

RNA-seq aligners

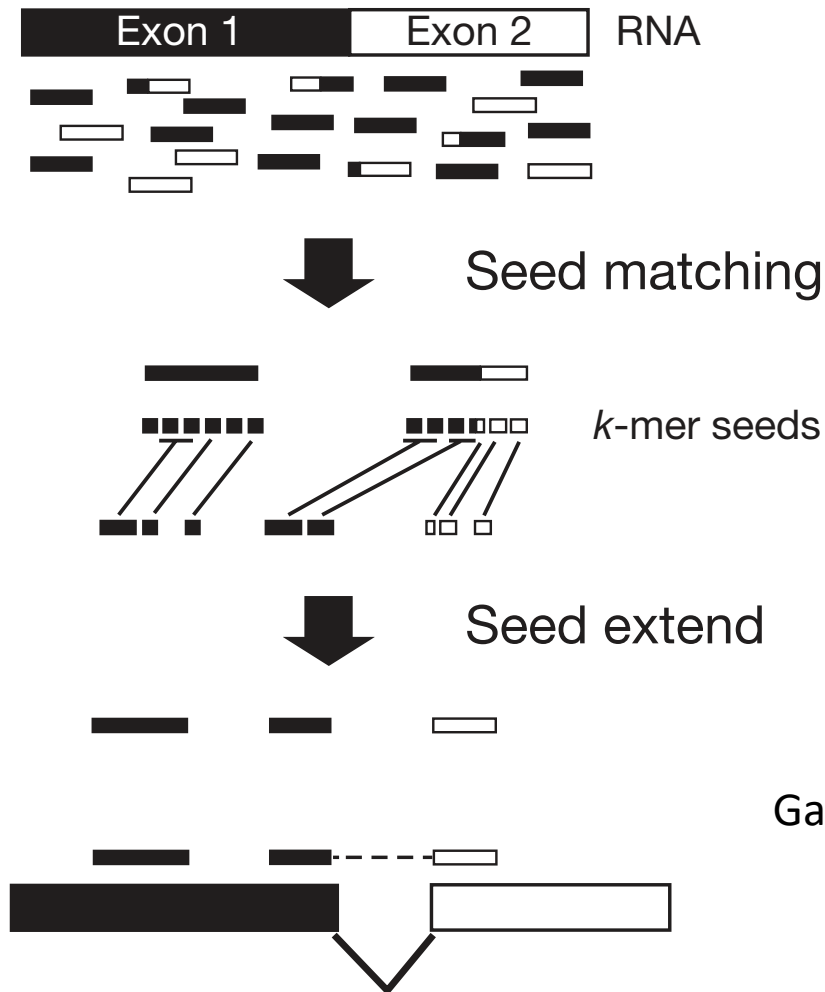
RNA-seq data analysis

Johan Reimegård | 15-November-2021

Innovations in RNA-seq alignment software

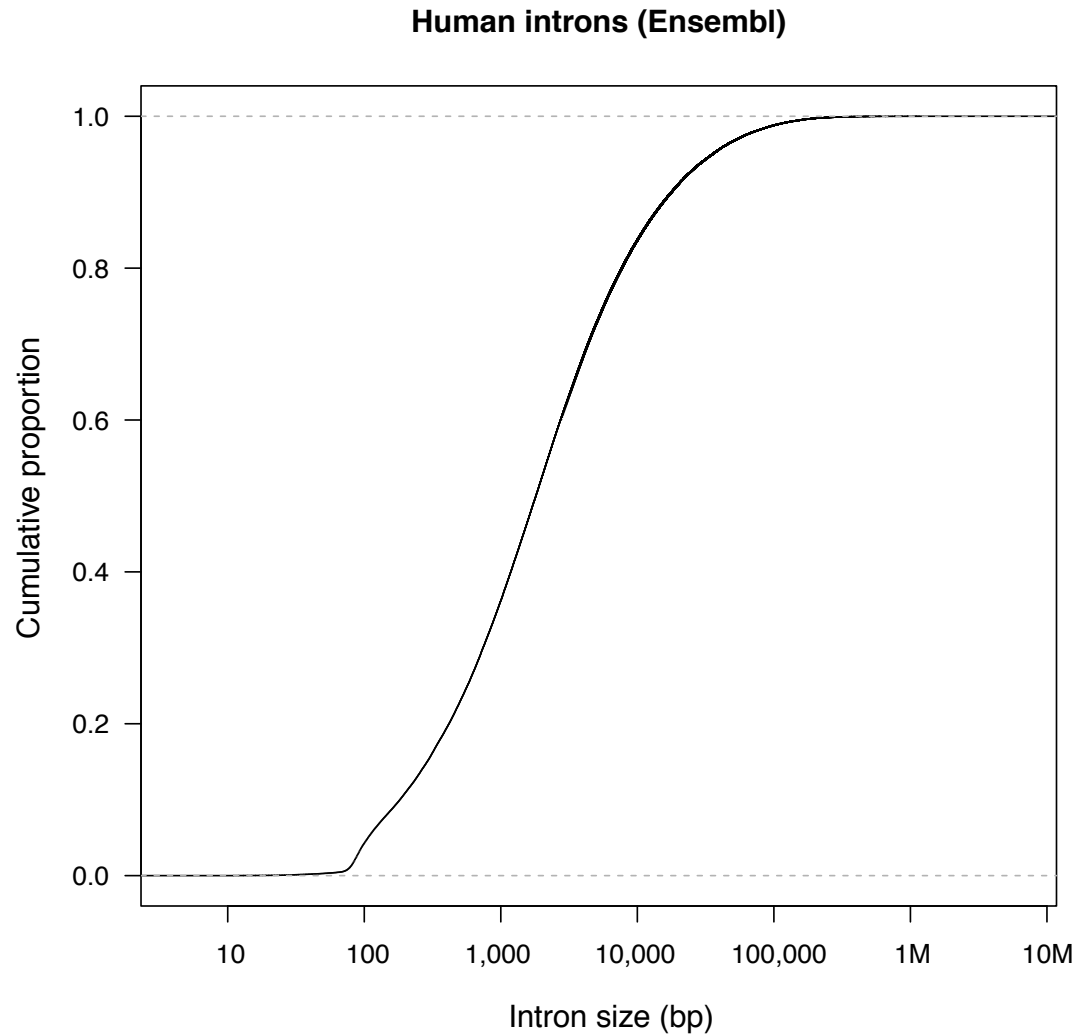
- Read pair alignment
- Consider base call quality scores
- Sophisticated indexing to decrease CPU and memory usage
- Map to genetic variants
- Resolve multi-mappers using regional read coverage
- Consider junction annotation
- Two-step approach (junction discovery & final alignment)

Most aligner use a seed and extend approach

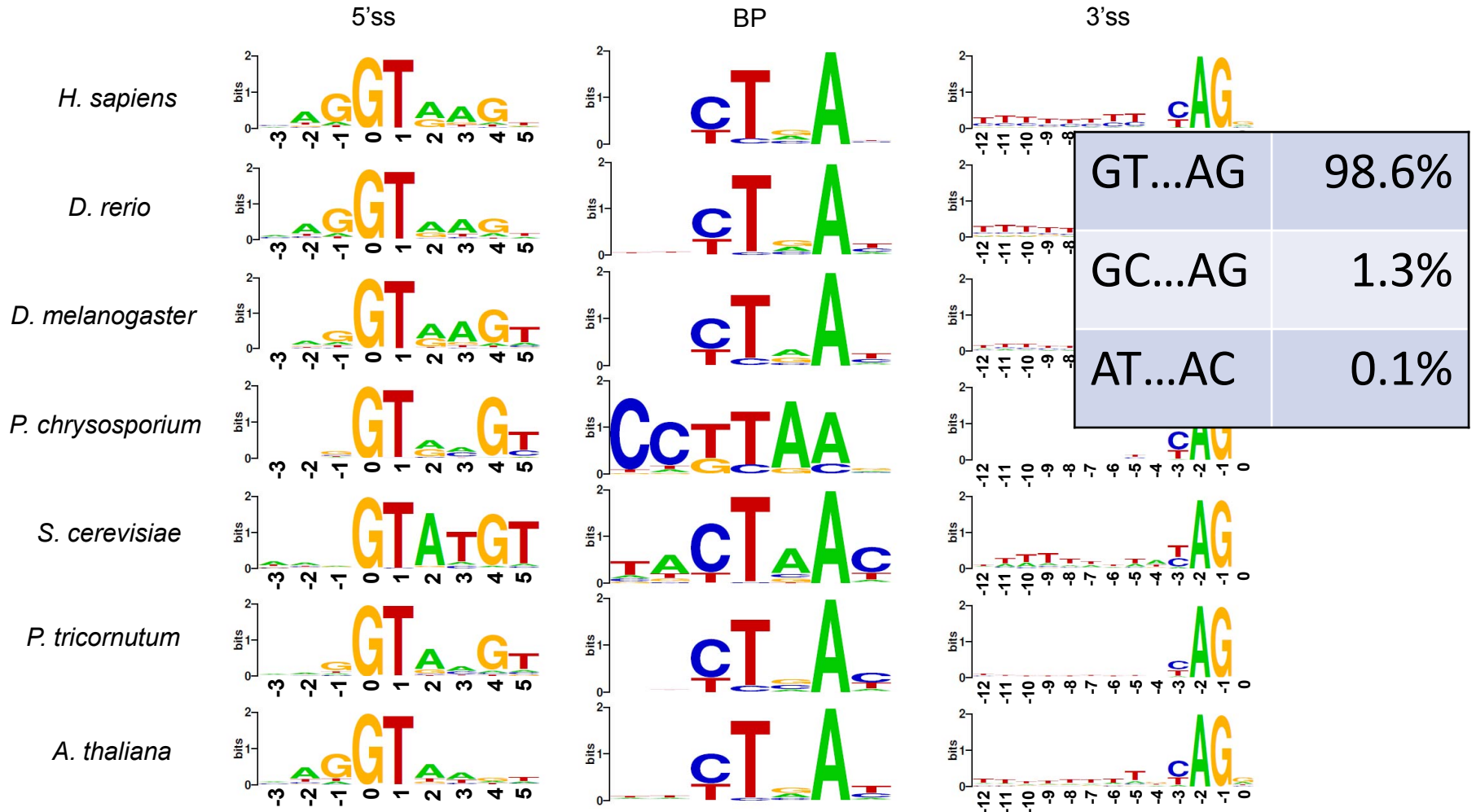


Garber et al. *Nature Methods* 2011

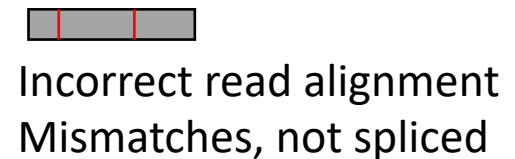
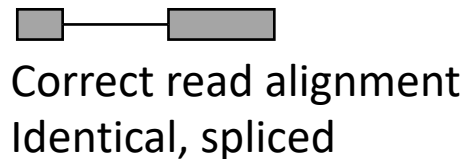
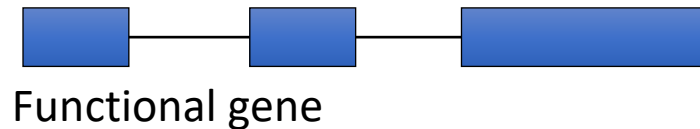
Introns can be very large!



Limited sequence signals at splice sites



Multi-mapping reads and pseudogenes



Note:

- An aligner may report both alignments or either
- Some search strategies and scoring schemes give preference to unspliced alignments

How important is mapping accuracy?



Depends what you want to do:

Identify novel genetic variants or RNA editing

Allele-specific expression

Genome annotation

Gene and transcript discovery

Differential expression

Current RNA-seq aligners

TopHat2	Kim et al. <i>Genome Biology</i> 2013
HISAT2	Kim et al. <i>Nature Methods</i> 2015
STAR	Dobin et al. <i>Bioinformatics</i> 2013
GSNAP	Wu and Nacu <i>Bioinformatics</i> 2010
OLego	Wu et al. <i>Nucleic Acids Research</i> 2013
HPG aligner	Medina et al. <i>DNA Research</i> 2016
MapSplice2	http://www.netlab.uky.edu/p/bioinfo/MapSplice2

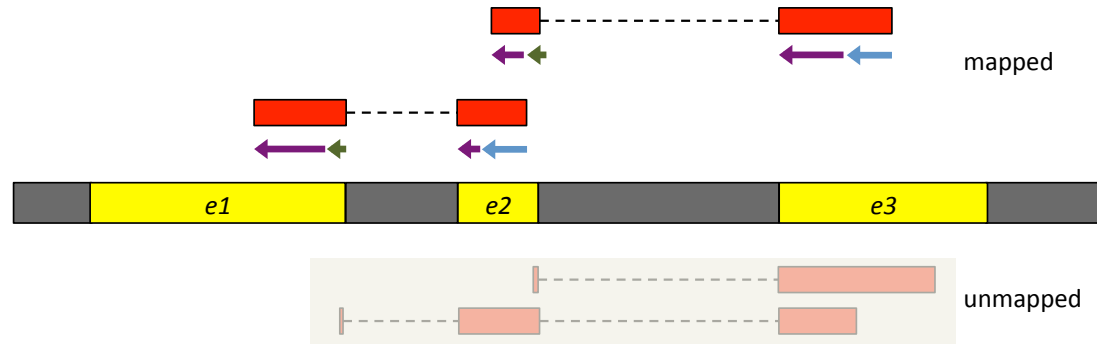
Compute requirements

Program	Run time (min)	Memory usage (GB)
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
OLego	989.5	3.7
TopHat2	1,170	4.3

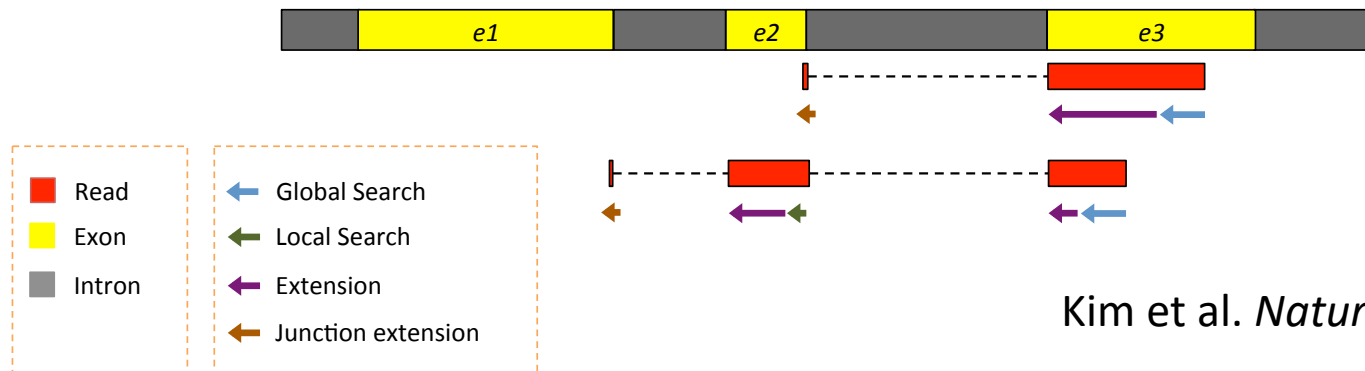
Run times and memory usage for HISAT and other spliced aligners to align 109 million 101-bp RNA-seq reads from a lung fibroblast data set. We used three CPU cores to run the programs on a Mac Pro with a 3.7 GHz Quad-Core Intel Xeon E5 processor and 64 GB of RAM.

Two-step RNA-seq read mapping

1st run of HISAT to discover splice sites

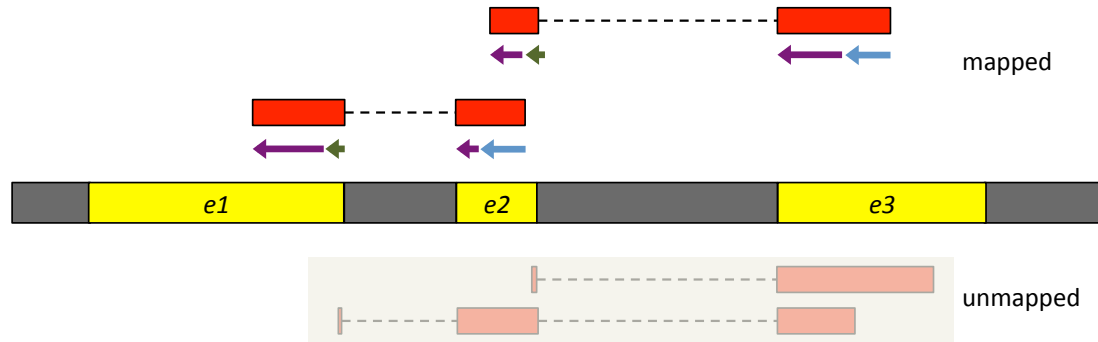


2nd run of HISAT to align reads by making use of the list of splice sites collected above

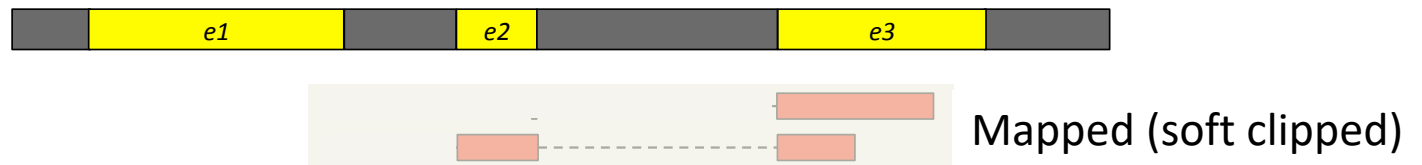


Kim et al. *Nature Methods* 2015

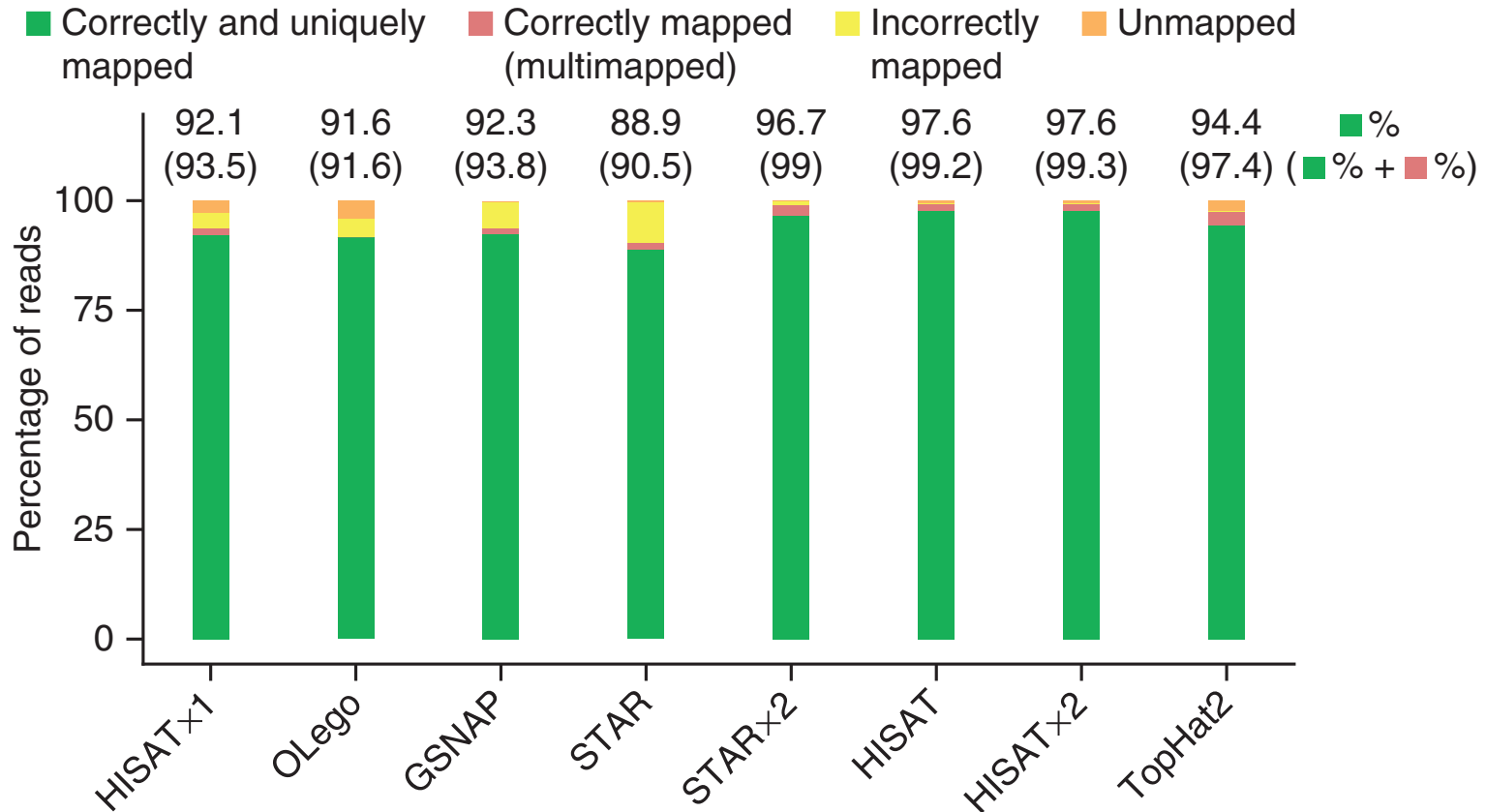
Alternative: soft clipping



Soft clipping

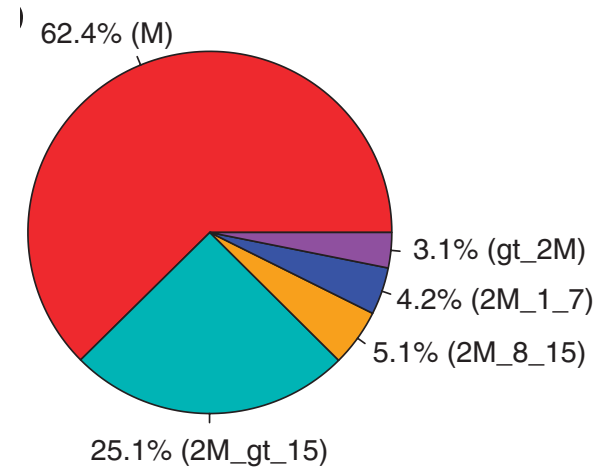
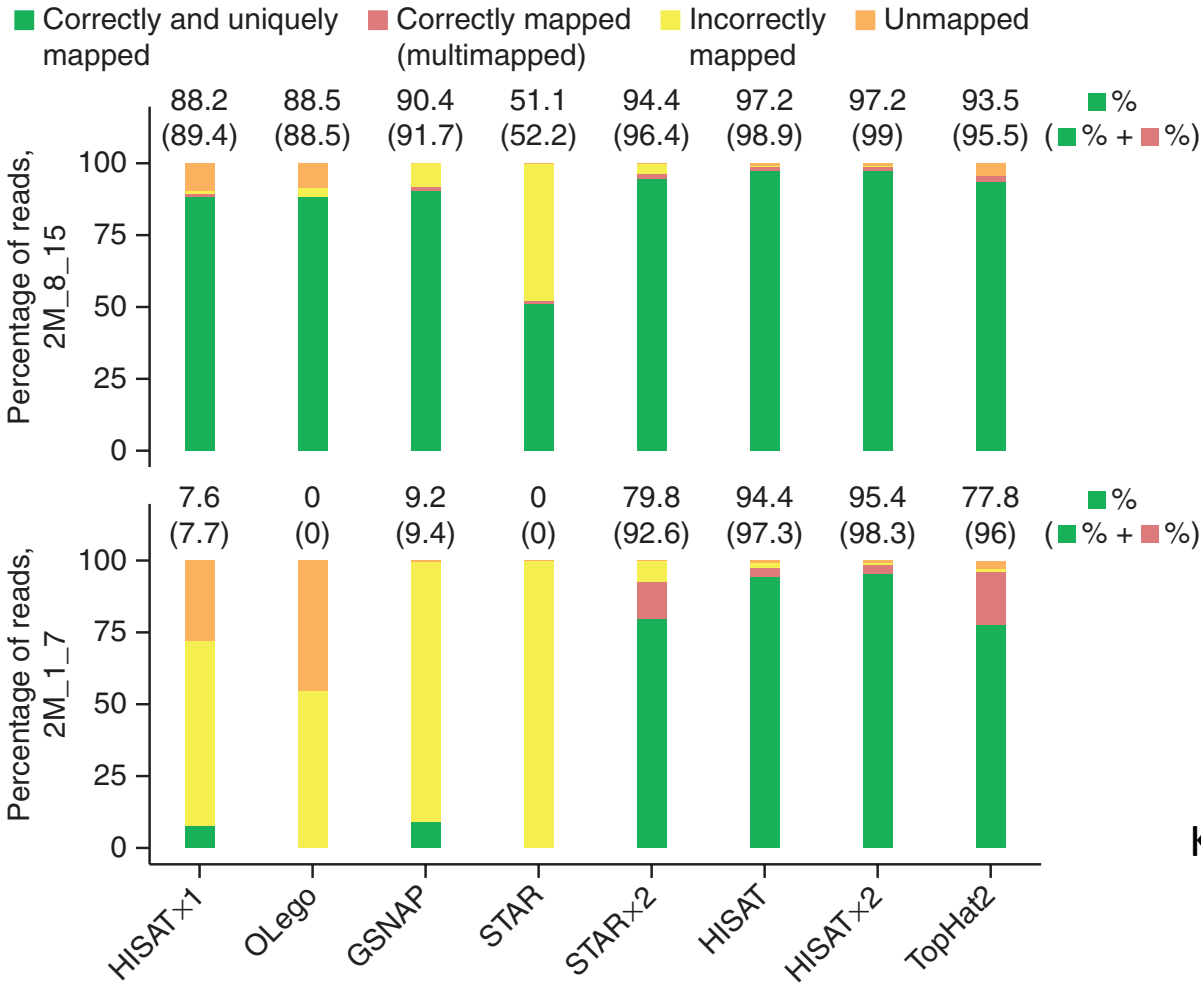


Mapping accuracy



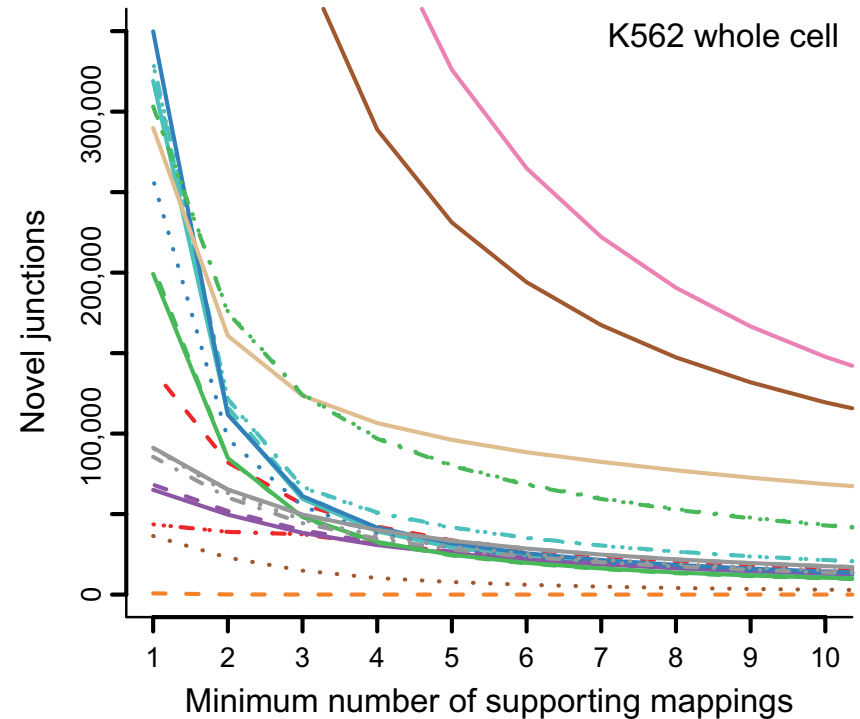
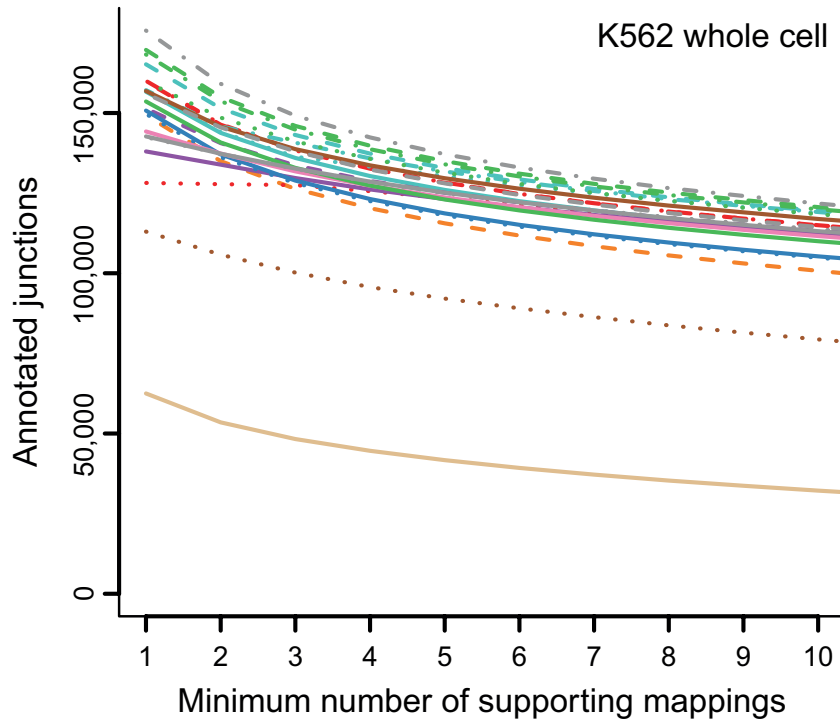
Accuracy for 20 million simulated human 100 bp reads with 0.5% mismatch rate

Mapping accuracy for reads with small anchors



Kim et al. *Nature Methods* 2015

Novel junctions are typically supported by few alignments



Each curve represents one RNA-seq read mapping protocol (program + settings).

Input: sequence reads (FASTQ format)

```
@HWI-ST1018:7:1101:16910:46835#0/1
CTTCATTTCCCTCCAGTCCCTGGAGGGGCTTCTAGTATTACTGGGACAATGACCACGCTGCCTGTTTGTCTGTGAGTTACGGGCAACCAGCCTCTTCAGCC
+
bbbeeeefgggghiiiiiiiiiiiiiiiihihihiiiihiiiiiiiiiiiiiiiiiiiiiggggdeeeebdddcbbccccccccccccccccccdbbX
```

Output: reads mapped to genome (SAM format)

```
HWI-ST1018:7:1101:16910:46835#0 97      chr1      150812084      255      96M5S      chr2      73300602
```

Initial steps in RNA-seq data processing

(for species with a reference genome)

1. Quality checks on reads
2. Index reference genome
3. Map reads to genome (output in SAM or BAM format)
4. Convert results to a sorted, indexed BAM file
5. Quality checks on mapped reads
6. Visualize read mappings on the genome

Followed by further analyses...

Browsing your results

Two main browsers:

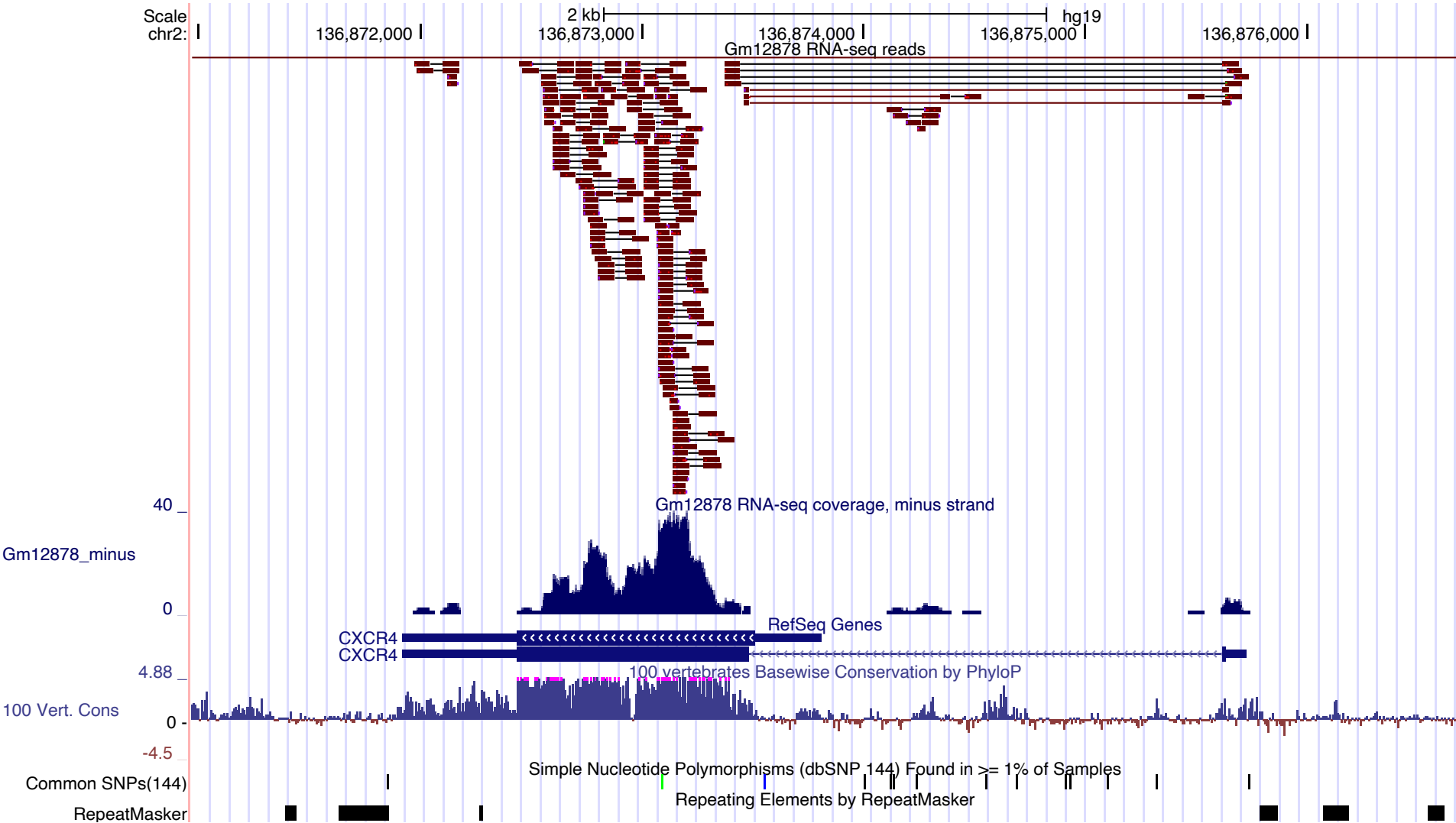
Integrative Genomics Viewer (IGV)

- + Fast response (runs locally)
- + Easy to load your data (including custom genomes)
- Limited functionality
- User interface issues

UCSC Genome Brower

- Sluggish (remote web site)
- Need to place data on web server (e.g. UPPMAX webexport)
- + Much public data for comparison
- + Good for sharing your data tracks (e.g. using track hubs)

Visualization of read alignments



Recommendations

- Use STAR, HISAT2 or GSNAP
- STAR and HISAT2 are the fastest
- HISAT2 uses the least memory
- Consider 2-pass read mapping (default in HISAT2 and TopHat2)
 - No need to supply annotation to mapper
 - Check that junction discovery criteria are conservative
- HISAT2 and GSNAP can use SNP data, which may give higher sensitivity
- For long (PacBio) reads, STAR, BLAT or GMAP can be used
- Don't trust novel introns supported by single reads
- Always check the results!



Thank you. Questions?

Johan Reimegård | 30-November-2020