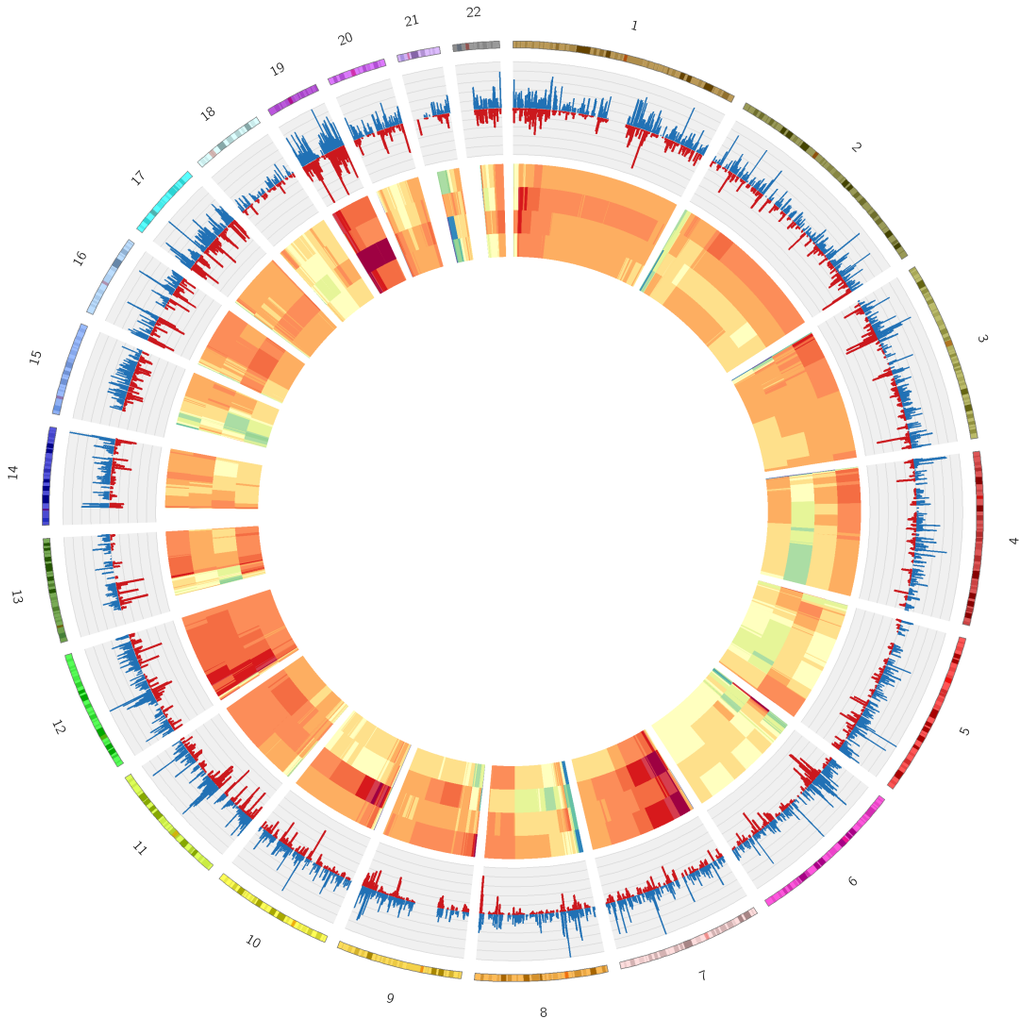


Short Reads Alignment to a Reference Genome

Joanna Krupka

CRUK Summer School in Bioinformatics

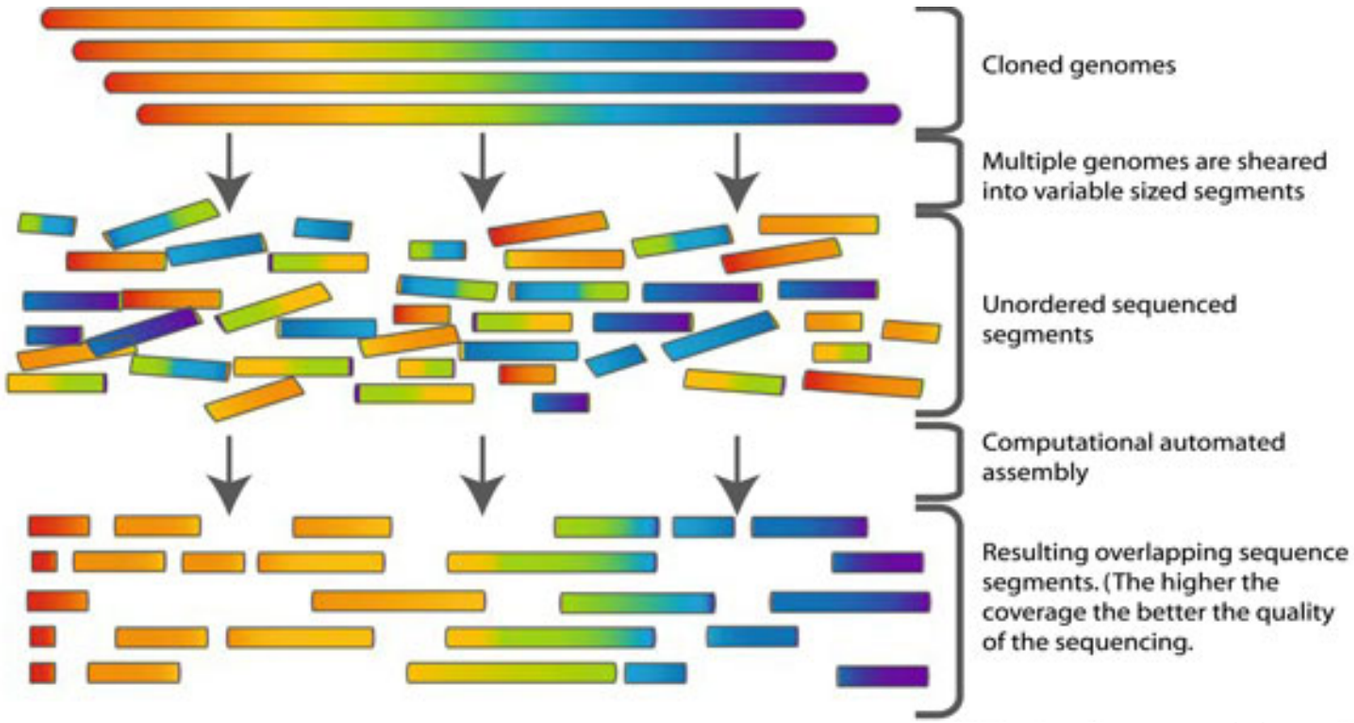
Cambridge, July 2019

CANCER RESEARCH UK

MRC | Cancer Unit

UNIVERSITY OF CAMBRIDGE

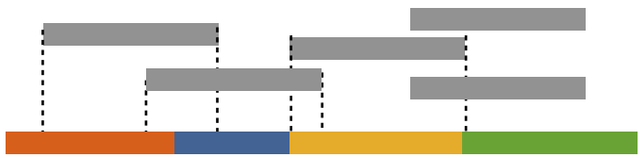
Shotgun Sequencing and sequence assembly approaches



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

Mapping to reference sequence

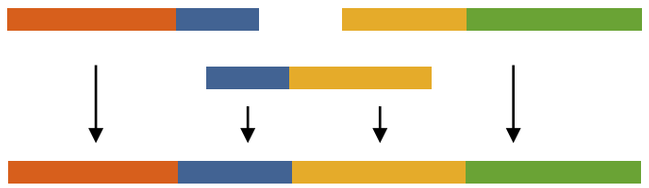
Recreate the genome with using prior knowledge as reference



Mapping is as good as reference used

De Novo assembly

Recreate the genome with no prior knowledge



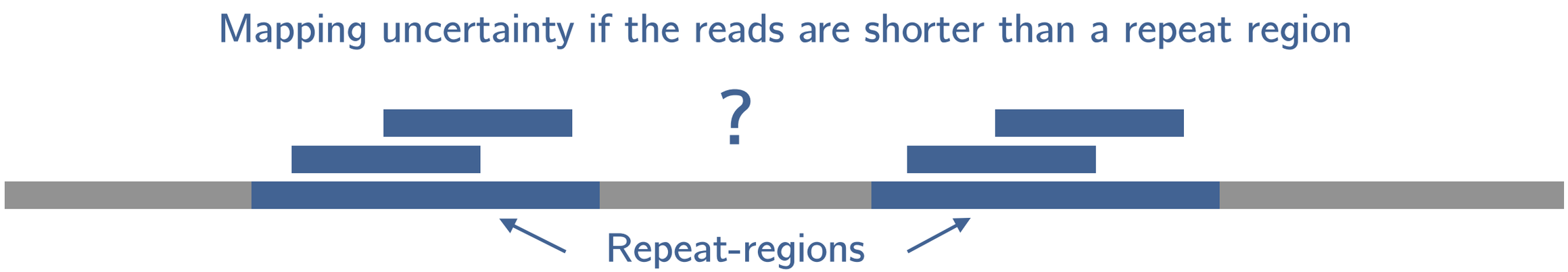
Problem with repeated regions, high coverage and long reads required

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%

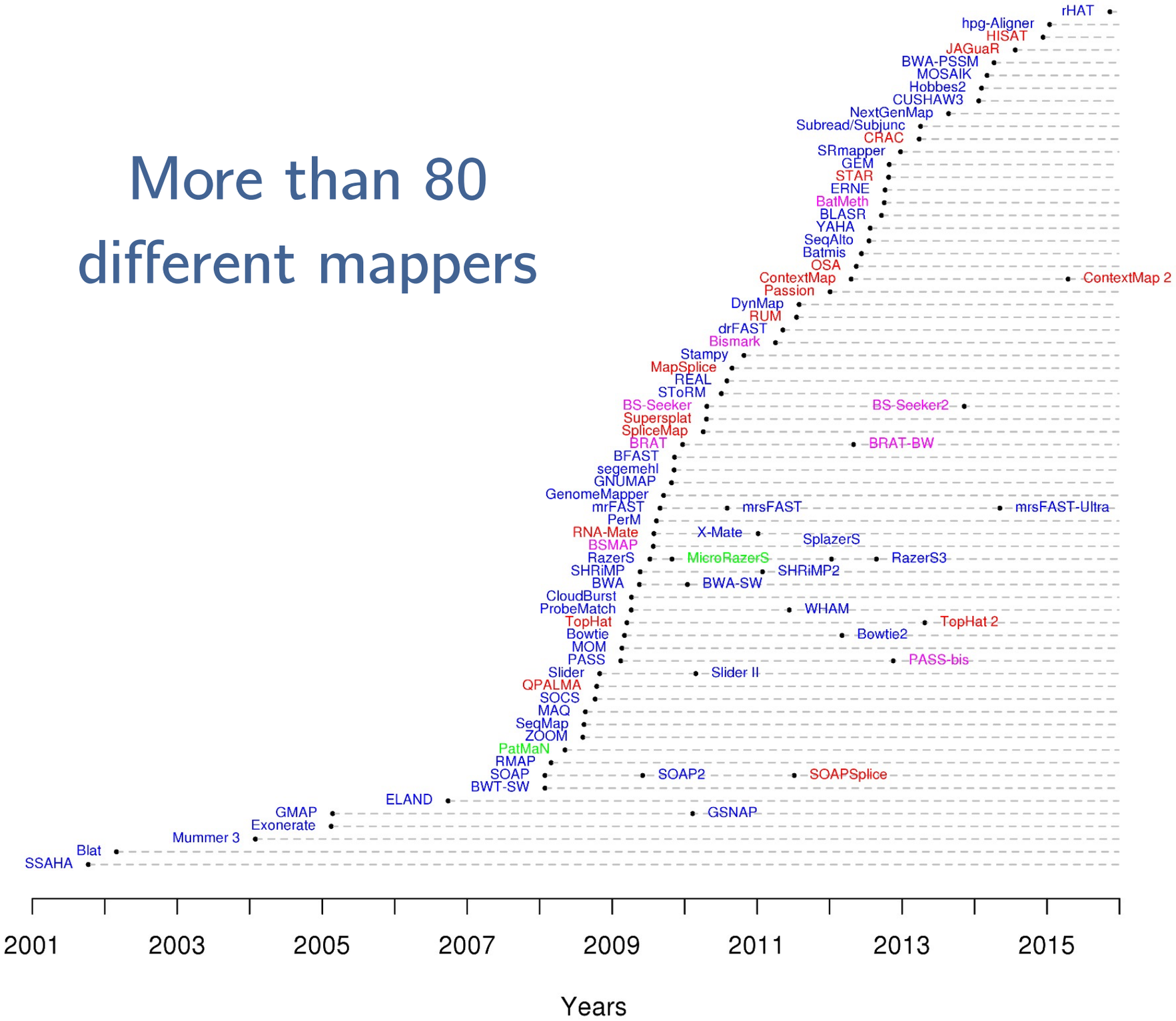
Rozowsky J. Et al. Nat Biotechnol 2009

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



Short sequence mapping tools

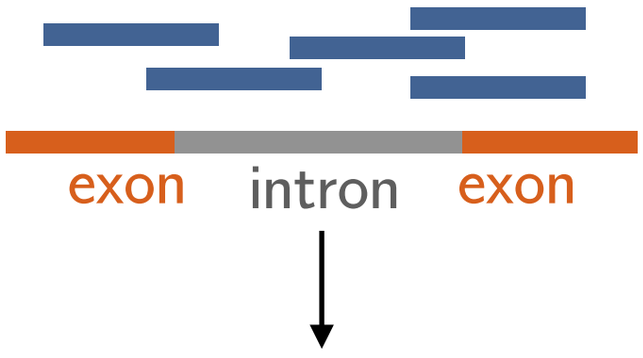
More than 80
different mappers



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

Short sequence mapping tools

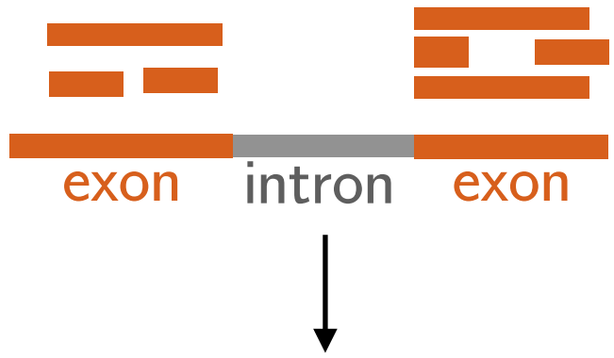
eg. Whole Genome Sequencing



Not splice aware

Bowtie2
BWA

eg. RNA-Seq



Splice aware

STAR
TopHat2
Hisat2

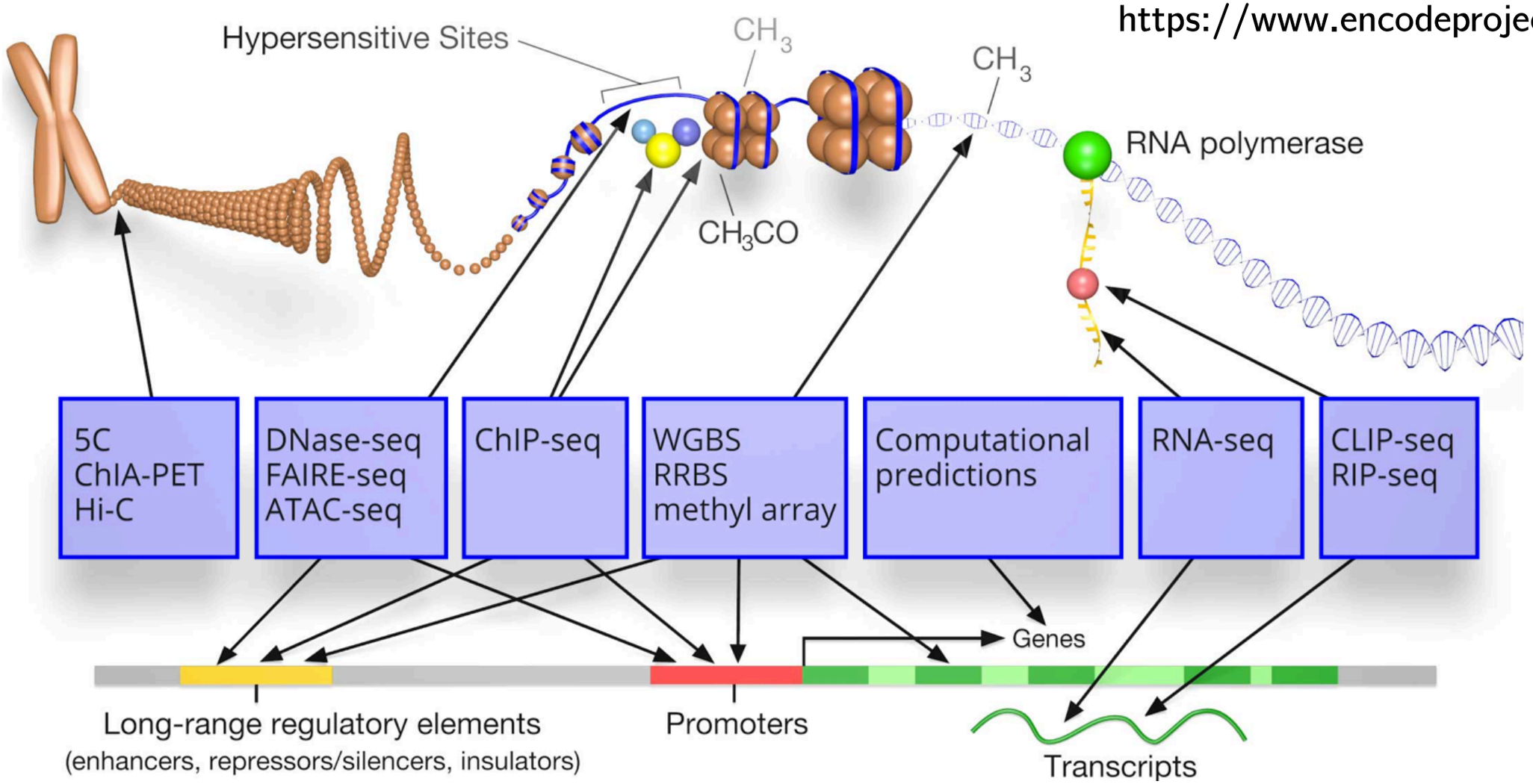
Reference genome
with exons genomic coordinates

Annotations
with exons genomic coordinates

Alternatively:
Reference transcriptome

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

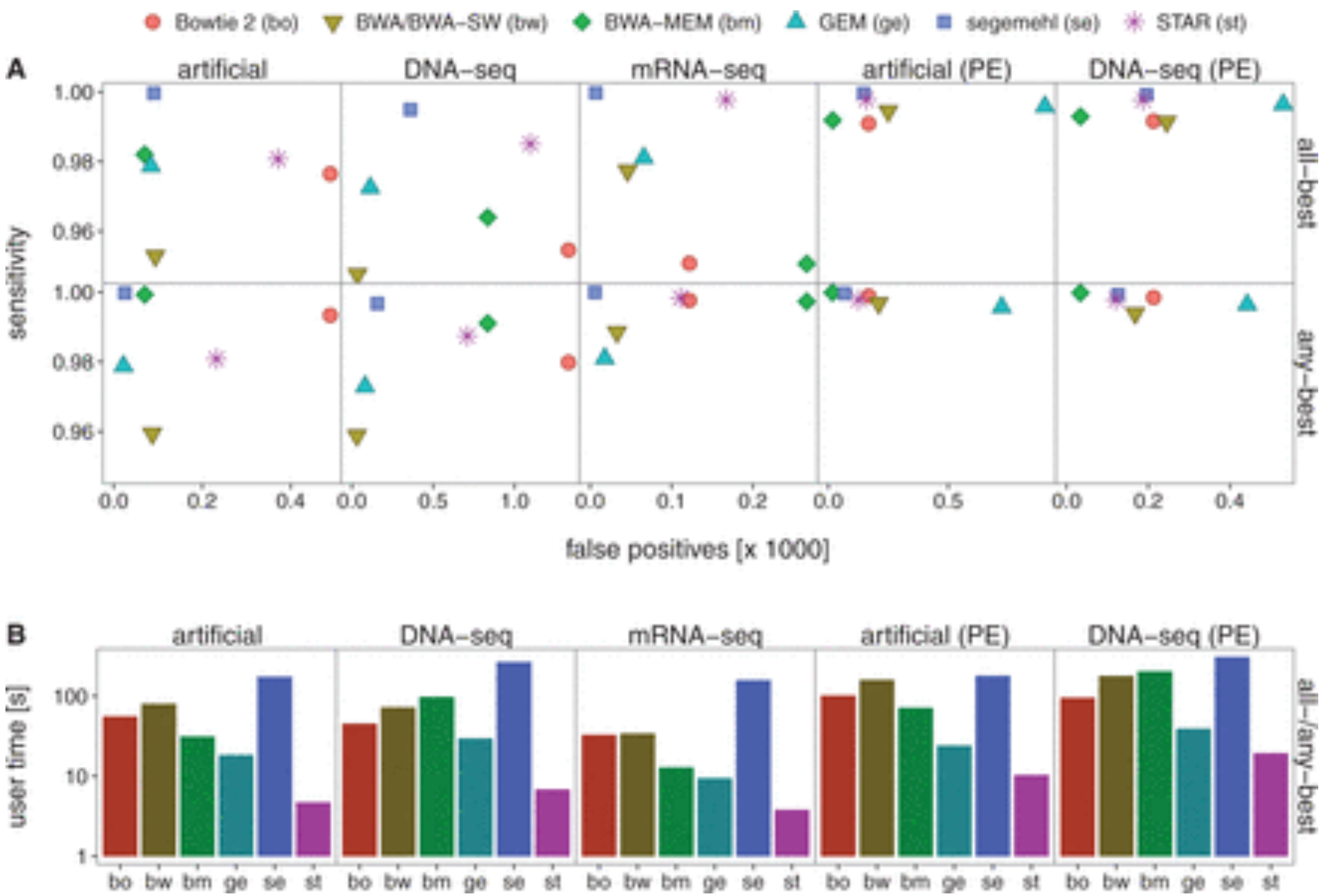
Ensembl

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

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

Short reads aligners comparison



Otto, C., Stadler, P. F., & Hoffmann, S. (2014). Lacking alignments? The next-generation sequencing mapper segemehl revisited. Bioinformatics, 30(13), 1837–1843.

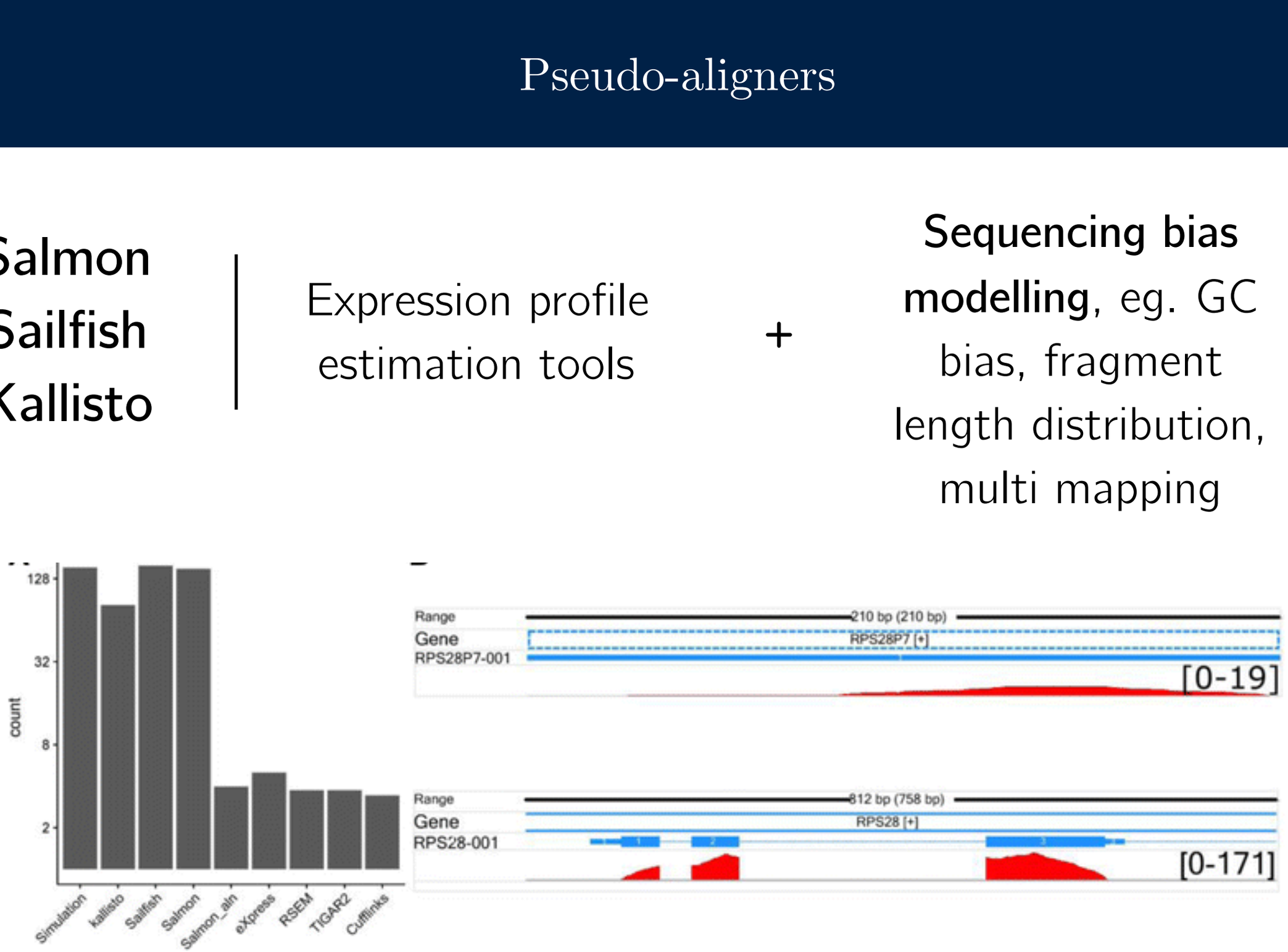
Pseudo-aligners

Salmon
Sailfish
Kallisto

Expression profile estimation tools

+

Sequencing bias modelling, eg. GC bias, fragment length distribution, multi mapping



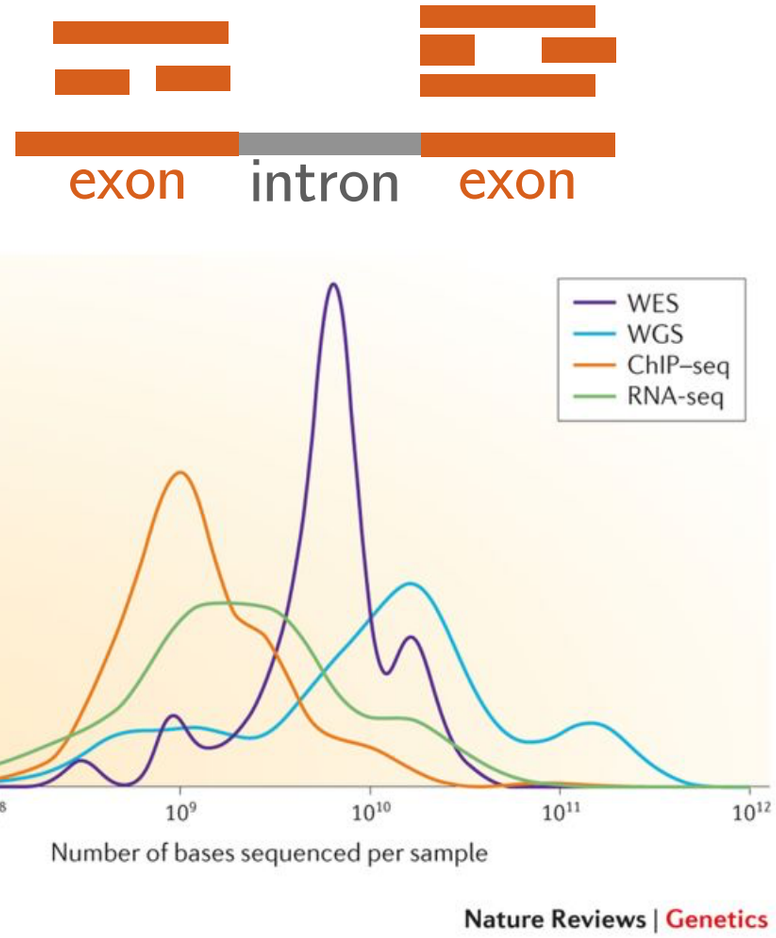
Read more: https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/08_salmon.html

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.

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.



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.

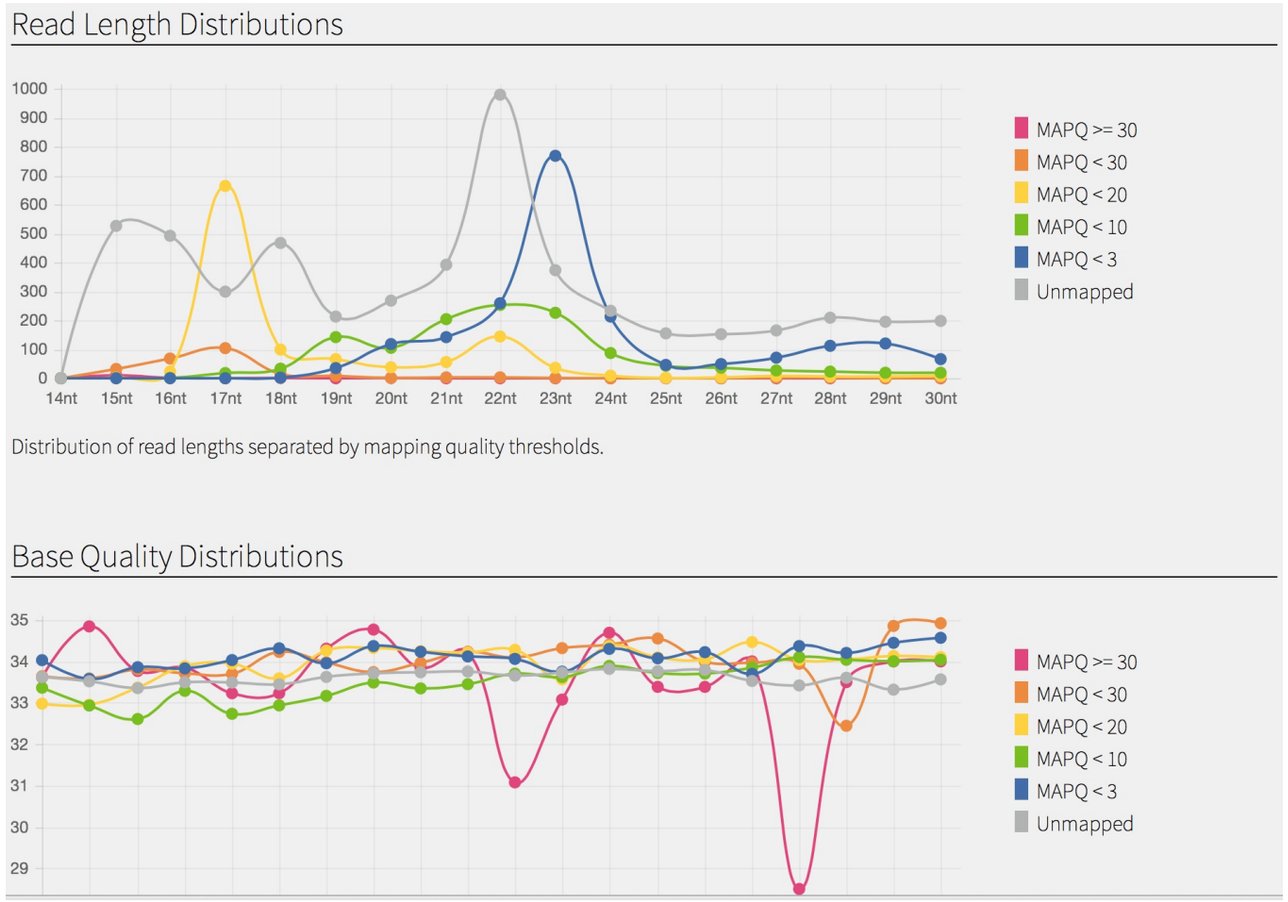


Table 1. Overview of SAMstat output

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.

Other possible QC measures: genomic regions distribution, reproducibility between replicates, observations consistent with experimental conditions etc.

Processing alignment files

Downstream analysis is highly depended on sequencing technique and biological question. Sometimes files need to be modified before using a specific bioinformatic tool.

Some useful software:

SAMtools (RSamTools): sorting, indexing - BAM/SAM files

Bedtools: *intersect*, *merge*, *count*, *complement*, and *shuffle* genomic intervals (BED files)

BBMap

Picard

Deeptools: coverage computation

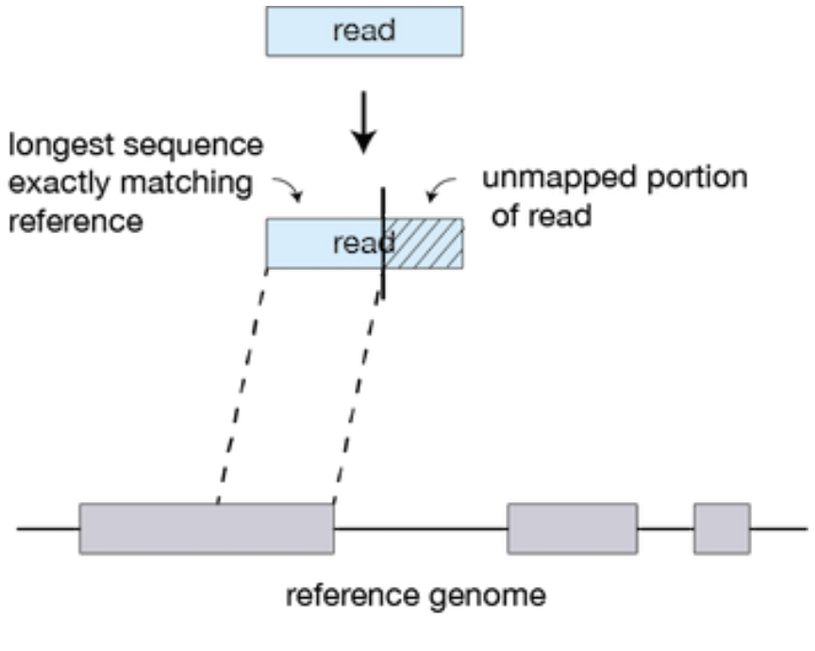
UCSC Genome Browser Utilities

...

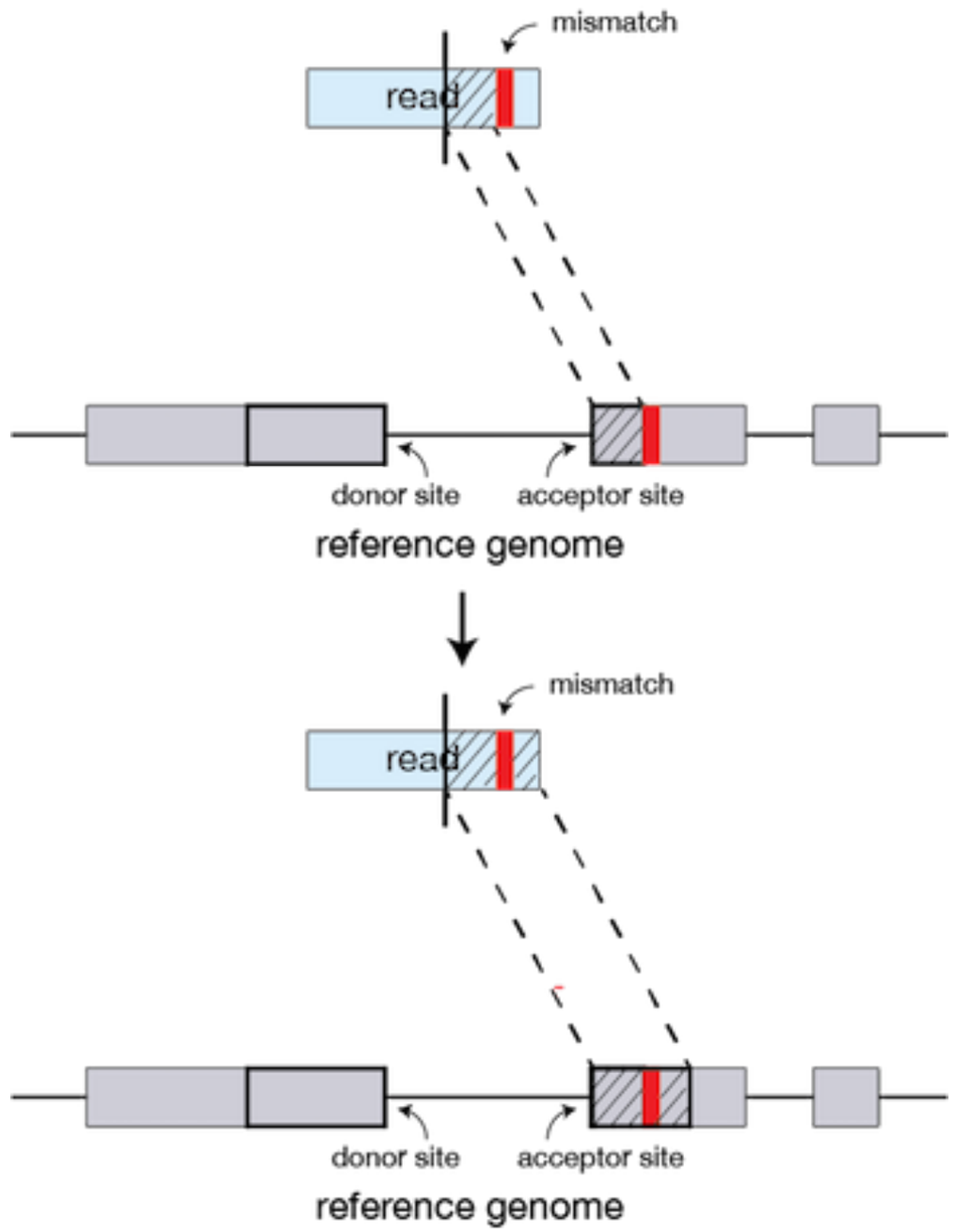
and many more

Appendix: Example mapping algorithm: STAR

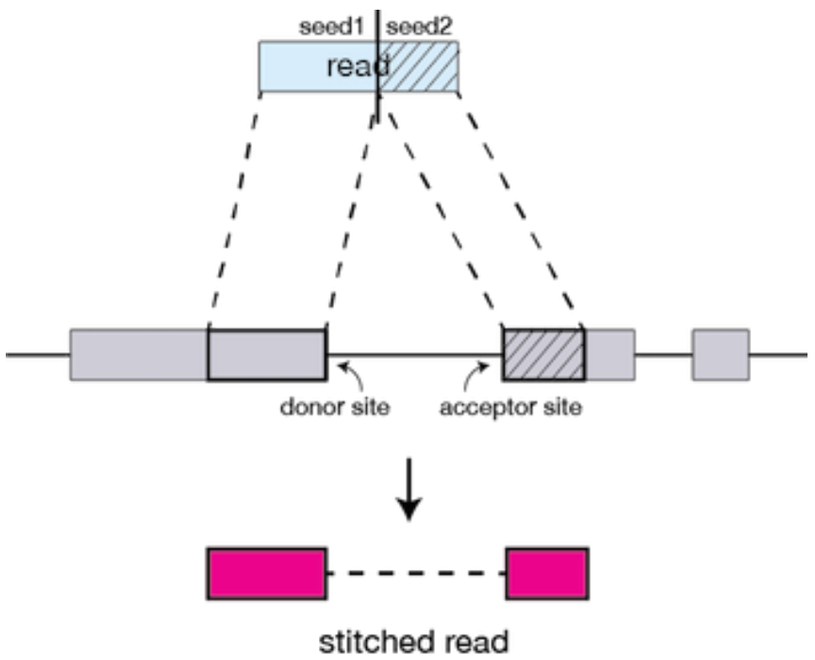
Seed searching



Extension/Mismatches recognition



Scoring



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html