RNA-seq read mapping

Pär Engström

SciLifeLab RNA-seq workshop

March 2017

Enabler for Life Sciences

Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet

SciLifeLab

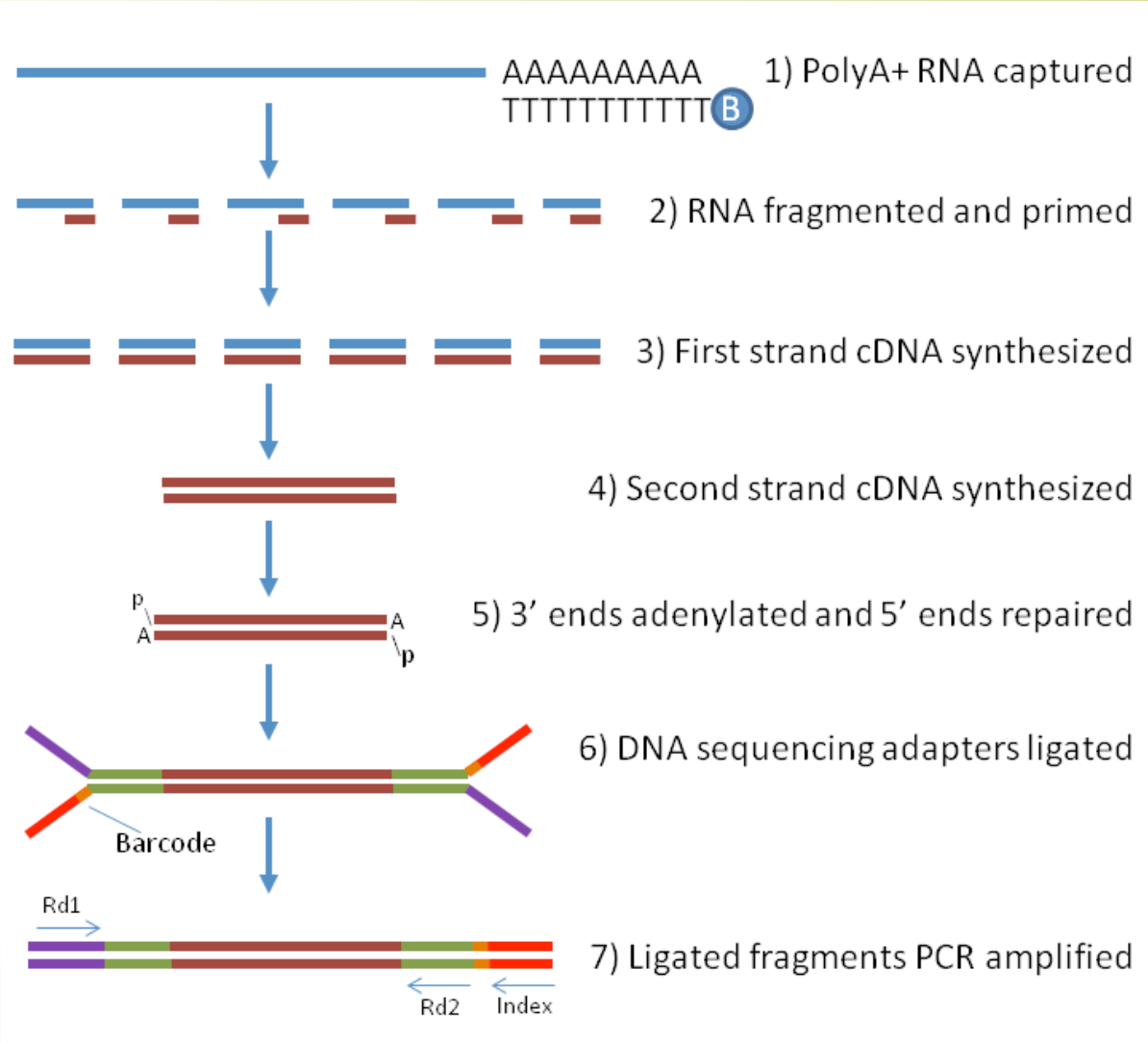
Initial steps in RNA-seq data processing

(for species with a reference genome)

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

Followed by further analyses…

RNA-seq library preparation



http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

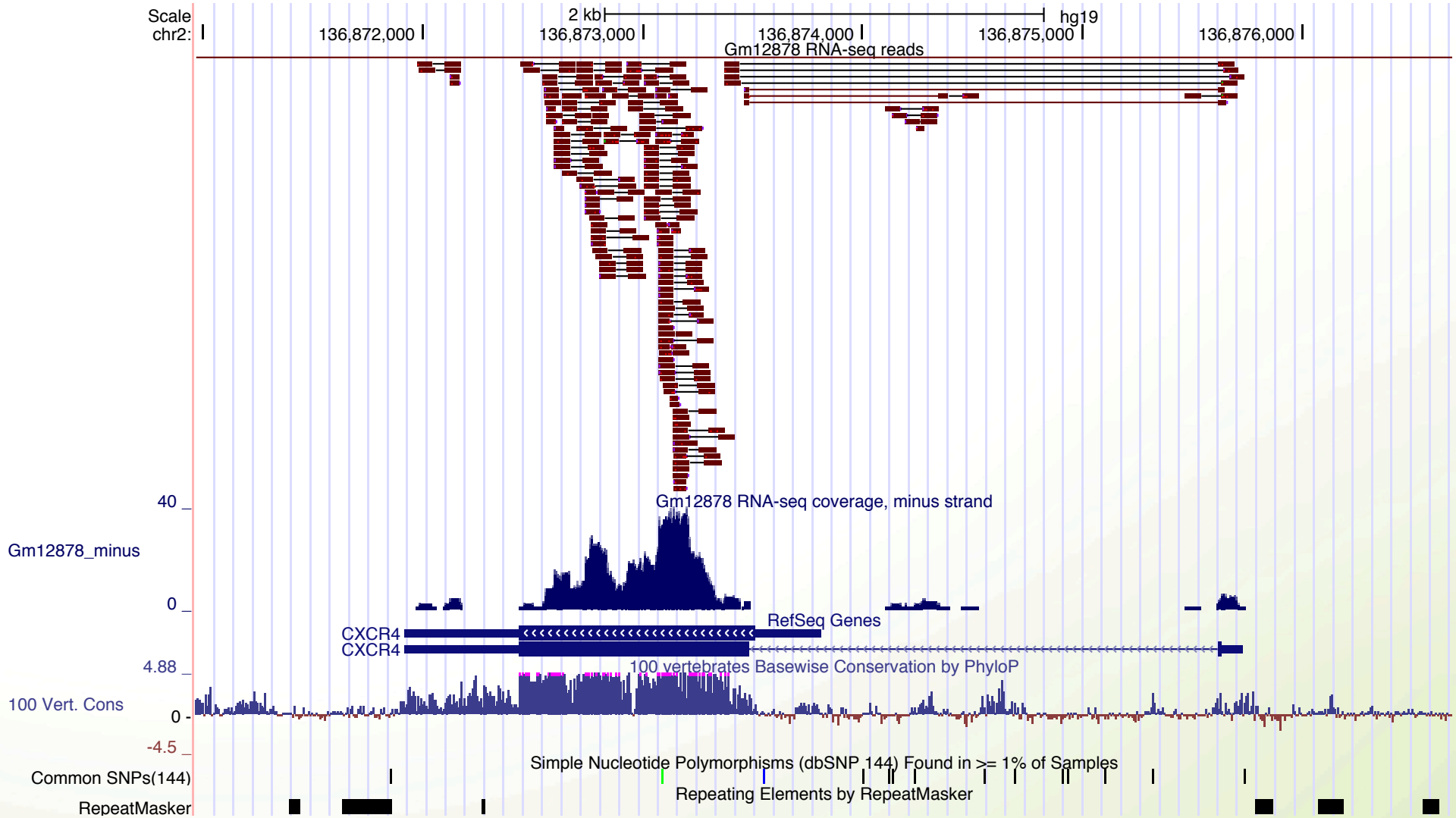
Input: sequence reads (FASTQ format)

```
@HWI-ST1018:7:1101:16910:46835#0/1
CTTCATTTCCCTCCAGTCCCTGGAGGGGCTTCTAGTATTACTGGGACAATGACCACGCTGCCTGTTTGTCTGTGAGTTACGGGCAACCAGCCTC
+
bbbeeeeefgggghiiiiiiiiiiiiiiiiihihihhiiiihiiiiiiiihiiiiiiiiigggdeeeebddddcbbbcccccccccccacccc
@HWI-ST1018:7:1101:2937:53143#0/1
CGACCAGCTGATCGTGTCTCCAAGGGCAGAAGCACAAGCGGGGAGGCTGGGGTGGCTGCAGCGAGGTCCTCCCTAAGTAGGGCAGGGGAGCCCC
+
bbbeeeeeggfggihihiiiiiiiiiihiiiihiiiihihigadcccdcccZaa^^_acccc_ac_bccccbb^bYabbcbc]a]aET]aca
@HWI-ST1018:7:1101:14544:66521#0/1
GGTGGCTGCAGCGAGGTCCTCCCTAAGTAGGGCAGGGGAGCCCCCAGGTGGGGAGGGCTCATGGGGGCCAGGGAGTAAGGCTGGCTCCCCTGGT
+
bbaeeeeeggggiifghiiiiiihfhfhihiifhigihhiiihigggdcecc^accccccccccaccccccccac^b_bcbccccbbaacba`Y
@HWI-ST1018:7:1101:15405:122666#0/1
CCCACCTGCAACTTTCCTCCAAGTGTGGCTCGGAGAAGAAACATCAACAAGGACCCTGGGCTTCGATTCAAAAACTCCTCTGAAGCCATCCATG
+
bbbeeeeeggggiiiiiiihiigieghiii_eU_^cbceghffdhhiiicg`\XaZ`ggcdecebcdbb`bcaW_]bbbb]bbbbcbc^`bbb
@HWI-ST1018:7:1101:14326:133684#0/1
CGCCTGCCCAGCAGTGTTTATCCTGGGATCCTCCTATTGGGGTTGAGGGAGGGGAAGACAGCAGGAAGGTTGAGGGAGCAGCAACTTGGCCAGA
+
^\\cccc^Y[Ybee^bfcegagX_^aeehhheebZPbf_RZeO^_ea]`Ye`[WYY^Q_Xab]ZZ^Z\_aY[GY^aNROW^PQXQX`a`XY`P^
...
```

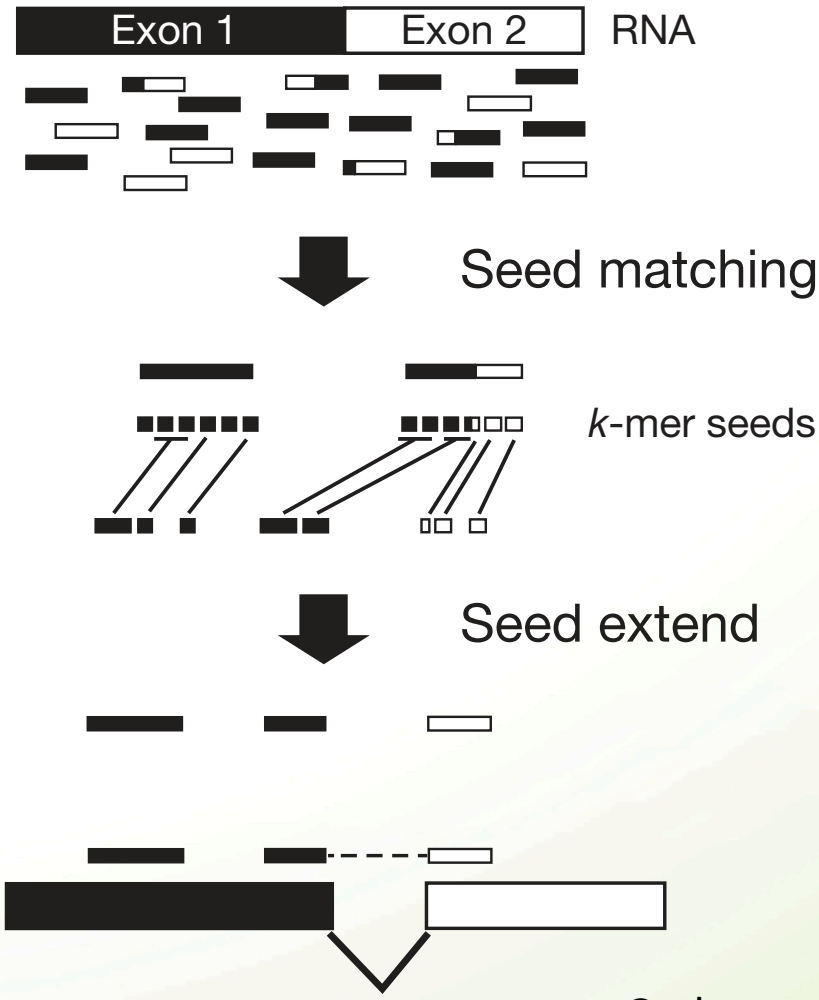
Goal: reads mapped to genome (SAM format)

```
HWI-ST1018:7:1206:3667:137198#0 97      chr1    150812084       255     47M2769N47M7S   chr2
HWI-ST1018:7:2305:11836:132357#0        177     chr12   13070344        255     11S90M  chr2
HWI-ST1018:7:1205:18018:8988#0  97      chr12   51637109        255     96M5S   chr2    733025
HWI-ST1018:7:1103:2457:70159#0  129     chr19   45504799        255     101M    chr2    733155
HWI-ST1018:7:1107:14230:146505#0        99      chr2    73300510        255     101M    =
HWI-ST1018:7:1106:16800:63390#0 163     chr2    73300524        255     101M    =       733006
HWI-ST1018:7:2306:19900:62130#0 99      chr2    73300547        255     101M    =       733007
HWI-ST1018:7:2305:8697:195892#0 163     chr2    73300561        255     4S97M   =       733006
HWI-ST1018:7:1208:10024:50258#0 99      chr2    73300563        255     98M3S   =       733006
HWI-ST1018:7:1107:14230:146505#0        147     chr2    73300572        255     101M    =
HWI-ST1018:7:1208:10123:71500#0 99      chr2    73300593        255     101M    =       733006
HWI-ST1018:7:2107:11555:46214#0 163     chr2    73300593        255     101M    =       733006
HWI-ST1018:7:1102:12130:87067#0 73      chr2    73300594        255     101M    =       733005
HWI-ST1018:7:1102:12130:87067#0 133     chr2    73300594        0       *       =       733005
HWI-ST1018:7:1206:3667:137198#0 145     chr2    73300602        255     101M    chr1    150812
HWI-ST1018:7:1208:16138:88503#0 99      chr2    73300603        255     101M    =       733007
HWI-ST1018:7:2206:7742:86872#0  163     chr2    73300621        255     101M    =       733006
HWI-ST1018:7:1308:14606:19516#0 99      chr2    73300623        255     1S100M  =       733008
HWI-ST1018:7:2301:14871:81110#0 99      chr2    73300623        255     101M    =       733007
HWI-ST1018:7:2201:13683:64077#0 145     chr2    73300623        255     11S90M  =       733006
...
```

Visualization of read alignments

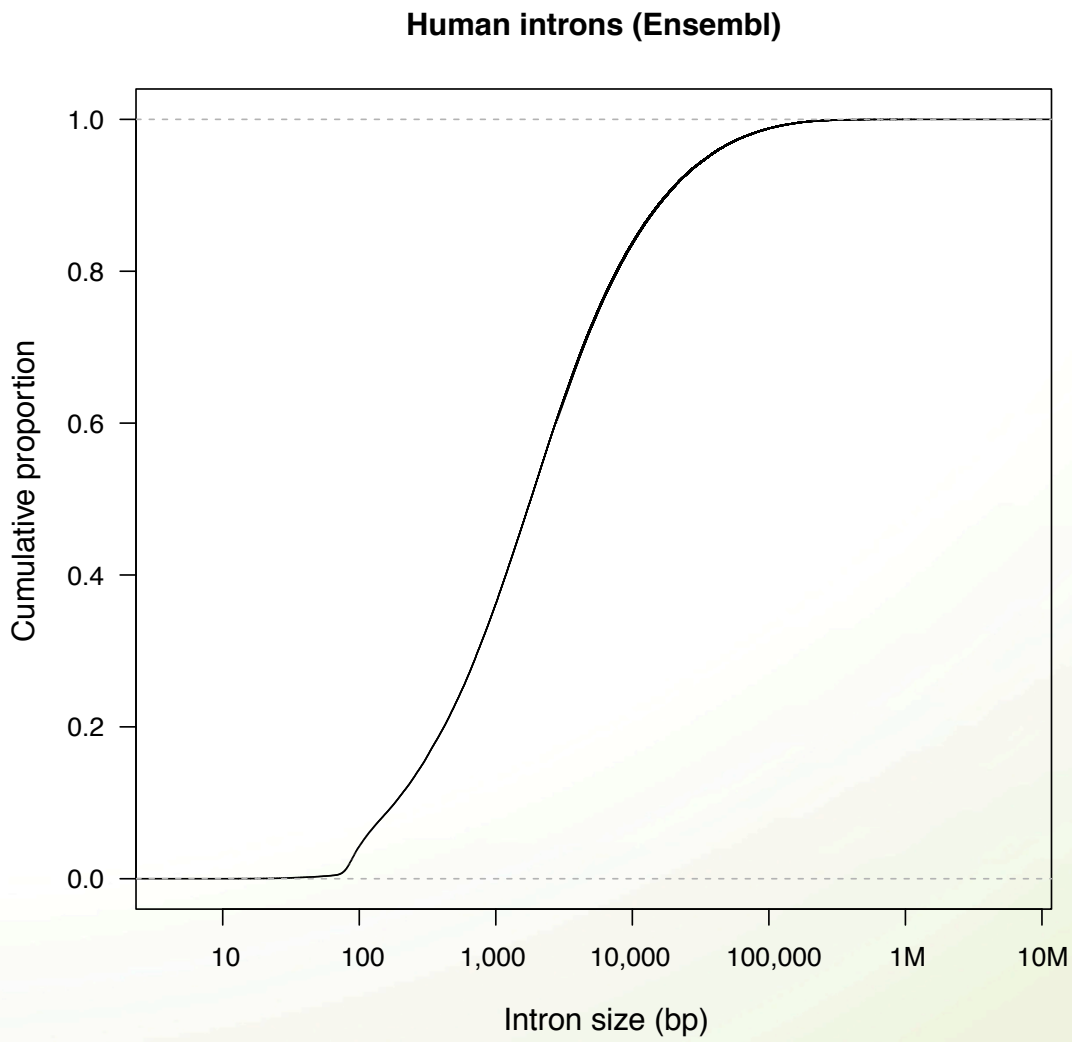


Spliced alignment

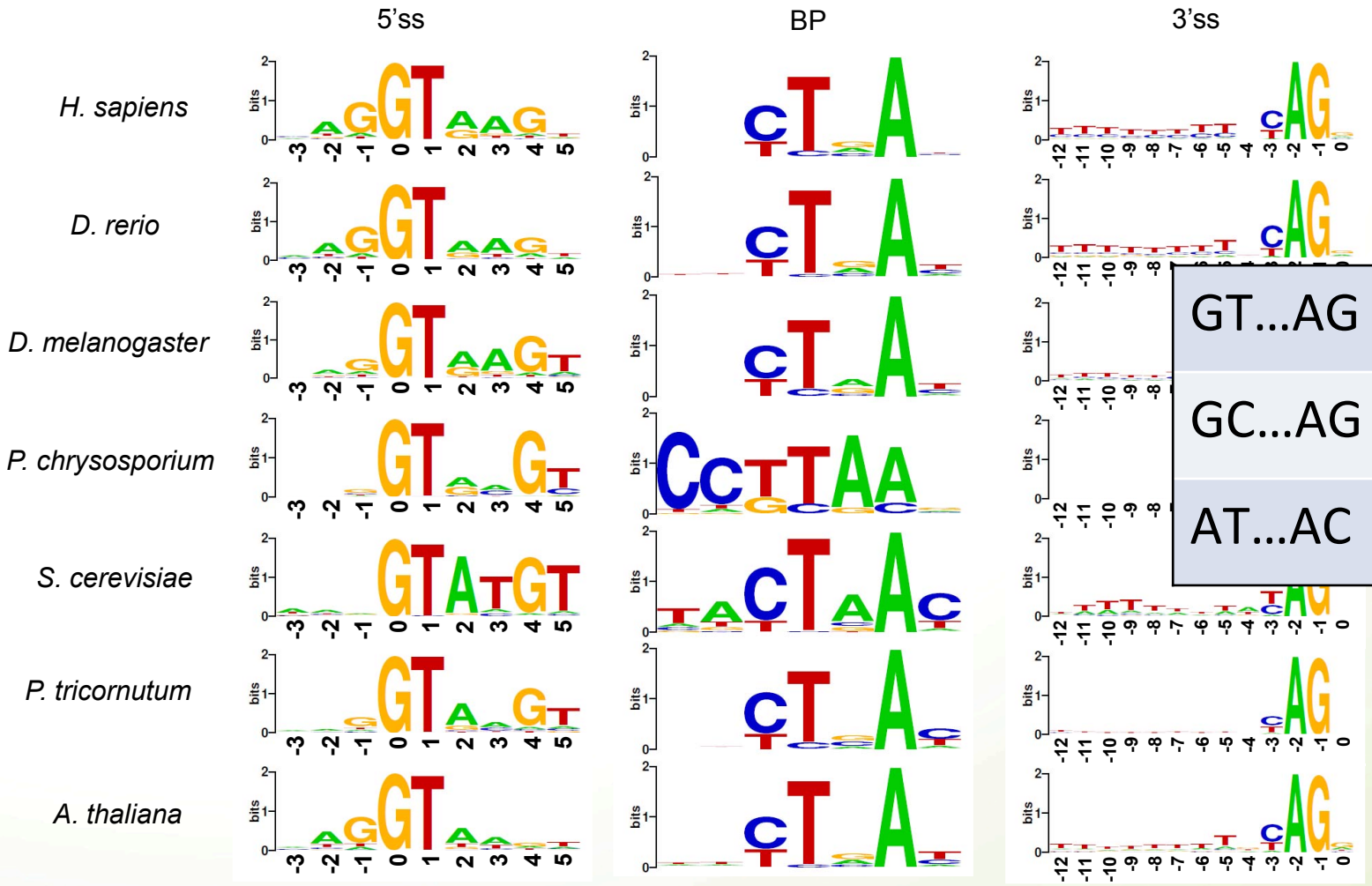


Garber et al. *Nature Methods* 2011

Introns can be very large!



Limited sequence signals at splice sites



GT…AG	98.6%
GC…AG	1.3%
AT…AC	0.1%

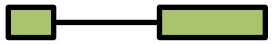
Iwata and Gotoh *BMC Genomics* 2011

Karolinska Institutet
KTH Vetenskap och konst
Stockholms universitet
Uppsala Universitet
SciLifeLab

Multi-mapping reads and pseudogenes

Functional gene

Processed pseudogene



Correct read alignment
Identical, spliced

Incorrect read alignment
Mismatches, not spliced

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. *Genome Biology* 2013
HISAT2	Kim et al. *Nature Methods* 2015
STAR	Dobin et al. *Bioinformatics* 2013
GSNAP	Wu and Nacu *Bioinformatics* 2010
OLego	Wu et al. *Nucleic Acids Research* 2013
HPG aligner	Medina et al. *DNA Research* 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.

Kim et al. *Nature Methods* 2015

The predecessor: BLAT

"In the process of assembling and annotating the human genome, I was faced with two very large-scale alignment problems: aligning three million ESTs and aligning 13 million mouse whole-genome random reads against the human genome. These alignments needed to be done in less than two weeks' time on a moderate-sized (90 CPU) Linux cluster in order to have time to process an updated genome every month or two. To achieve this I developed a very-high-speed mRNA/DNA and translated protein alignment algorithm. "

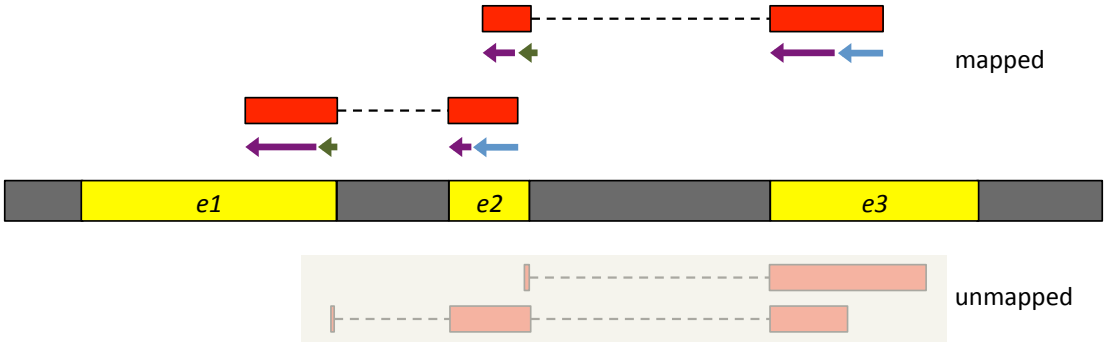
(Kent *Genome Research* 2002)

Innovations in RNA-seq alignment software

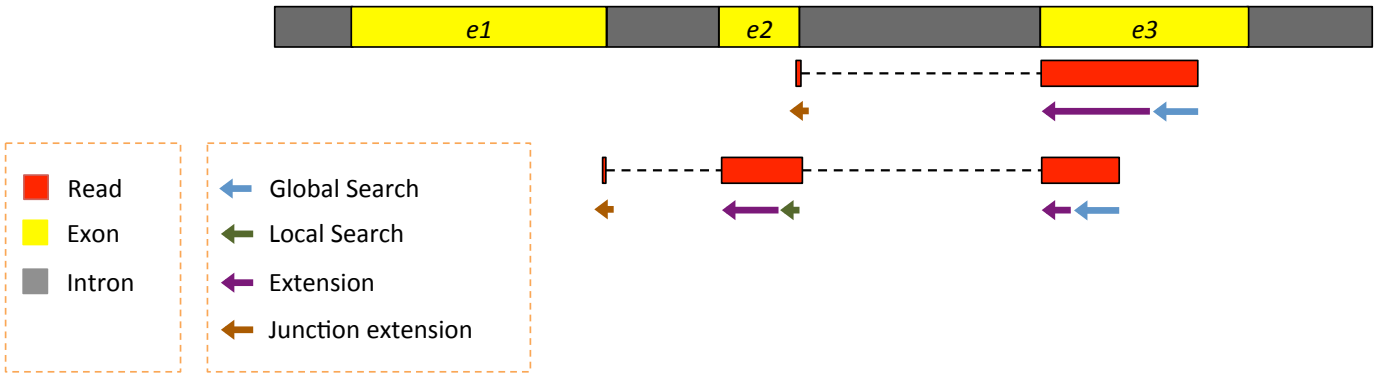
- Read pair alignment
- Consider base call quality scores
- Sophisticated indexing to decrease CPU and memory usage
- Map to genetic variants
- Resolve multi-mappers using regional read coverage
- Consider junction annotation
- Two-step approach (junction discovery & final alignment)

Two-step RNA-seq read mapping

1st run of HISAT to discover splice sites

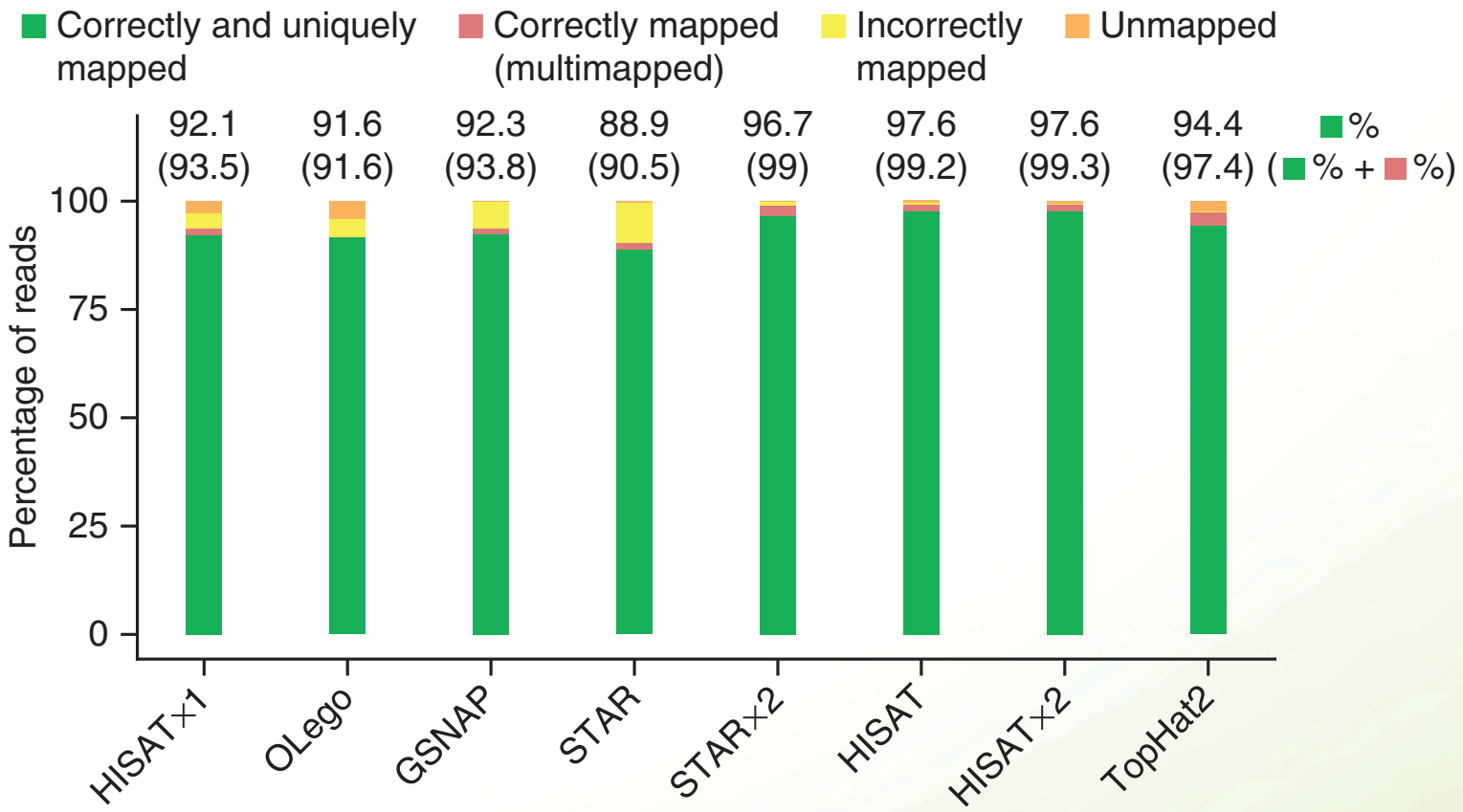


2nd run of HISAT to align reads by making use of the list of splice sites collected above



Kim et al. *Nature Methods* 2015

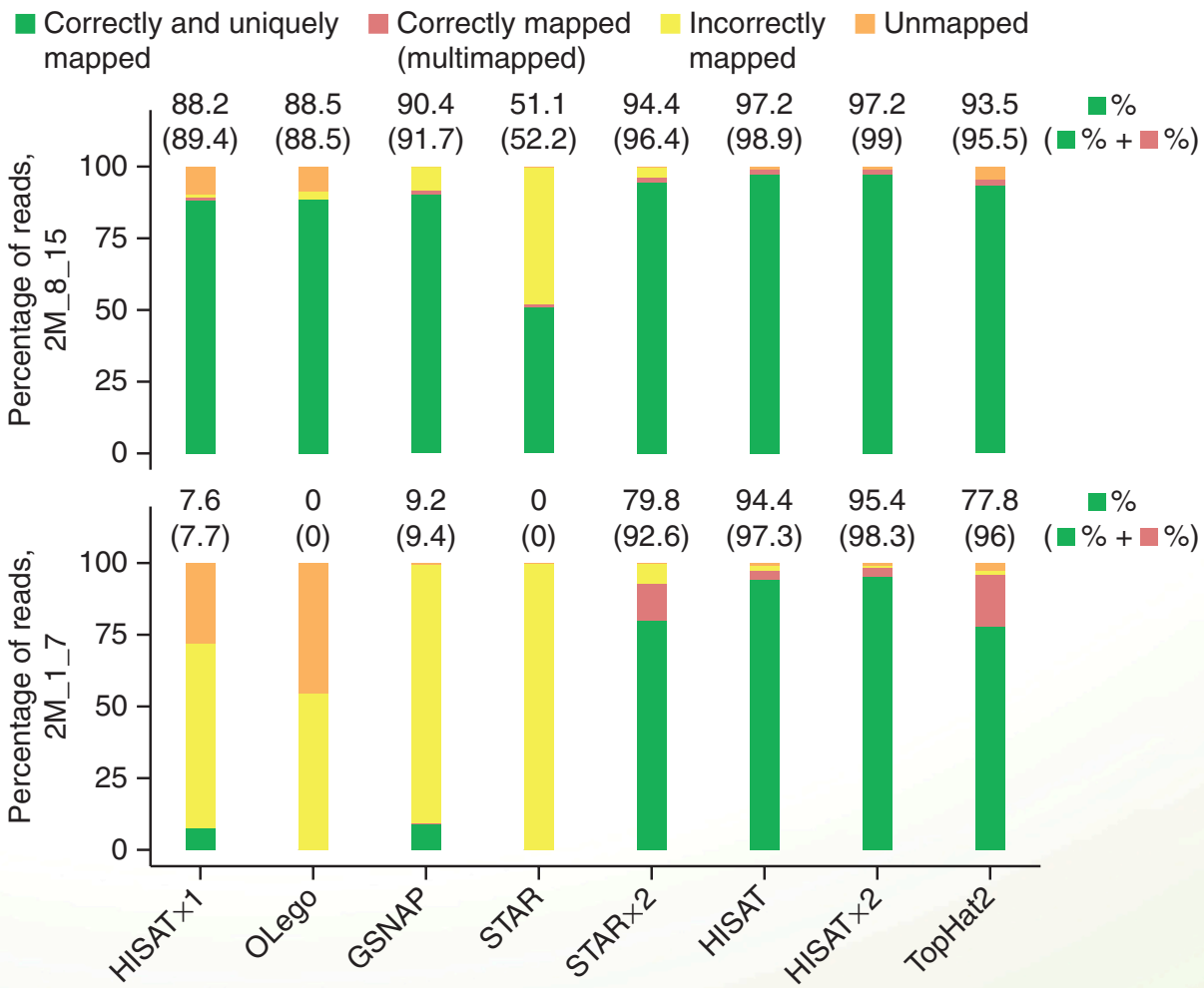
Mapping accuracy

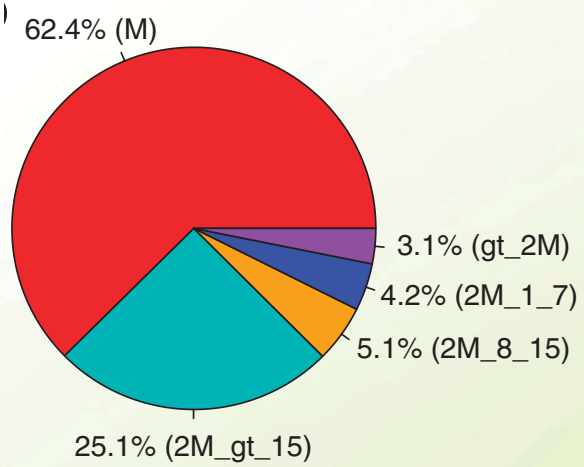


Accuracy for 20 million simulated human 100 bp reads with 0.5% mismatch rate

Kim et al. *Nature Methods* 2015

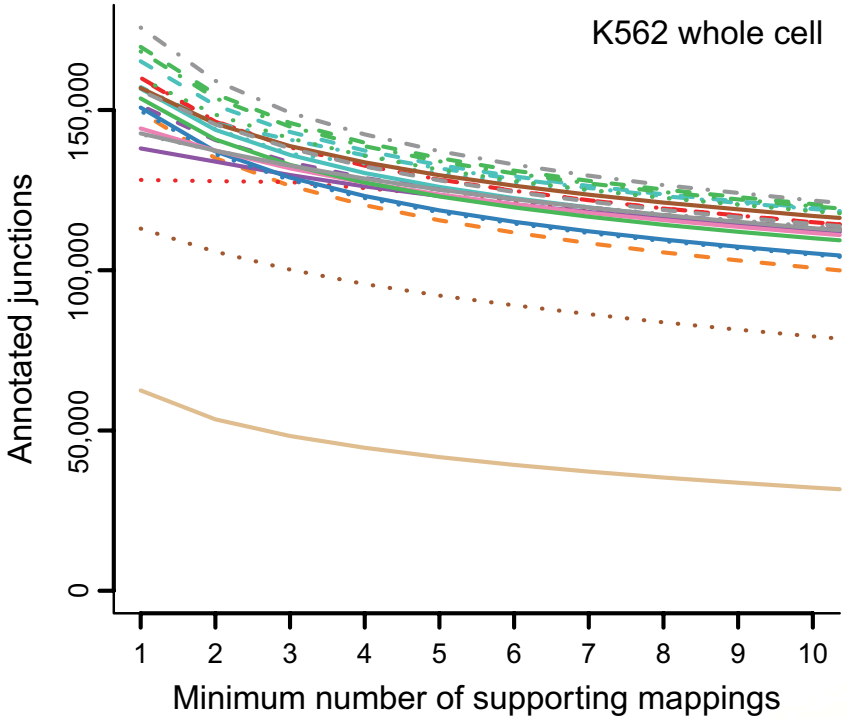
Mapping accuracy for reads with small anchors

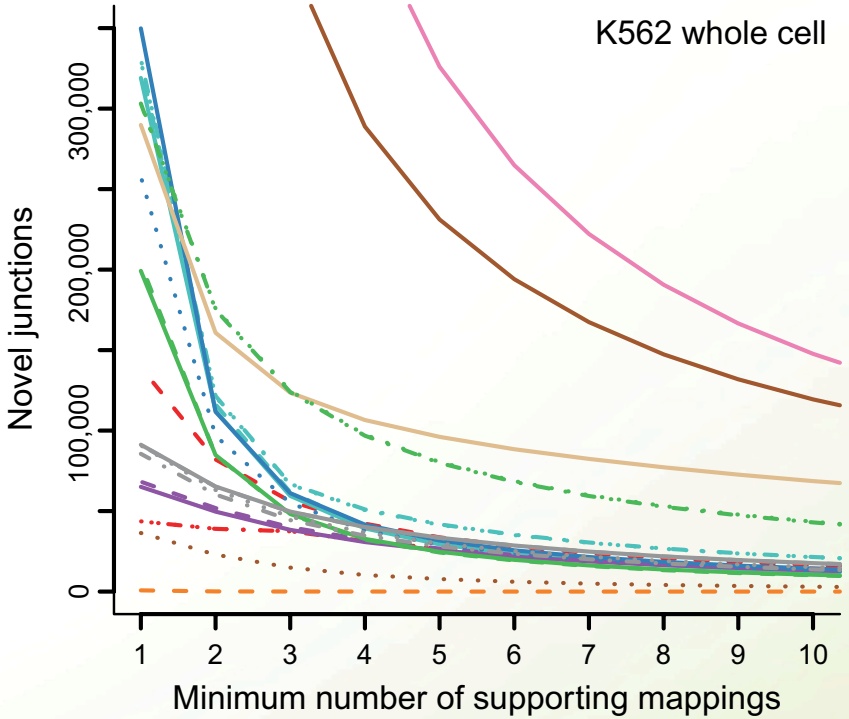




Kim et al. *Nature Methods* 2015

Novel junctions are typically supported by few alignments

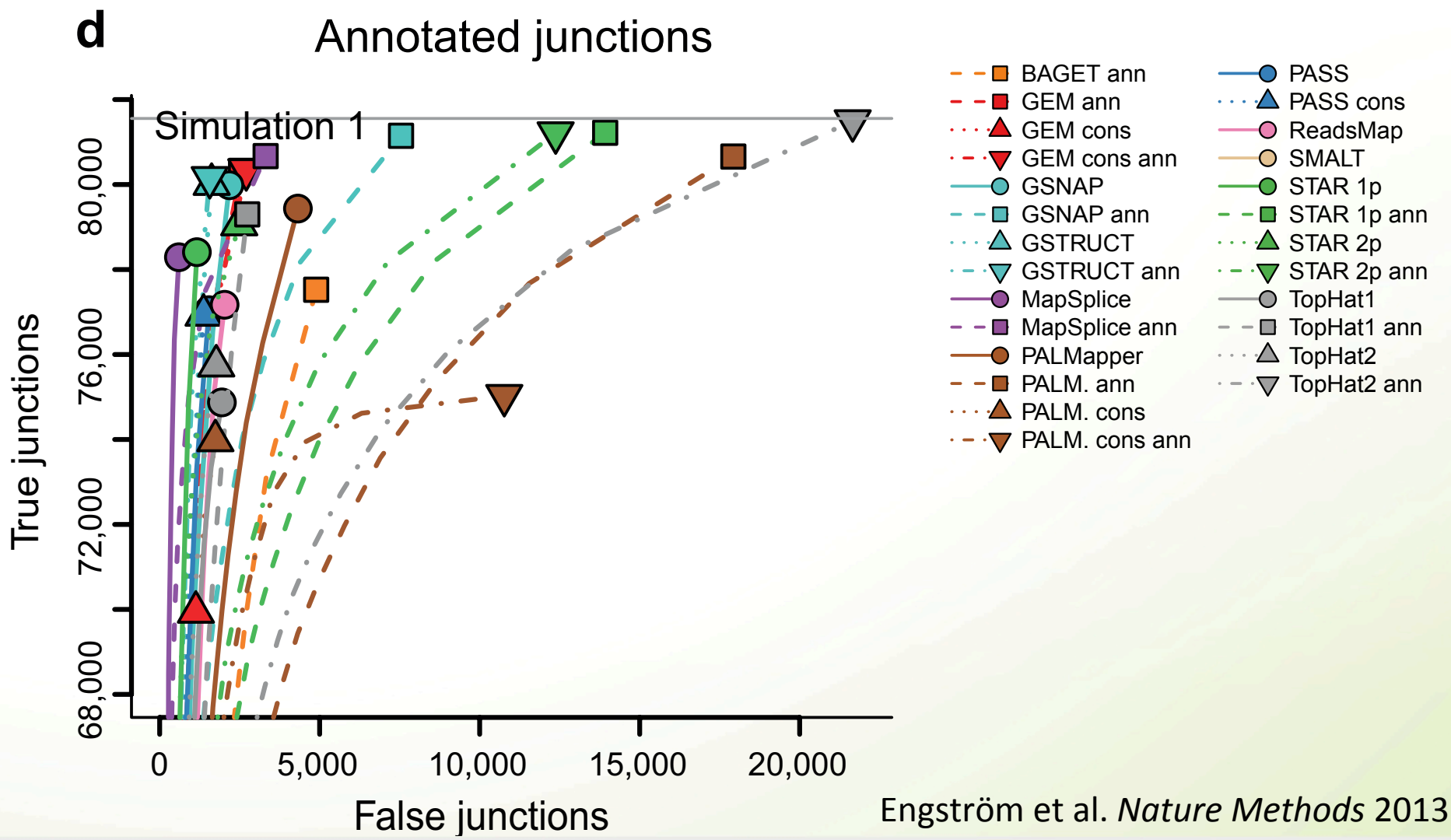




Each curve represents one RNA-seq read mapping protocol (program + settings).

Engström et al. *Nature Methods* 2013

Several methods show over-confidence in annotation



Engström et al. *Nature Methods* 2013

Recommendations

- Use STAR, HISAT2 or GSNAP
- STAR and HISAT2 are the fastest
- HISAT2 uses the least memory
- If you want to run Cufflinks, use TopHat2 (but don't)
- 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!

Inspecting a BAM file

Command:

```
samtools view -X file.bam
```

Paper:

Li et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-9

SAM format specification:

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

Visualizing reads mapped to genome

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

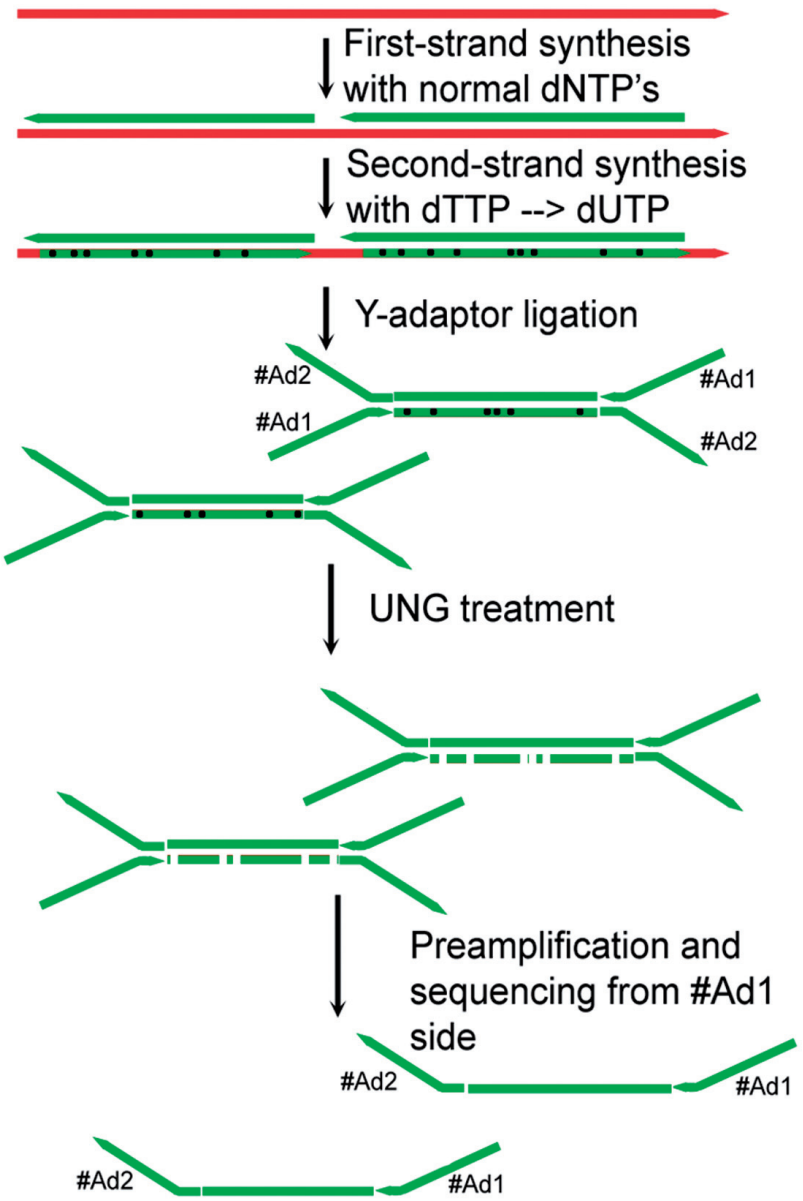
− Sluggish (remote web site)
− Need to place data on web server (e.g. UPPMAX webexport)
+ Much public data for comparison
+ Good for sharing your data tracks (e.g. using track hubs)

Unsolved problems in RNA-seq read mapping

- Determine correct location of multimapping reads
- Accurate alignment of indels
- Use gene annotation in an unbiased fashion
- Cross-species mapping

Thanks for listening!

The dUTP method for strand-specific RNA-seq



Parkhomchuk et al. *Nucleic Acids Research* 2011
Borodina et al. *Methods in Ezymology* 2011

SciLifeLab

Important SAM fields

Command:

```
samtools view -X file.bam
```

Perfectly and uniquely aligned read pair:

```
HWI-ST1018:3:1305:21090:45397#0  pPR1   chr1  4426  255  101M        =  4435   110  GT…  C@…
NH:i:1  HI:i:1  AS:i:200  nM:i:0

HWI-ST1018:3:1305:21090:45397#0  pPr2   chr1  4435  255  101M        =  4426  -110  CG…  5<…
NH:i:1  HI:i:1  AS:i:200  nM:i:0
```

Problematic read pair:

```
HWI-ST1018:3:2109:6170:66353#0  pPR2s  chr1  5058    3  65M36S      =  5058    95  CA…  B@…
NH:i:2  HI:i:2  AS:i:135  nM:i:9

HWI-ST1018:3:2109:6170:66353#0  pPr1s  chr1  5058    3  7S73M1D21M  =  5058   -95  CC…  ##…
NH:i:2  HI:i:2  AS:i:135  nM:i:9
```