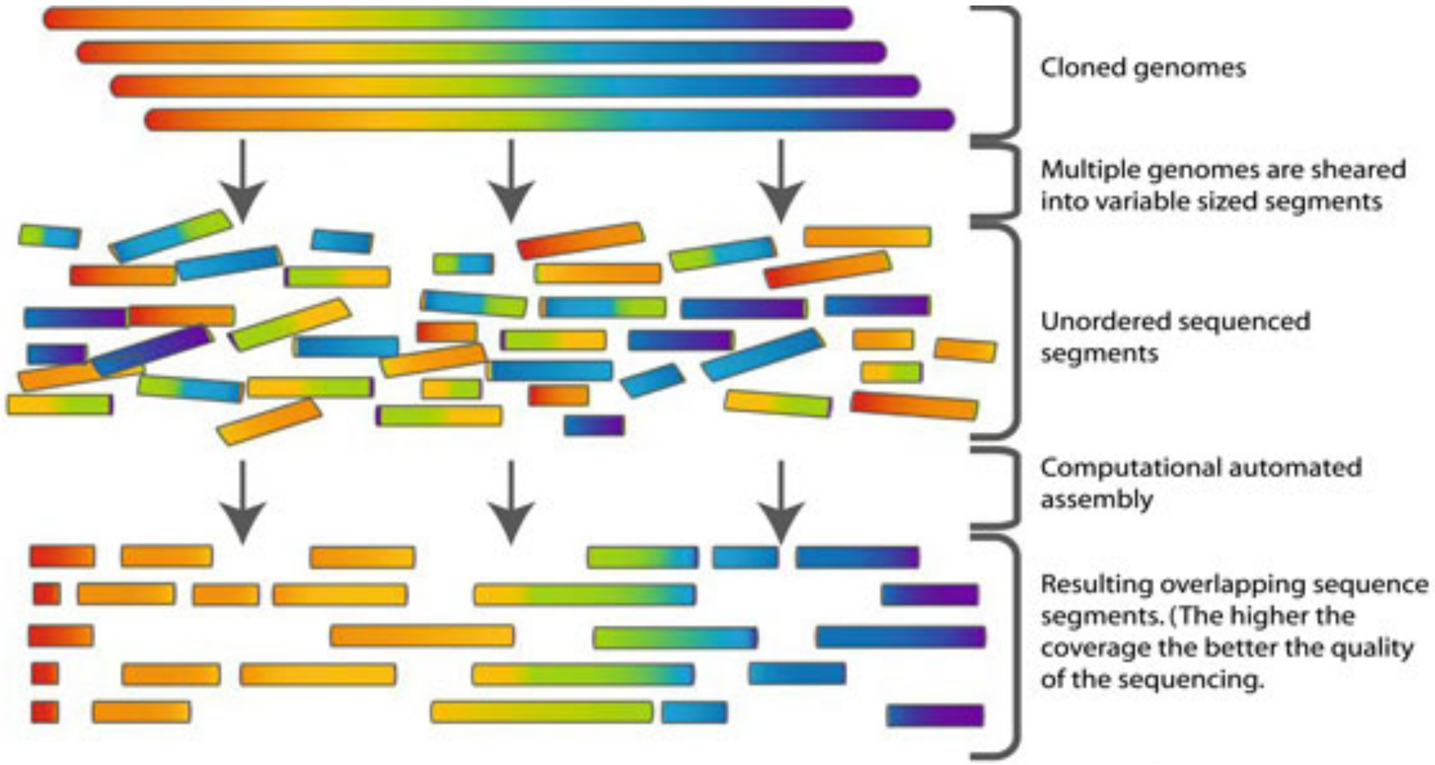
Cambridge, July 2020

CANCER RESEARCH UK

MRC Cancer Unit

UNIVERSITY OF CAMBRIDGE

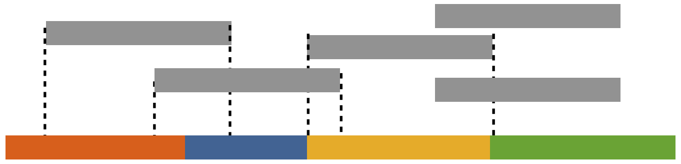
Shotgun Sequencing and sequence assembly approaches


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.

Commins J. et al, Biol Proced Online 11(1) 2015

Mapping to reference sequence

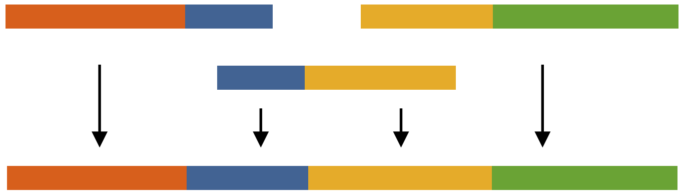
Recreate the genome with using prior knowledge as reference



Mapping is as good as reference used

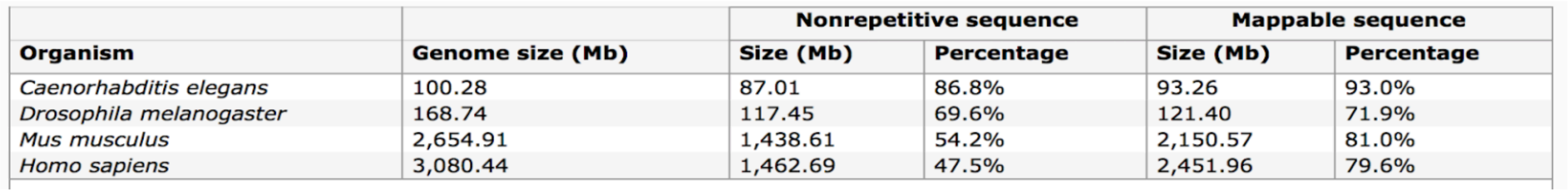
De Novo assembly

Recreate the genome with no prior knowledge



Problem with repeated regions, high coverage and long reads required

Mappability

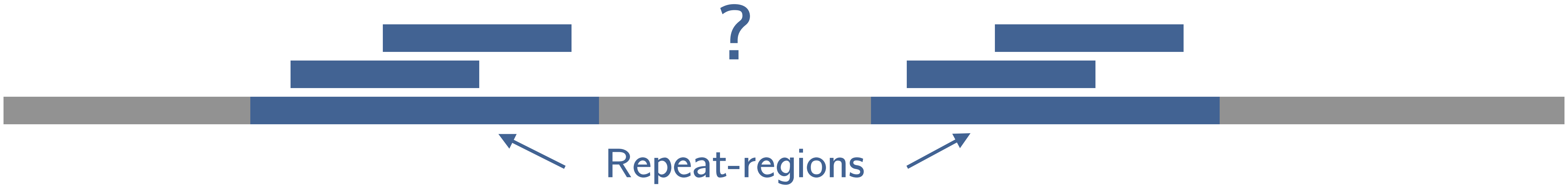


		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

Rozowsky J. Et al. Nat Biotechnol 2009

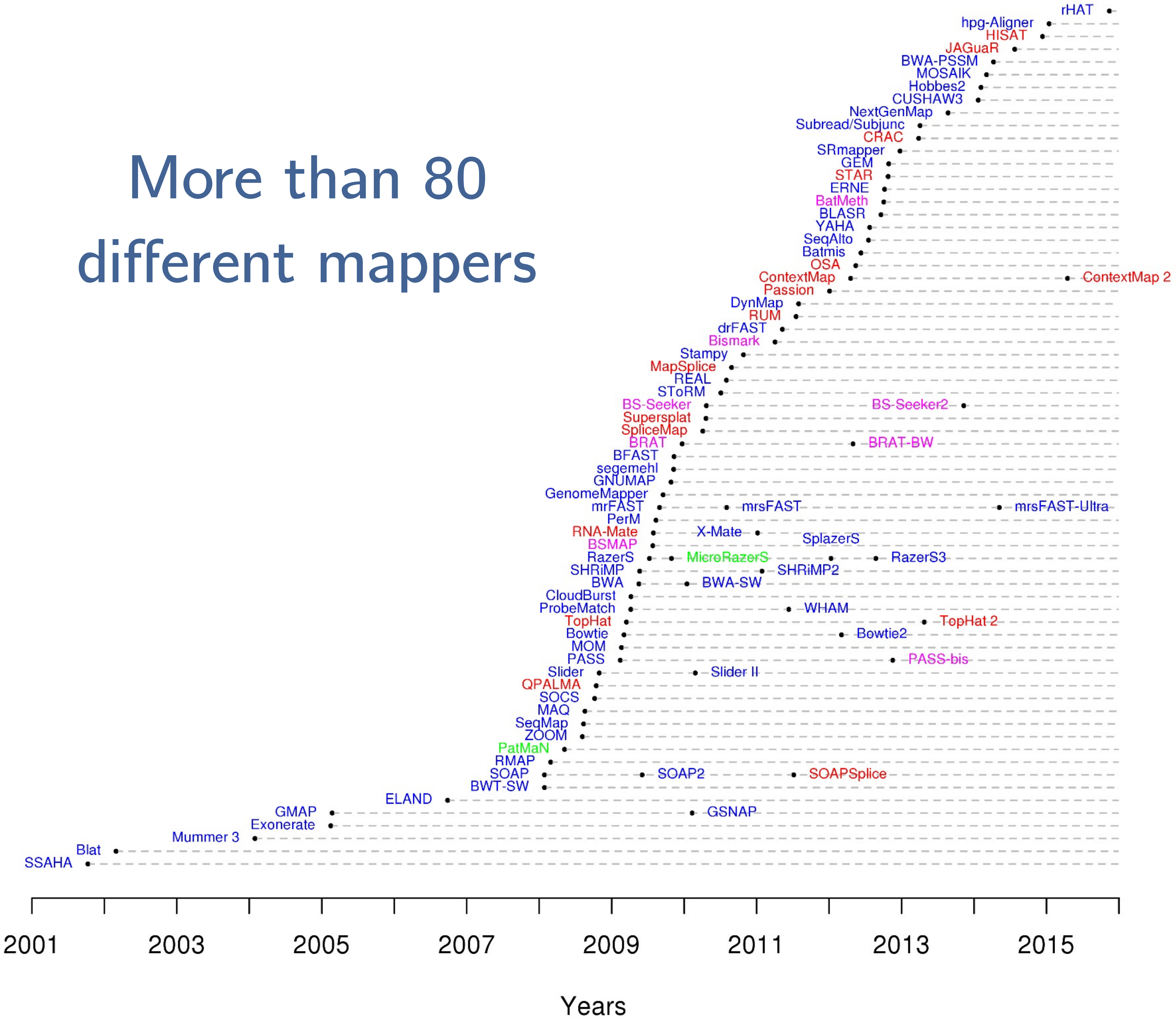
Mappability (or uniqueness) is a measure of the ability of aligning the short reads to a unique location in the reference genome.

Mapping uncertainty if the reads are shorter than a repeat region



Short sequence mapping tools

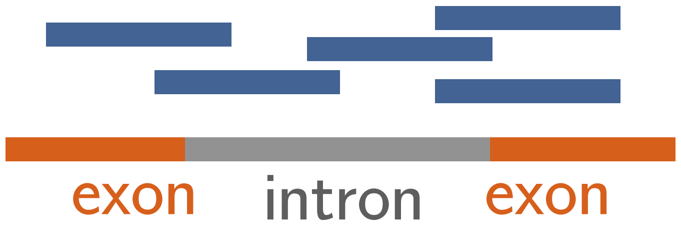
More than 80 different mappers



https://www.ecseq.com/support/ngs/what-is-the-best-ngs-alignment-software

Short sequence mapping tools

eg. Whole Genome Sequencing

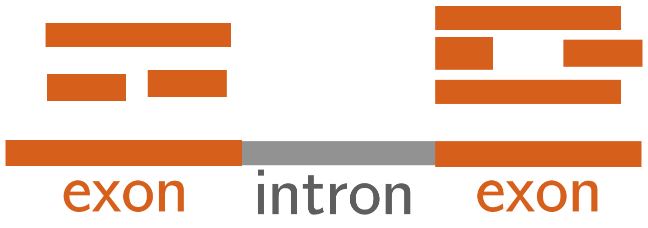


Not splice aware

Bowtie2

BWA

eg. RNA-Seq



Splice aware

STAR

TopHat2

Hisat2

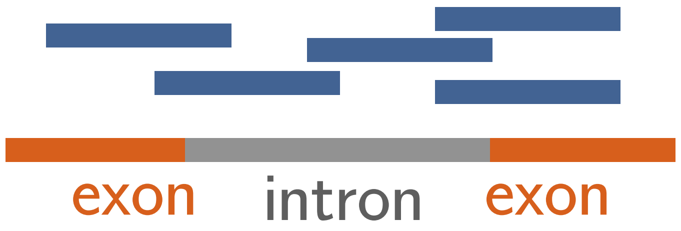
Reference genome
with exons genomic
coordinates

Alternatively:
Reference transcriptome

Annotations
with exons genomic
coordinates

Short sequence mapping tools

eg. Whole Genome Sequencing, ChIP-Seq

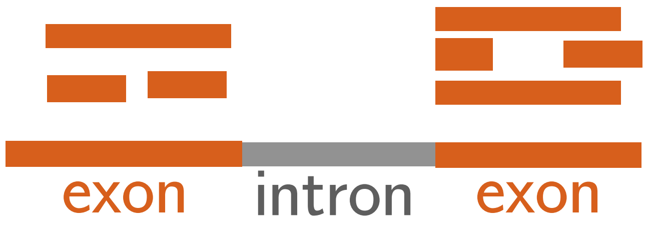


Not splice aware

Bowtie2

BWA

eg. RNA-Seq



Splice aware

STAR

TopHat2

Hisat2

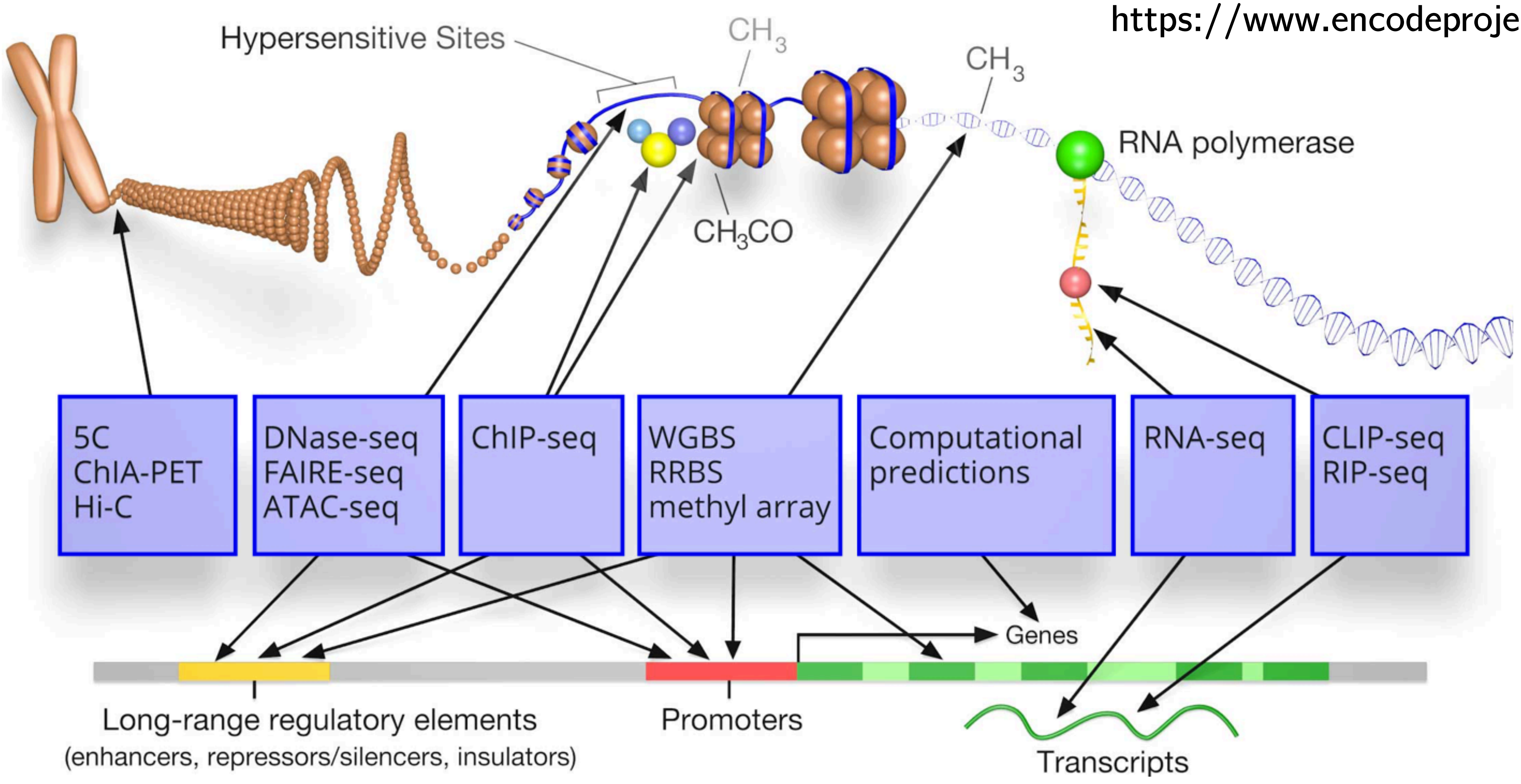
Reference genome
with exons genomic
coordinates

Alternatively:
Reference transcriptome

Annotations
with exons genomic
coordinates

ENCODE: encyclopedia of DNA elements

https://www.encodeproject.org



The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome employing variety of assays and techniques.

Annotations: GTF/GFF file

Resources:

GENCODE

RefSeq

Ensembl

GENCODE annotation is made by merging the manual gene annotation produced by the Ensembl-Havana team and the Ensembl-genebuild automated gene annotation.

exon
intron
exon

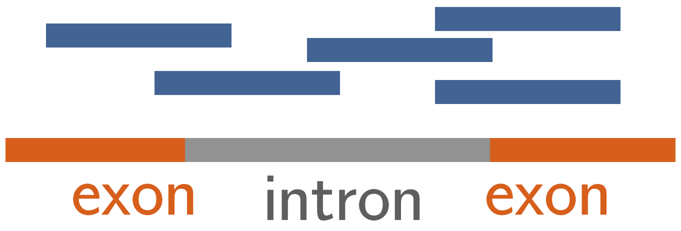
Gencode vs. Ensembl

- The gene annotation is the same in both files. The only exception is that the genes which are common to the human chromosome X and Y PAR regions can be found twice in the GENCODE GTF, while they are shown only for chromosome X in the Ensembl file.
- GENCODE GTF contains also APPRIS tags and the annotation are on the reference chromosomes only

Always make sure that annotations match the genome FASTA file (the same version & source)

Short sequence mapping tools

eg. Whole Genome Sequencing, ChIP-Seq

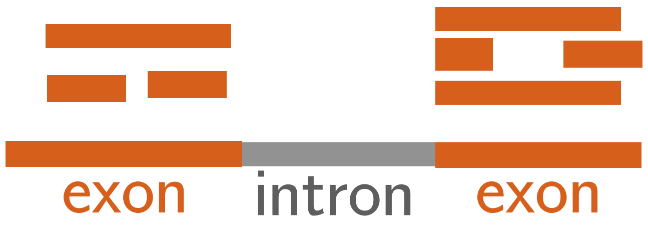


Not splice aware

Bowtie2

BWA

eg. RNA-Seq



Splice aware

STAR

TopHat2

Hisat2

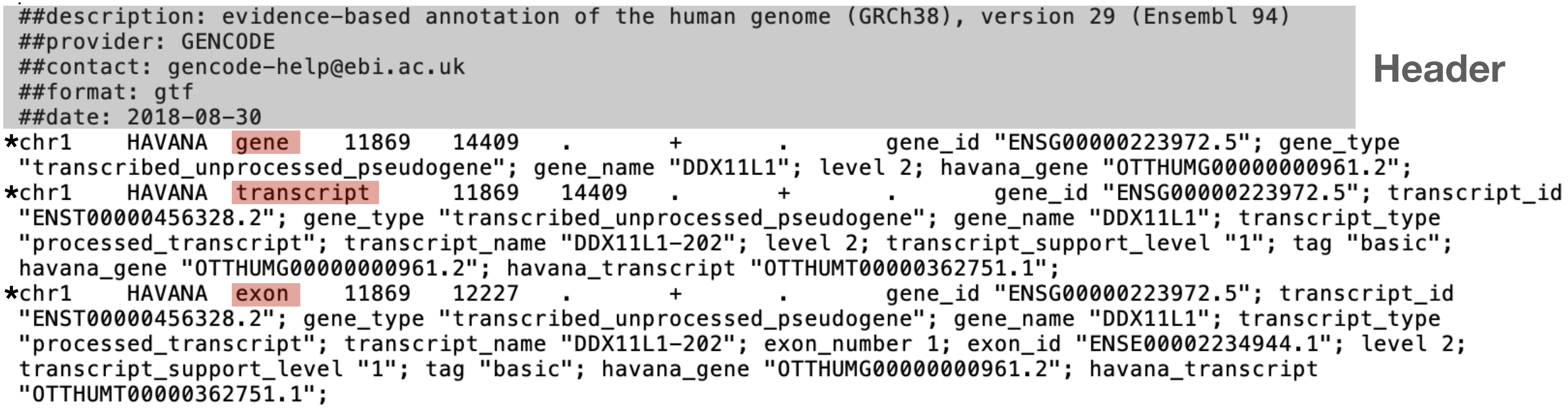
Pseudo-aligners

Reference genome
with exons genomic
coordinates

Alternatively:
Reference transcriptome

Annotations
with exons genomic
coordinates

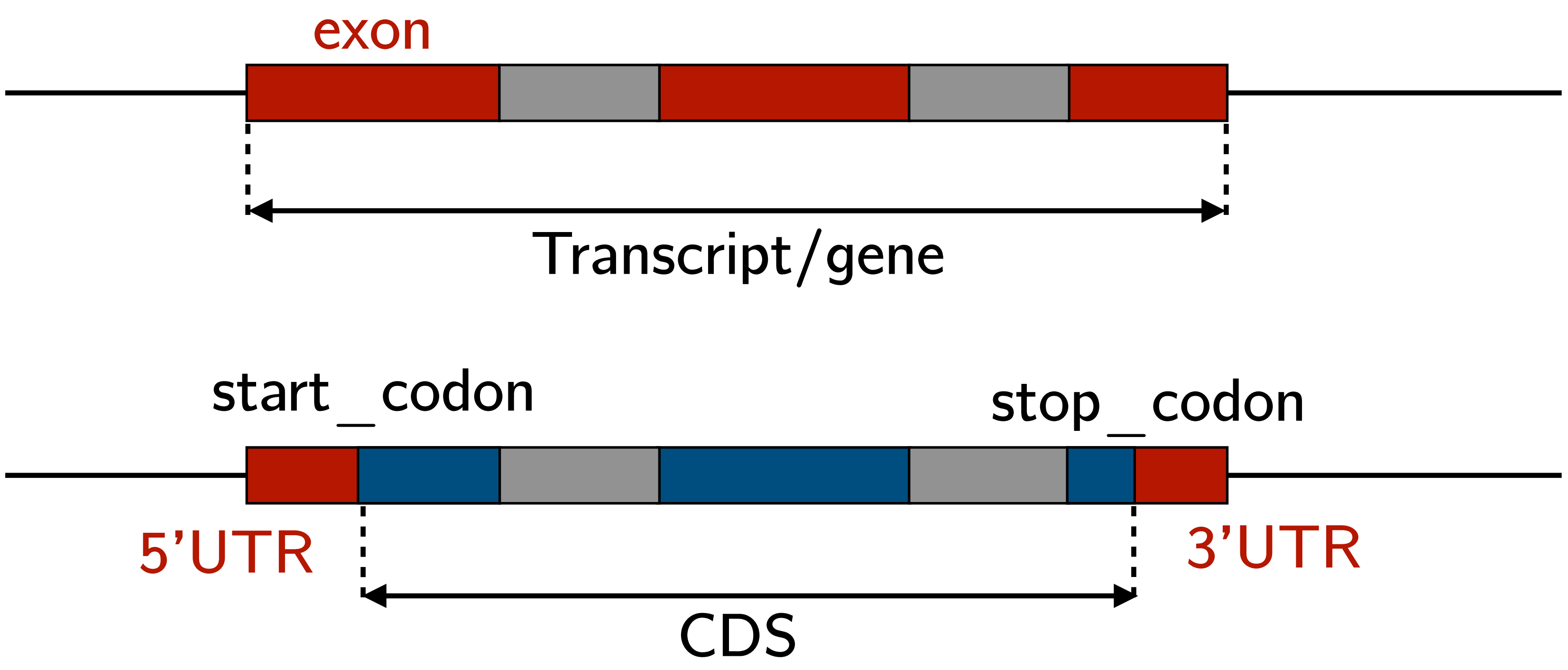
Annotations: GTF/GFF file



```
##description: evidence-based annotation of the human genome (GRCh38), version 29 (Ensembl 94)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2018-08-30
*chr1    HAVANA  gene     11869   14409   .       +       .       gene_id "ENSG00000223972.5"; gene_type
 "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana_gene "OTTHUMG00000000961.2";
*chr1    HAVANA  transcript       11869   14409   .       +       .       gene_id "ENSG00000223972.5"; transcript_id
 "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; transcript_type
 "processed_transcript"; transcript_name "DDX11L1-202"; level 2; transcript_support_level "1"; tag "basic";
 havana_gene "OTTHUMG00000000961.2"; havana_transcript "OTTHUMT00000362751.1";
*chr1    HAVANA  exon     11869   12227   .       +       .       gene_id "ENSG00000223972.5"; transcript_id
 "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; transcript_type
 "processed_transcript"; transcript_name "DDX11L1-202"; exon_number 1; exon_id "ENSE00002234944.1"; level 2;
 transcript_support_level "1"; tag "basic"; havana_gene "OTTHUMG00000000961.2"; havana_transcript
 "OTTHUMT00000362751.1";
```

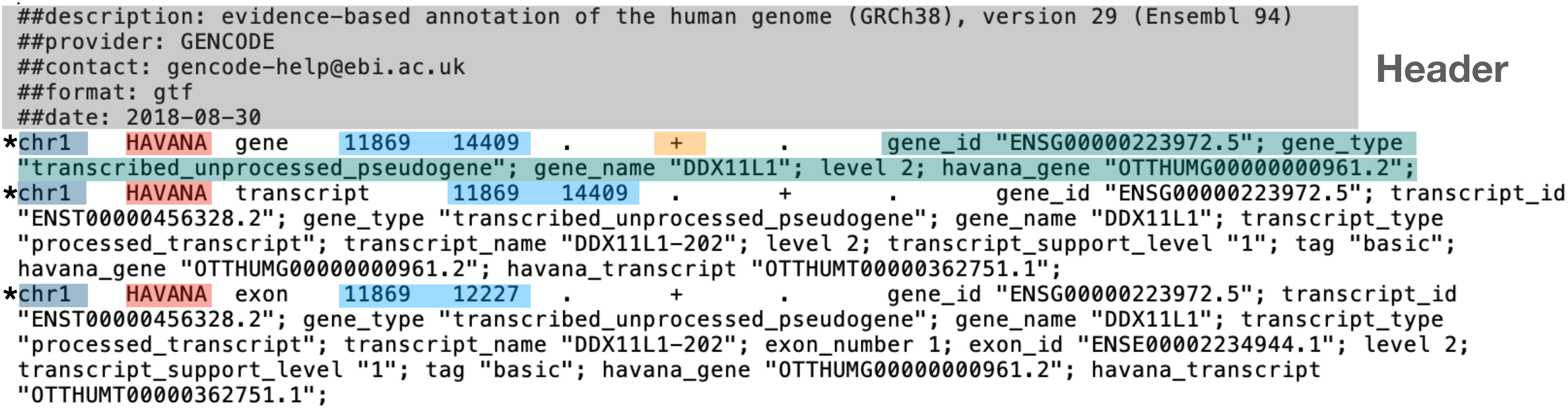
Header

feature type {gene,transcript,exon,CDS,UTR,start_codon,stop_codon}



*** New line**

Annotations: GTF/GFF file



```
##description: evidence-based annotation of the human genome (GRCh38), version 29 (Ensembl 94)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2018-08-30
*chr1   HAVANA  gene    11869   14409   .       +       .       gene_id "ENSG00000223972.5"; gene_type
 "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; level 2; havana_gene "OTTHUMG00000000961.2";
*chr1   HAVANA  transcript      11869   14409   .       +       .       gene_id "ENSG00000223972.5"; transcript_id
 "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; transcript_type
 "processed_transcript"; transcript_name "DDX11L1-202"; level 2; transcript_support_level "1"; tag "basic";
 havana_gene "OTTHUMG00000000961.2"; havana_transcript "OTTHUMT00000362751.1";
*chr1   HAVANA  exon    11869   12227   .       +       .       gene_id "ENSG00000223972.5"; transcript_id
 "ENST00000456328.2"; gene_type "transcribed_unprocessed_pseudogene"; gene_name "DDX11L1"; transcript_type
 "processed_transcript"; transcript_name "DDX11L1-202"; exon_number 1; exon_id "ENSE00002234944.1"; level 2;
 transcript_support_level "1"; tag "basic"; havana_gene "OTTHUMG00000000961.2"; havana_transcript
 "OTTHUMT00000362751.1";
```

Header

Genomic coordinates

Annotation source

Strand

Additional information

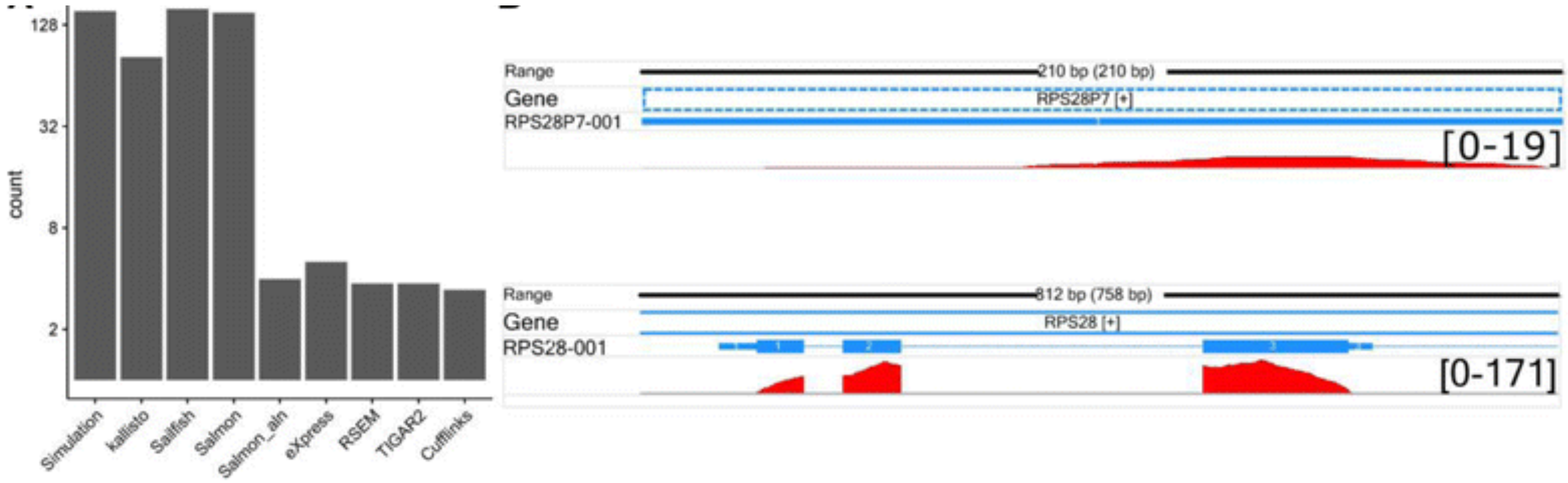
Gene id	Gene name	Exon number
Transcript id	Transcript type	Exon id
Gene type	Transcript status	Level
Gene status	Transcript status	

*** New line**

Pseudo-aligners

Salmon
Sailfish
Kallisto

- Quantification estimates rather than base-to-base alignment
- Can model sequencing bias, eg. GC-bias, fragment length
- Can handle multi mapping
- Faster
- Improved accuracy at the transcript level

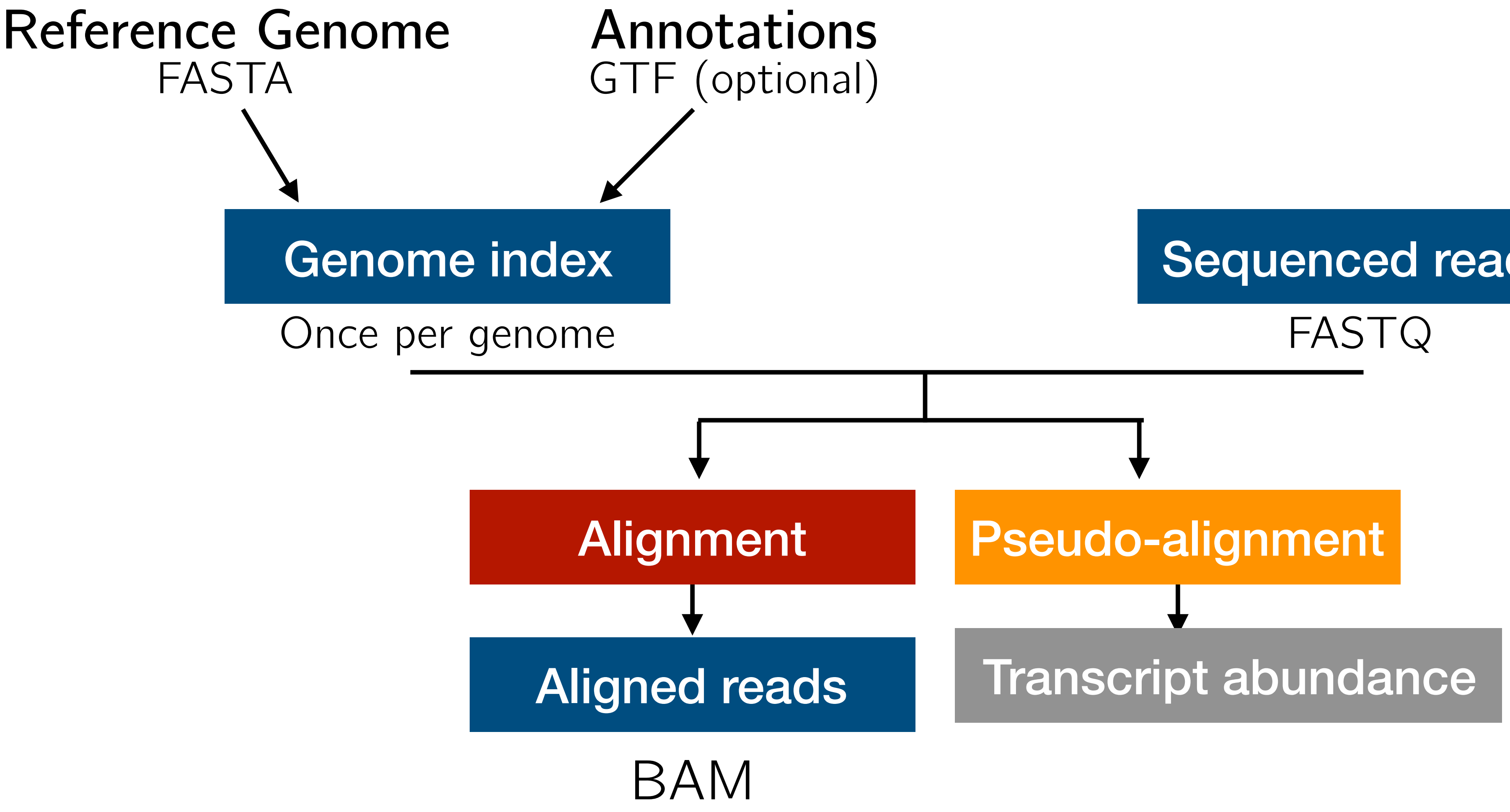


Zhang, C., Zhang, B., Lin, L. L., & Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics, 18(1), 1–11.

Before you align checklist & standard workflow

- Do I need splice-aware aligner?
- Am I using right genome version? (hg38 - human, mm10 -mouse?)
- Do annotations match the reference genome?
- Read manual, select parameters, check default settings

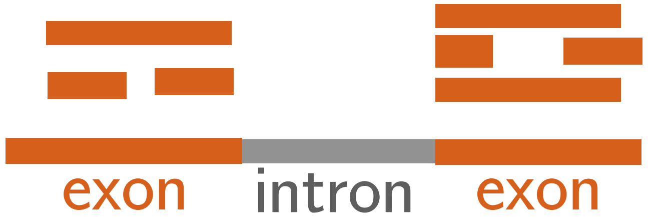
Standard alignment workflow

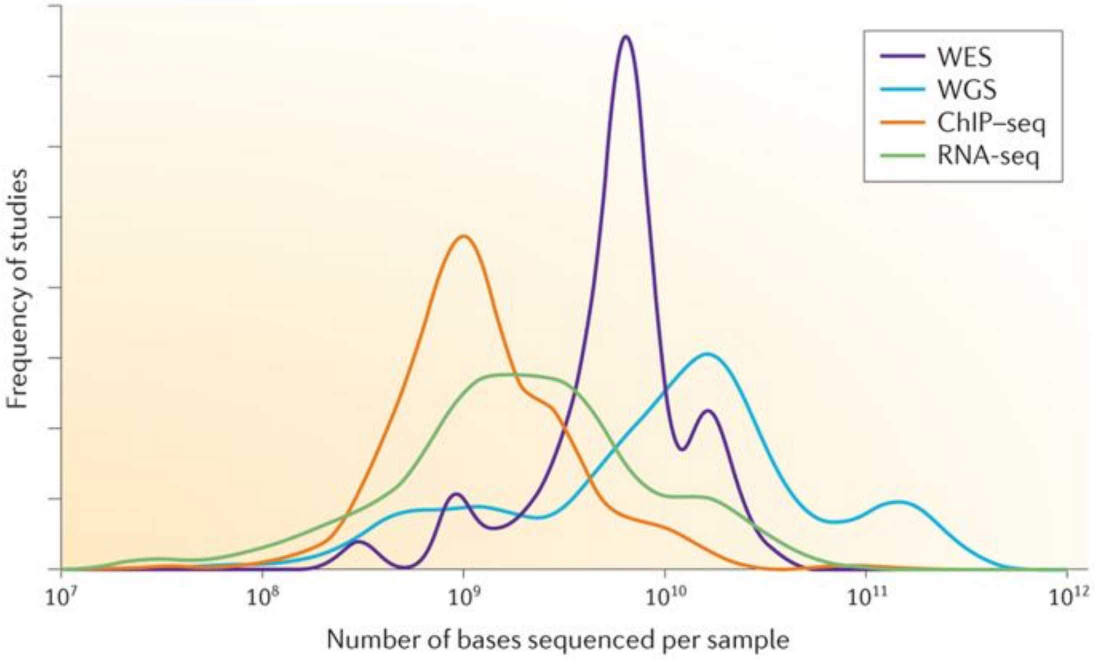


Coverage and Depth

Coverage: average number of reads of a given length that align to given region.

Depth: redundancy of coverage or the total number of bases sequenced and aligned at a given reference position.





Nature Reviews | Genetics

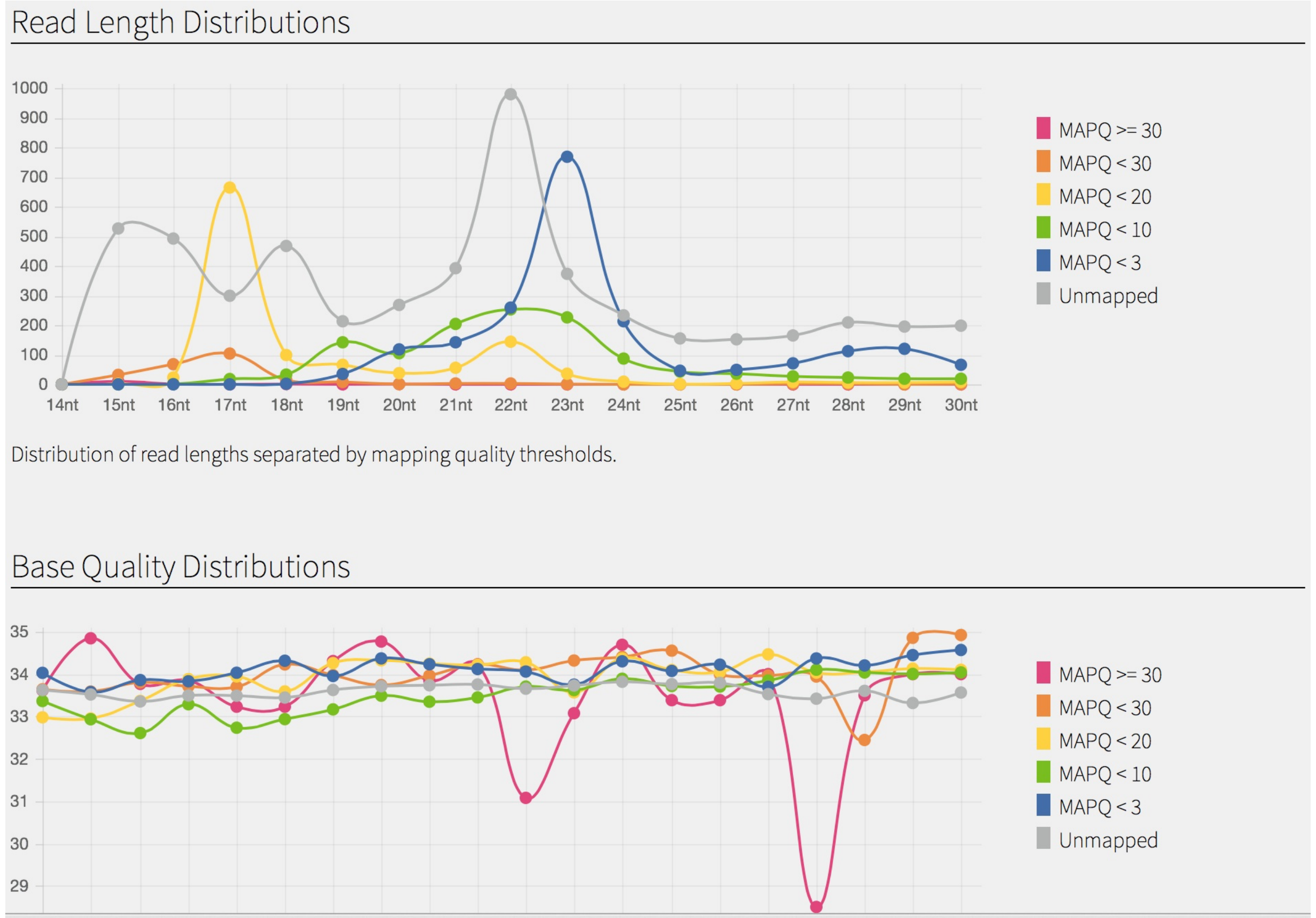
The average depth of sequencing coverage can be defined theoretically as *LN/G*, where *L* is the read length, *N* is the number of reads and *G* is the haploid genome length.

Example: If we sequence a genome with total length of 100 nucleotides and we have 500 reads, 25 nucleotides length each - the average depth of sequencing is 125

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics, 15(2),

Mapping quality check

SAMstat is a C program that plots nucleotide overrepresentation and other statistics in mapped and unmapped reads and helps understand the relationship between potential protocol biases and poor mapping.


Read Length Distributions

Distribution of read lengths separated by mapping quality thresholds.

Base Quality Distributions

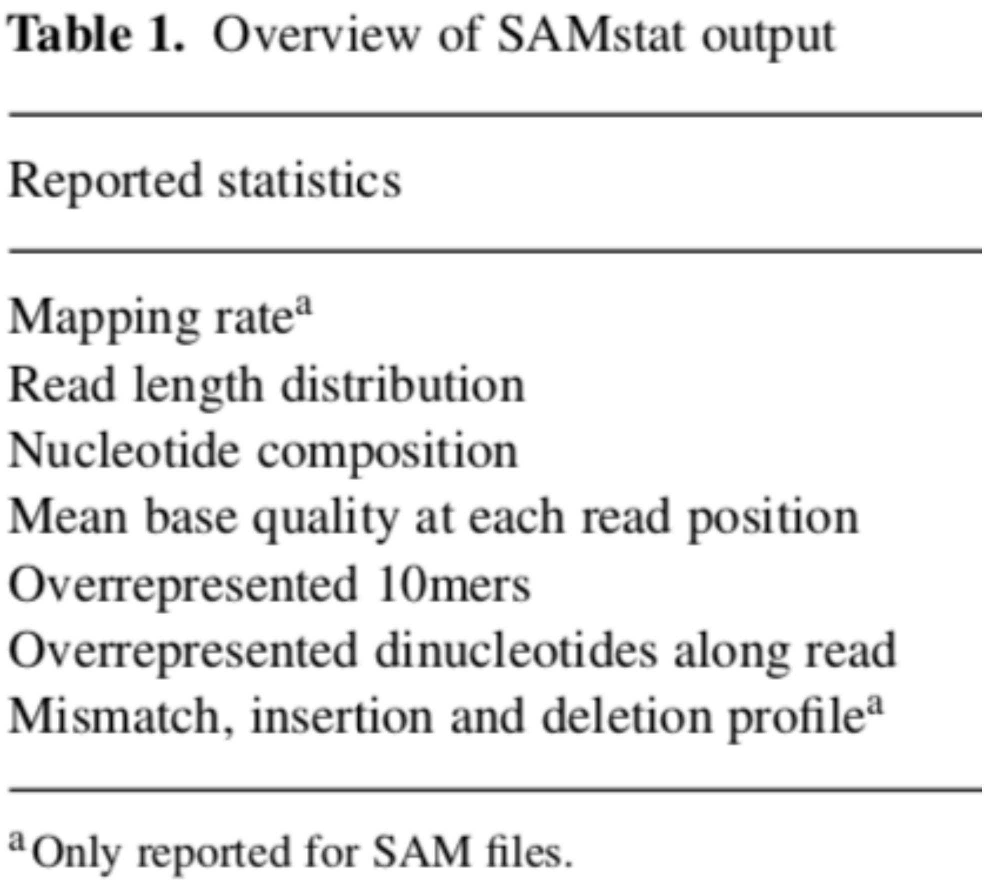


Table 1. Overview of SAMstat output

Reported statistics
Mapping rate[a]
Read length distribution
Nucleotide composition
Mean base quality at each read position
Overrepresented 10mers
Overrepresented dinucleotides along read
Mismatch, insertion and deletion profile[a]

[a] Only reported for SAM files.

Log files returned by aligner, eg Log.final.out file from STAR

FastQC

Let’s practice!