RNA-seq read mapping

Pär Engström

SciLifeLab RNA-seq workshop

April 2016

Enabler for Life Sciences

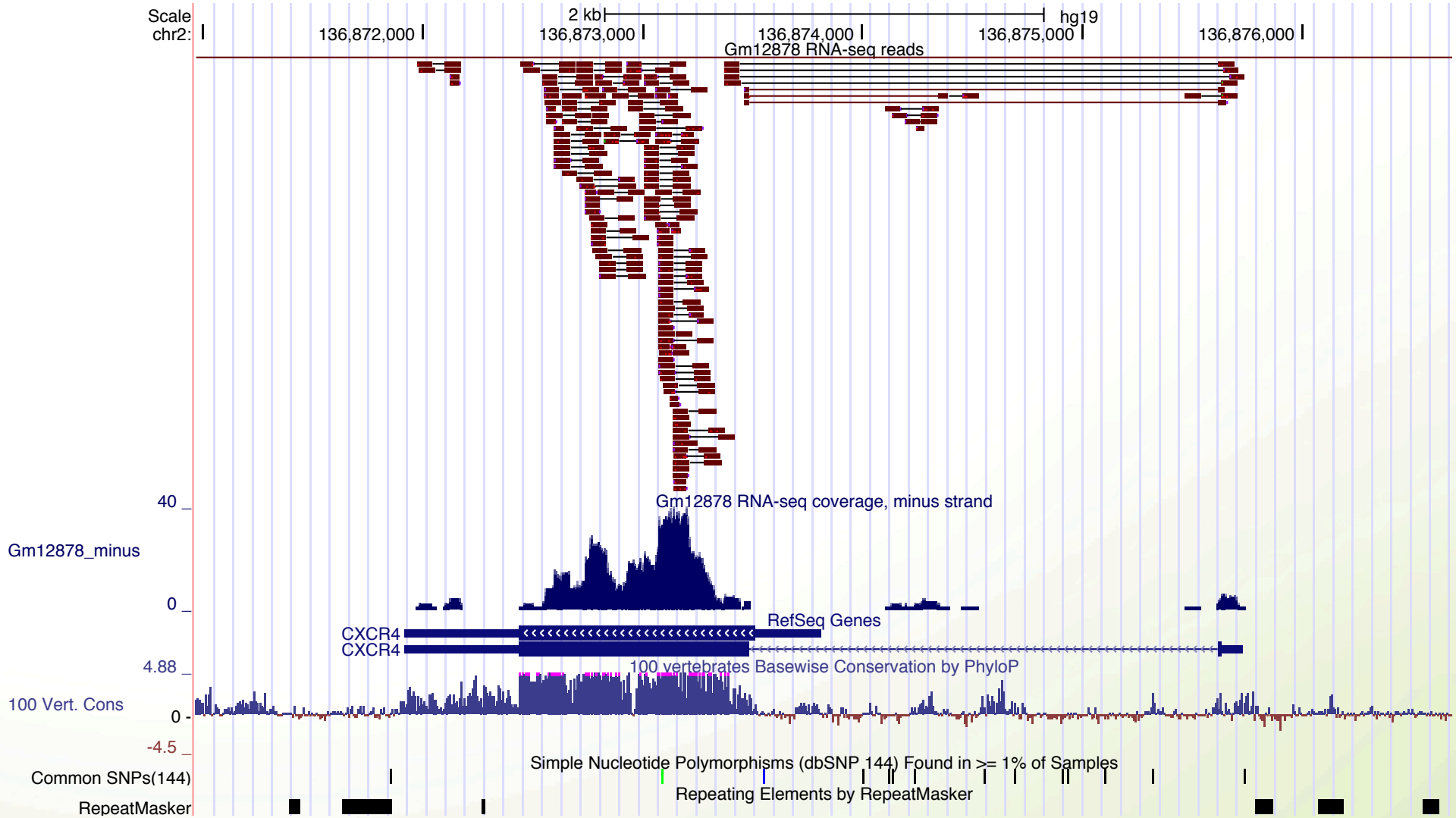
Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet

SciLifeLab

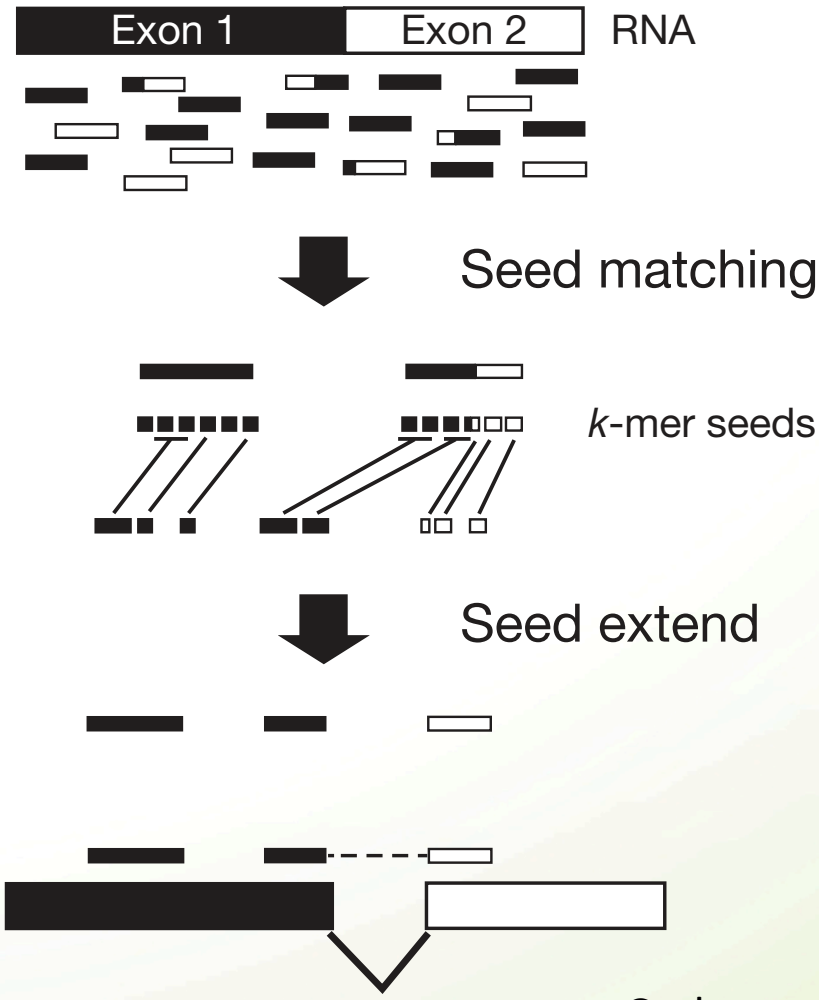
Input: sequence reads (FASTQ format)

```
@D2BJQQN1:142:C03FNACXX:5:1101:12935:2174 1:N:0:GCCAAT
ATGCTGTGCAGGGCCTTGAGAACATGCGGGGGAATACATGTGGGTTTTTGG
+
=??DDDDD2,4<D2+4AE##################################
@D2BJQQN1:142:C03FNACXX:5:1101:13035:2084 1:N:0:GCCAAT
TTTCTATAGTTCGTTACTAGAGAAGTTTTTTTGATTGTGTGGGGGTCCGGG
+
=??DD2=23AD,3CE223,3+3+33<CF+4)0?)*19?DD##########
@D2BJQQN1:142:C03FNACXX:5:1101:13322:2013 1:Y:0:GCCAAT
CGTTCCCGTGGTGGGATTTTTGGGTGGCAGGGGACTTCGGTTGGGGGATTT
+
:>>A2+20<AC@CC)+4@2<+1?AAA########################
@D2BJQQN1:142:C03FNACXX:5:1101:13460:2061 1:Y:0:GCCAAT
CCGCGGTCGGGGGGGGGGGGCGGGGGGGGGGGGGGGGGGGTGGGGTTTTTG
+
=??DDDD)<C):?####################################
…
```

Goal: reads mapped to genome

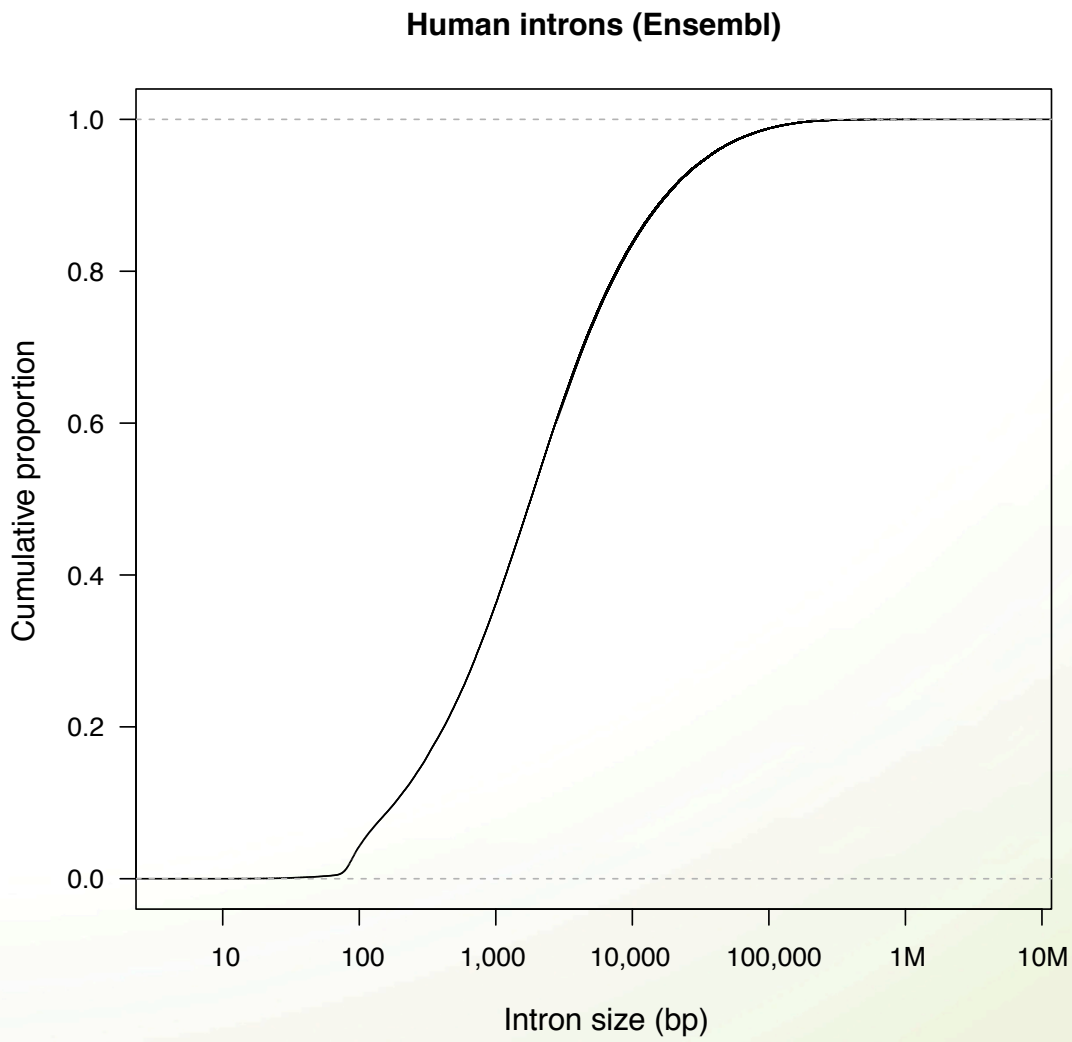


Spliced alignment

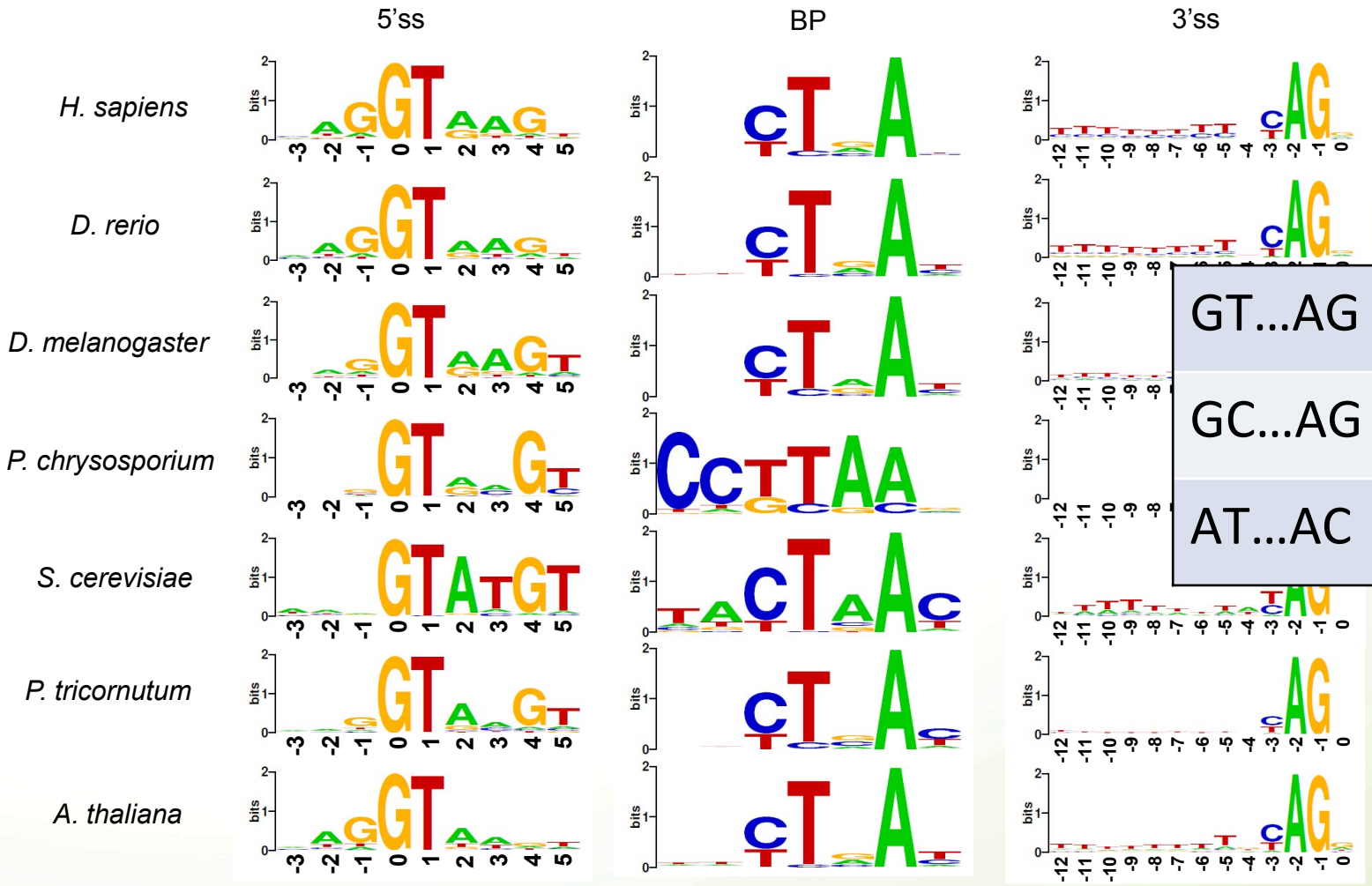


Garber et al. *Nature Methods* 2011

Introns can be very large!



Limited sequence signals at splice sites



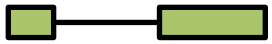
GT…AG	98.6%
GC…AG	1.3%
AT…AC	0.1%

Iwata and Gotoh *BMC Genomics* 2011

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

Current RNA-seq aligners

TopHat2	Kim et al. *Genome Biology* 2013
HISAT & 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

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. "

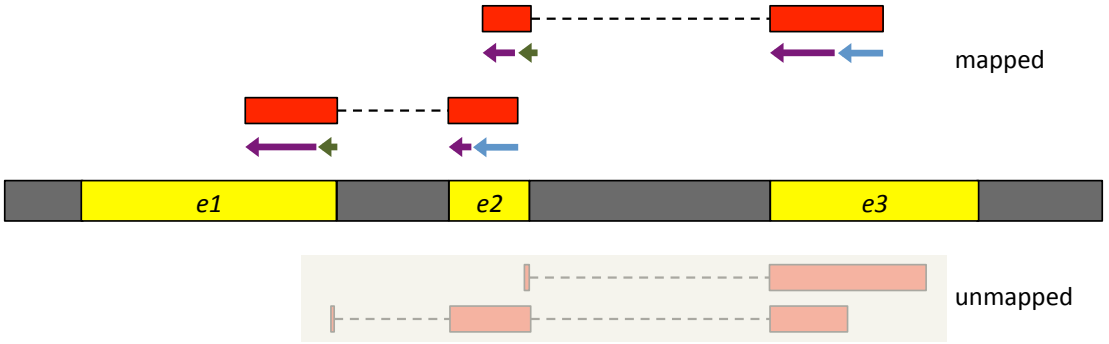
(Kent *Genome Research* 2002)

Innovations in RNA-seq alignment software

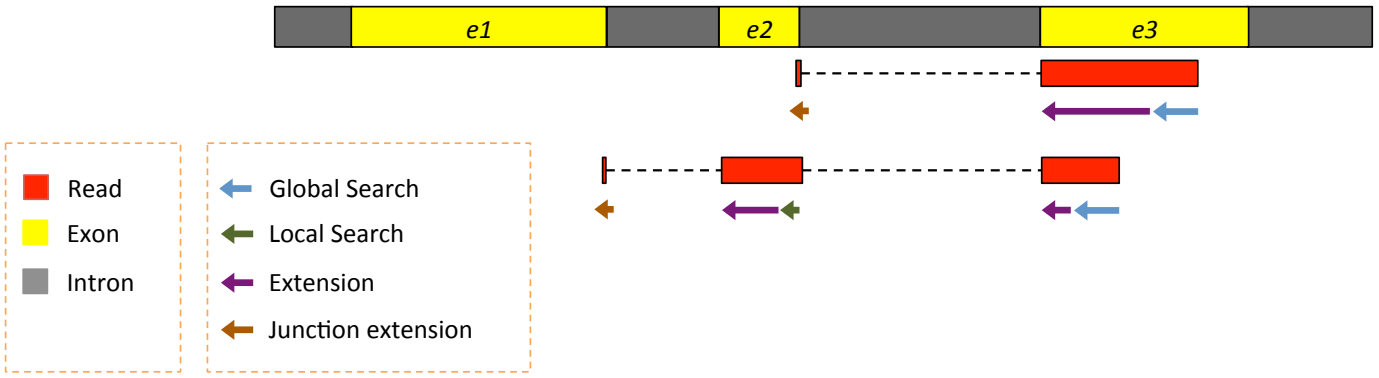
- Read pair alignment
- Consider base call quality scores
- Sophisticated indexing to decrease CPU and memory usage
- Map to genetic variants
- Resolve multi-mappers using regional read coverage
- Consider junction annotation
- Two-step approach (junction discovery & final alignment)

Two-step RNA-seq read mapping

1st run of HISAT to discover splice sites

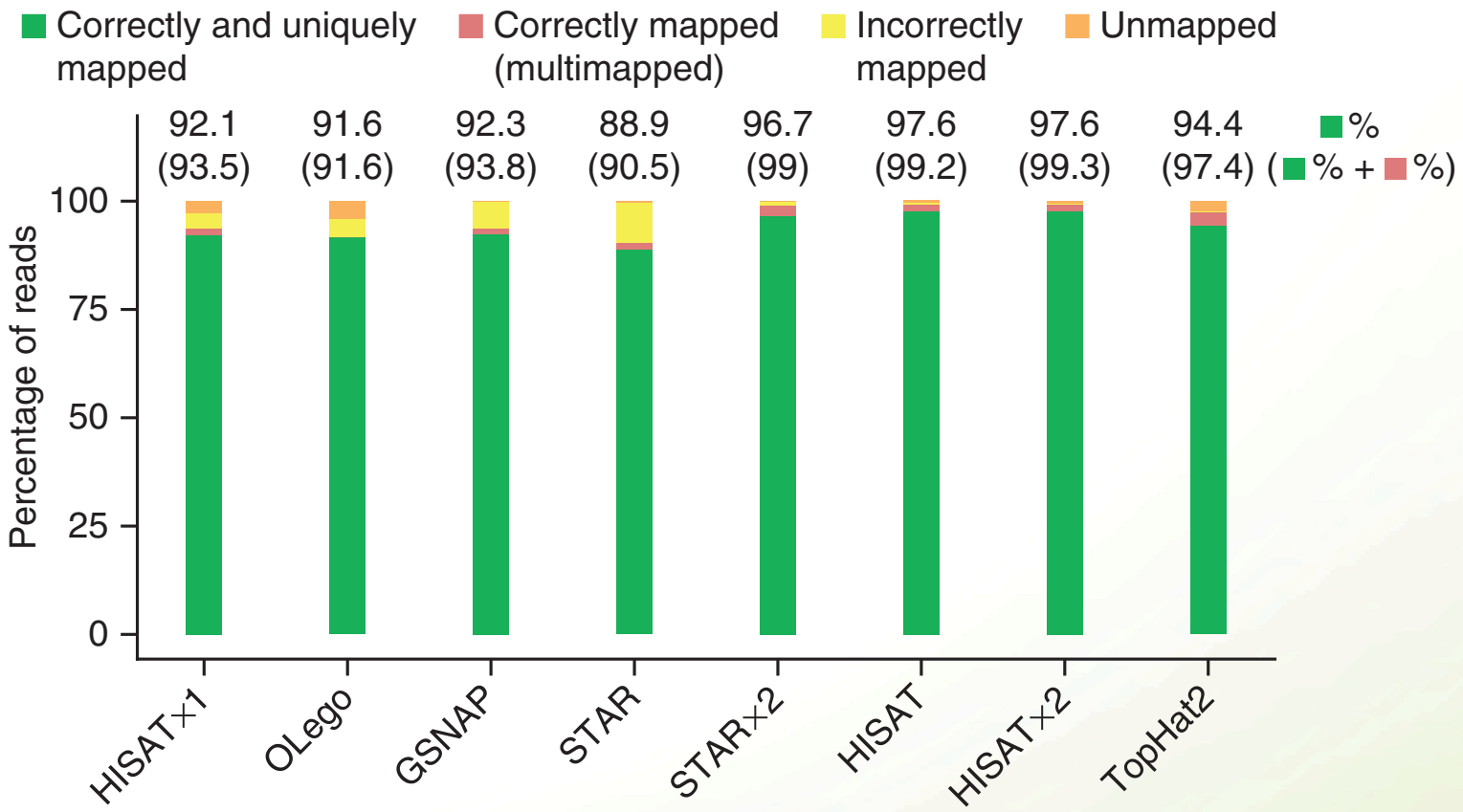


2nd run of HISAT to align reads by making use of the list of splice sites collected above



Kim et al. *Nature Methods* 2015

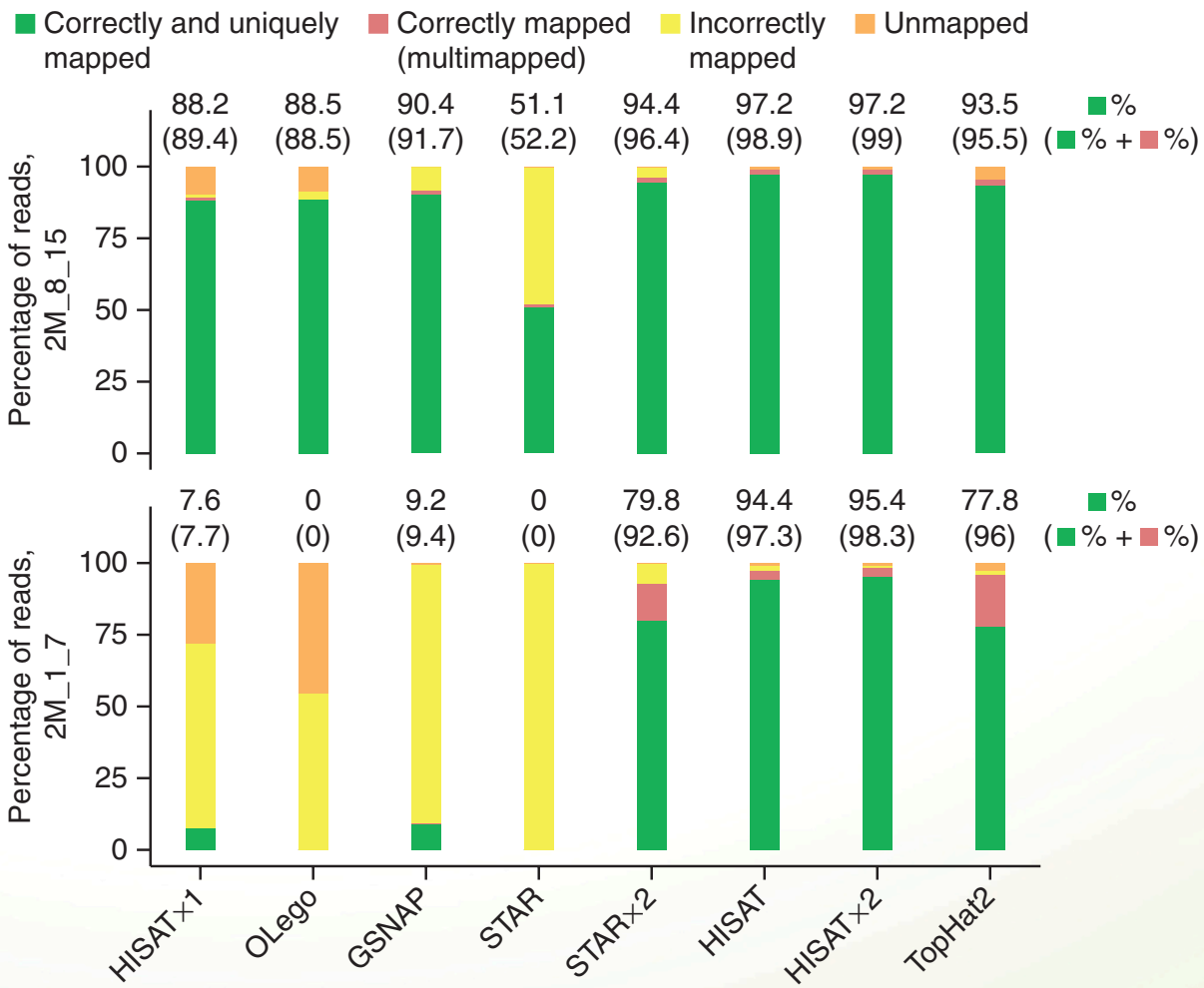
Mapping accuracy

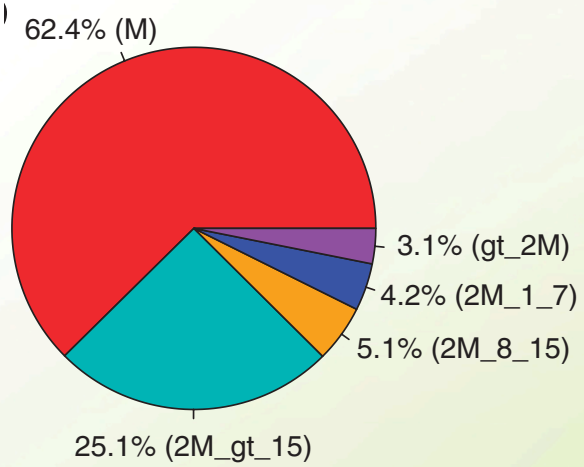


Accuracy for 20 million simulated human 100 bp reads with 0.5% mismatch rate

Kim et al. *Nature Methods* 2015

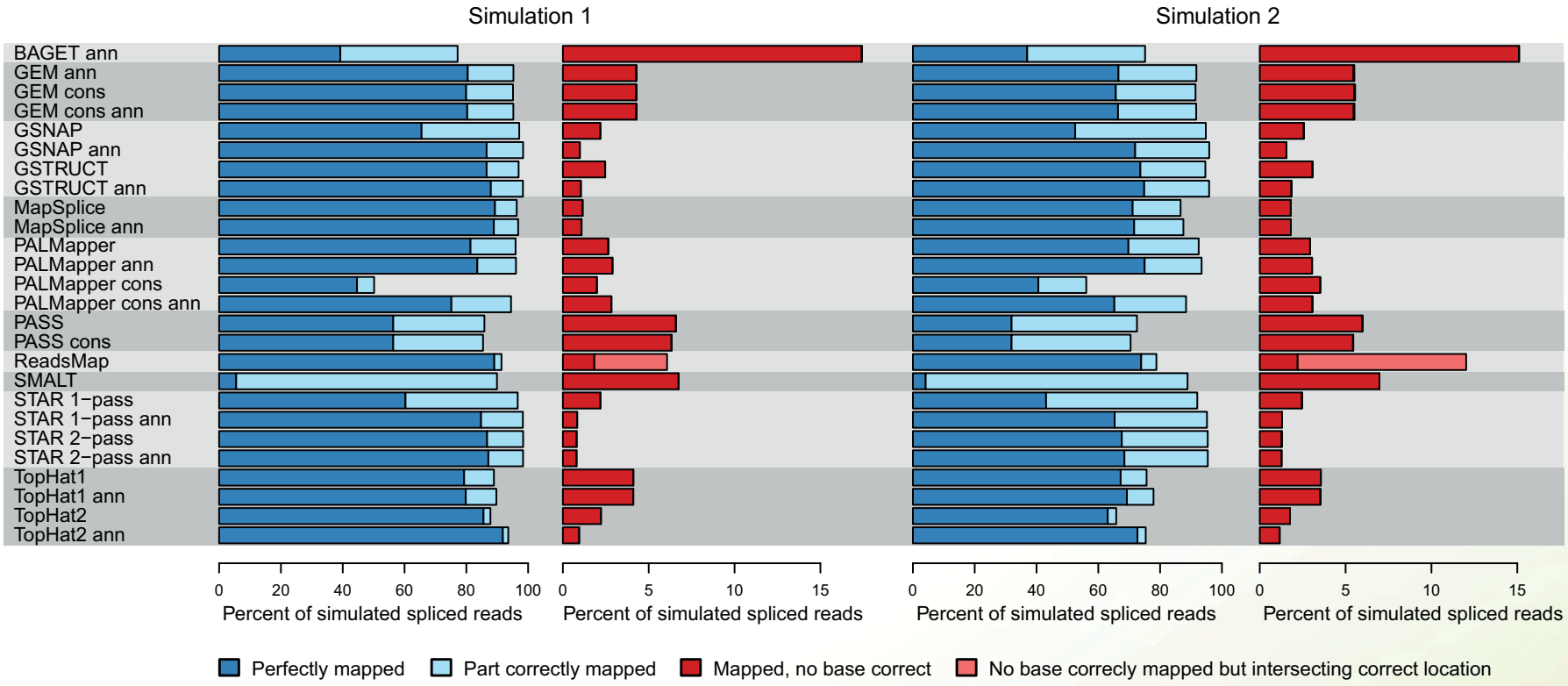
Mapping accuracy for reads with small anchors





Kim et al. *Nature Methods* 2015

Mapping accuracy for spliced RNA-seq reads



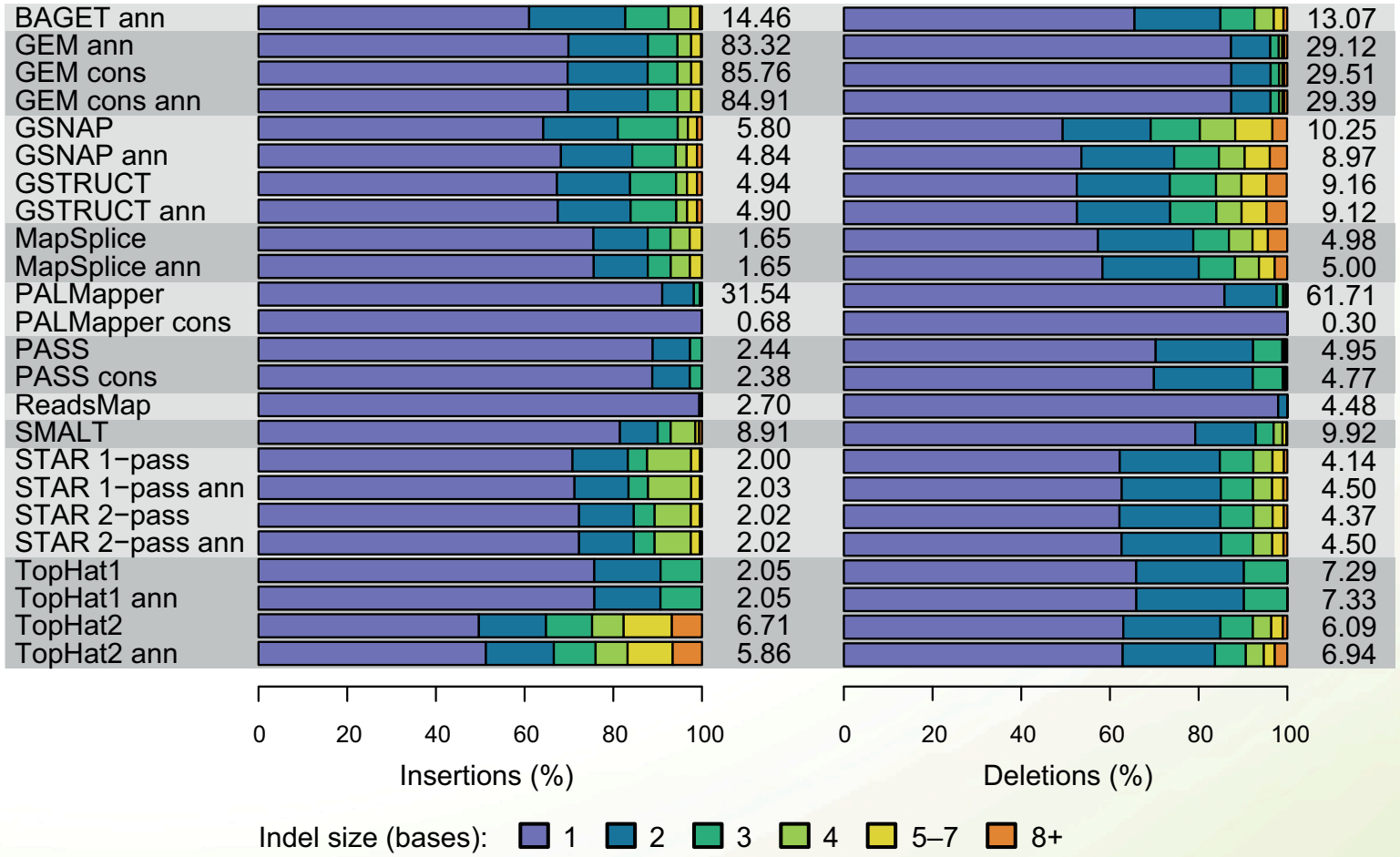
High accuracy at mapping to correct locus: GSNAP, GSTRUCT, MapSplice, STAR

High rate of perfect spliced alignments: ReadsMap, TopHat2 ann

Engström et al. *Nature Methods* 2013

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

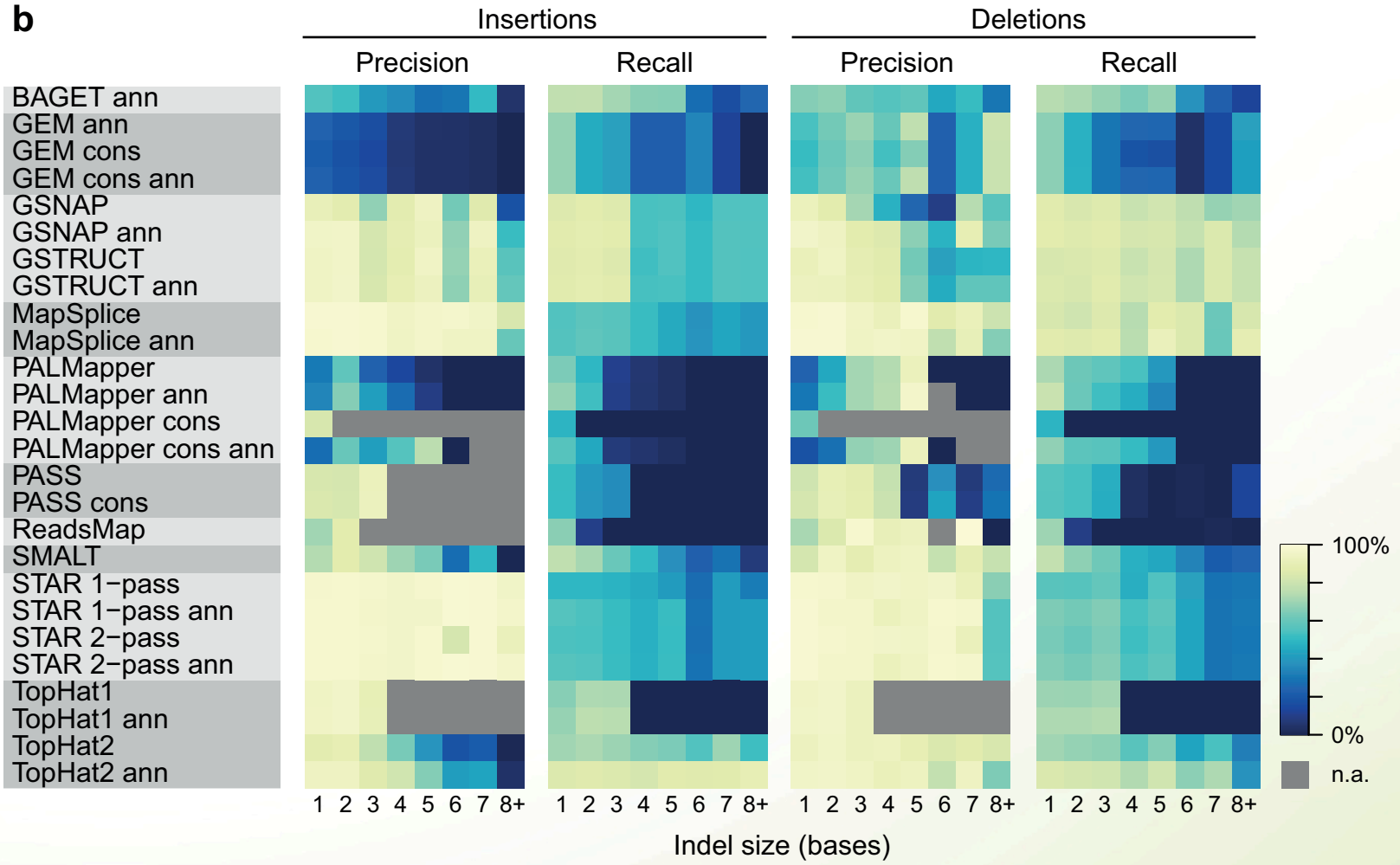
Major differences in indel frequencies



Indel frequencies are tabulated (number of indels per thousand sequenced reads). Data set: K562 (mean).

Engström et al. *Nature Methods* 2013

Indel accuracy on simulated data

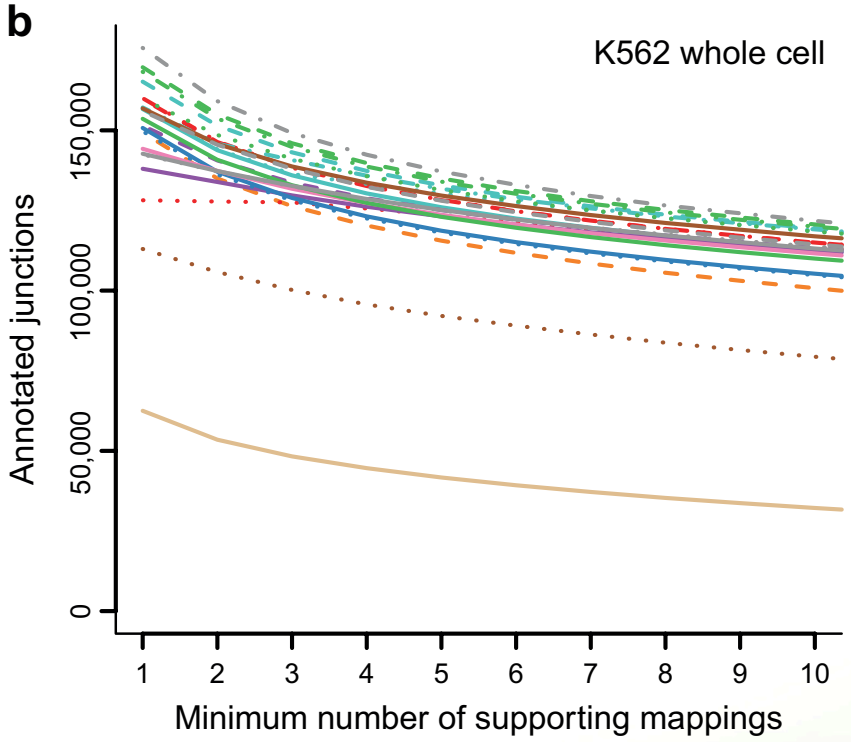


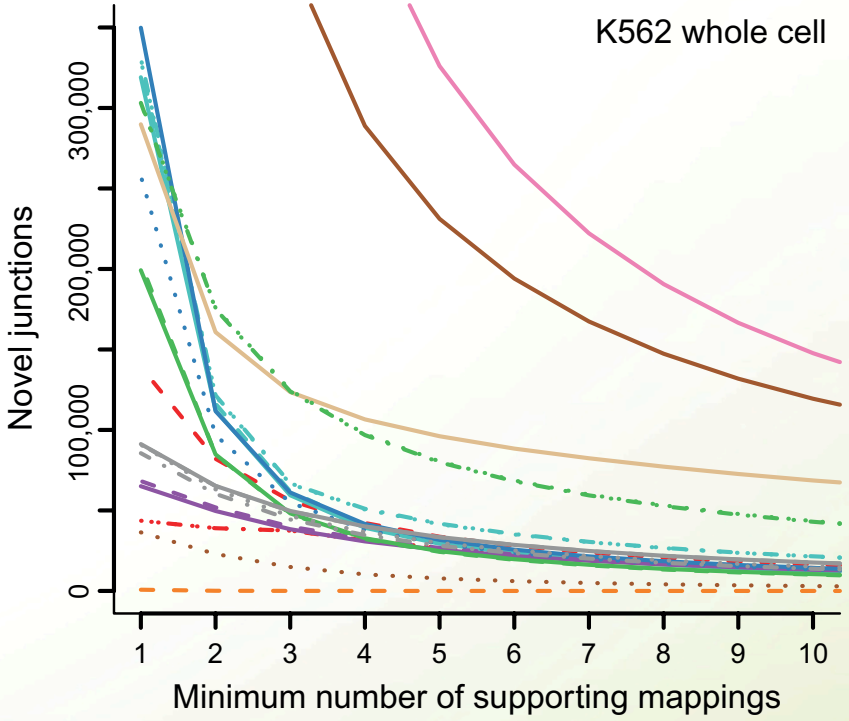
- GSNAP and GSTRUCT exhibit high sensitivity for both long and short deletions
- TopHat2 ann is most sensitive for long insertions

Engström et al. *Nature Methods* 2013

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

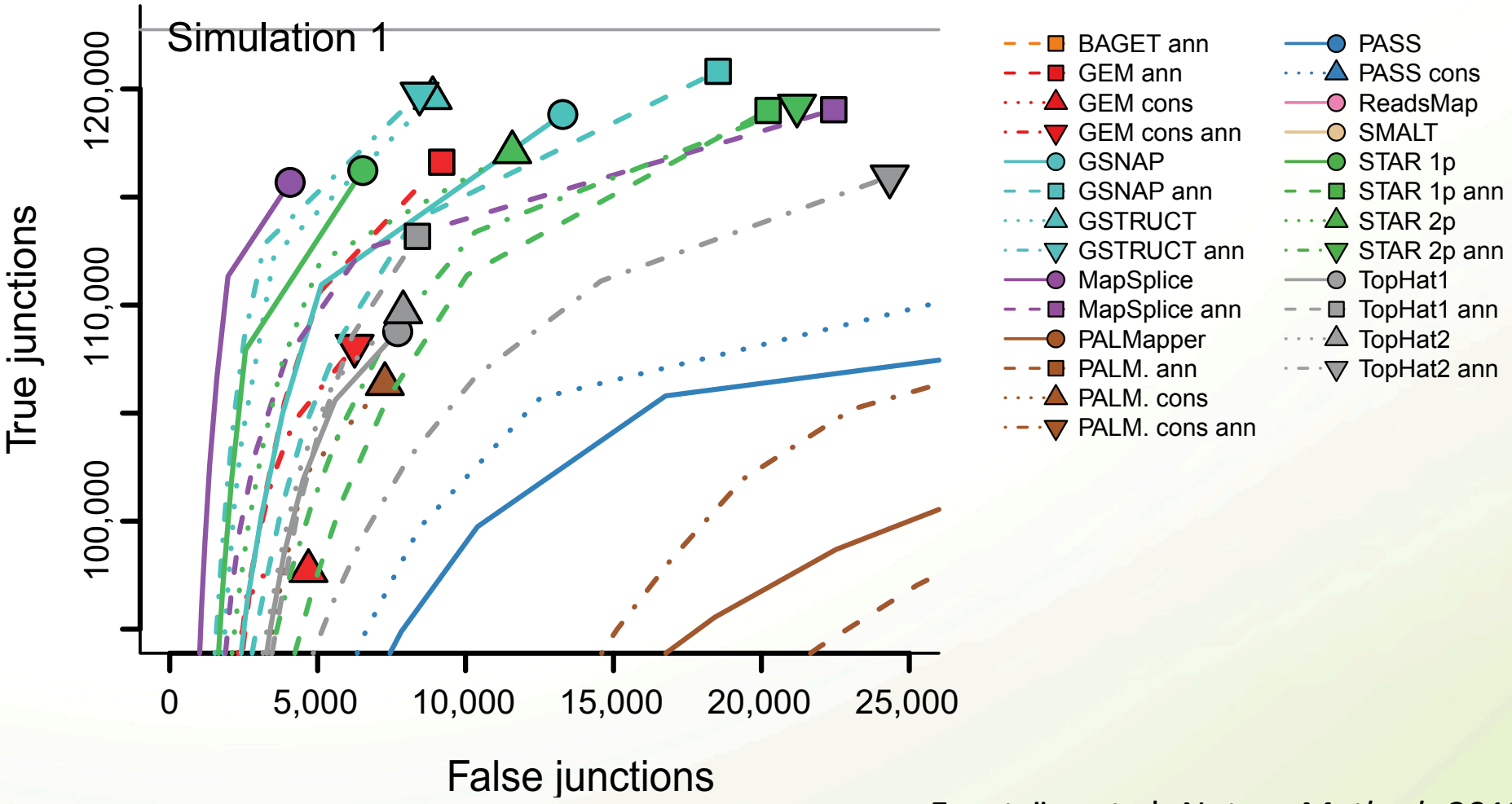
Novel junctions are typically supported by few alignments





Engström et al. *Nature Methods* 2013

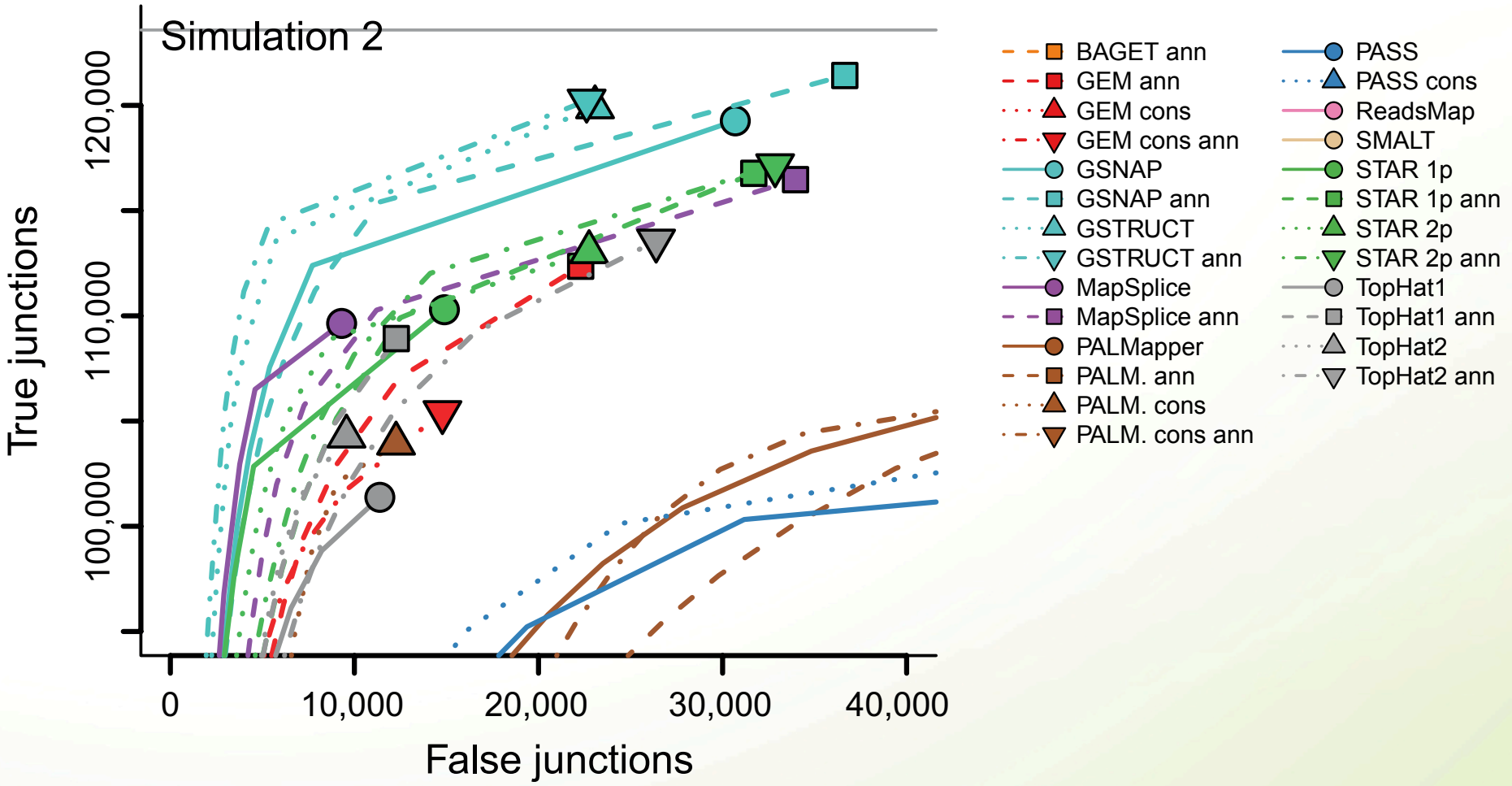
Improved junction accuracy by filtering on coverage



Engström et al. *Nature Methods* 2013

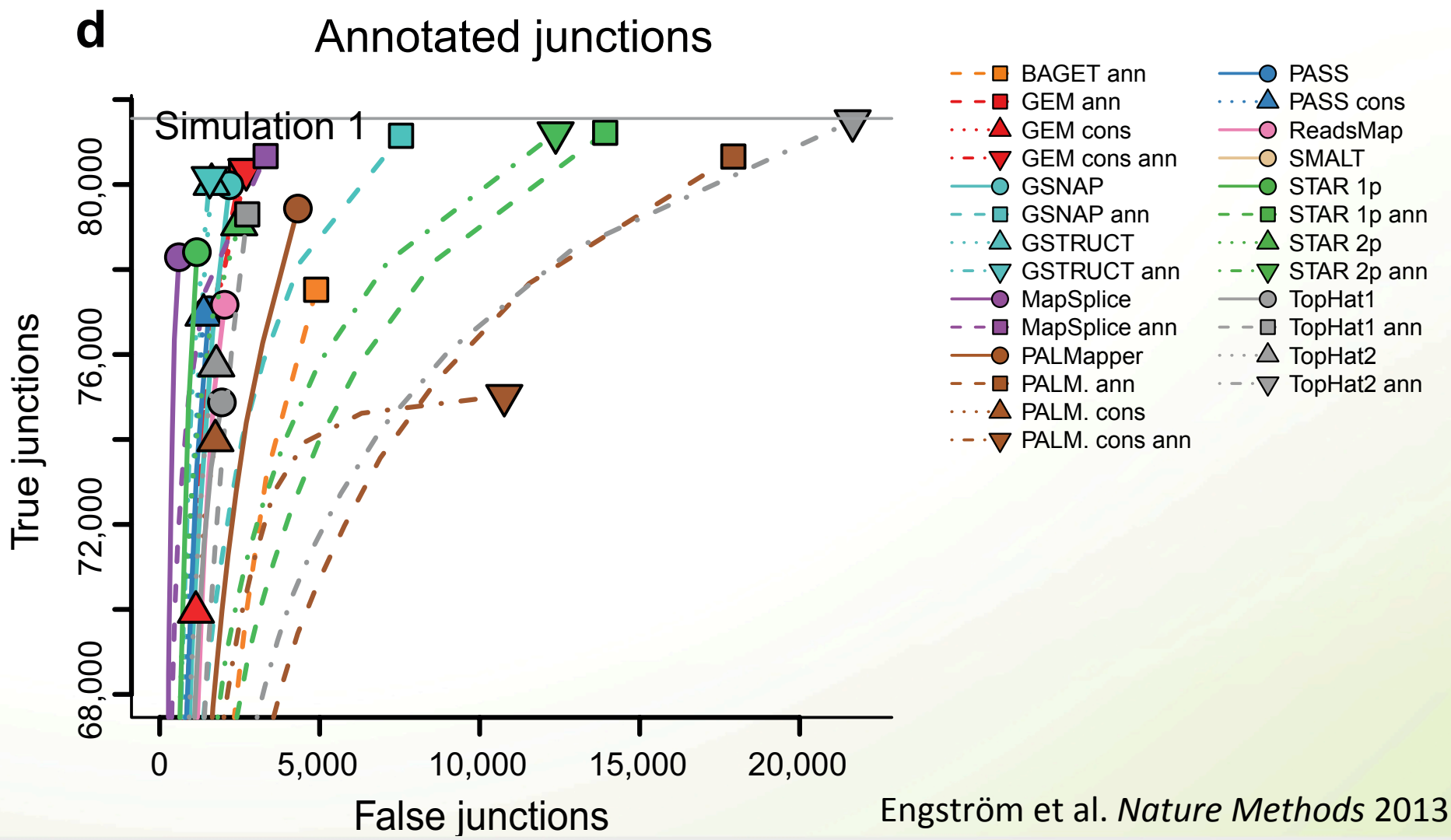
Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Improved junction accuracy by filtering on coverage



Engström et al. *Nature Methods* 2013

Several methods show over-confidence in annotation



Engström et al. *Nature Methods* 2013

Top performers (RGASP)

In general, GSNAP, GSTRUCT, MapSplice and STAR compared favorably to the other methods, but also displayed certain weaknesses:

- MapSplice is a conservative aligner, both with respect to mismatch frequency, indel calls and exon junction calls.

- The largest issue with GSNAP, GSTRUCT and STAR is the presence of many false exon junctions in the output.

Engström et al. *Nature Methods* 2013

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

Recommendations

- Use a two-pass workflow
- STAR and GSNAP generally perform well
- HISAT also seems to do well
- HISAT and STAR are the fastest
- HISAT2 has not been evaluated but the authors recommend it over HISAT
- HISAT2 and GSNAP can use SNP data, which may give higher sensitivity
- If you want to run Cufflinks, use TopHat2 or HISAT2
- For long (PacBio) reads, STAR, BLAT or GMAP can be used

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
```

Thanks for listening!