Mapping RNA-seq reads

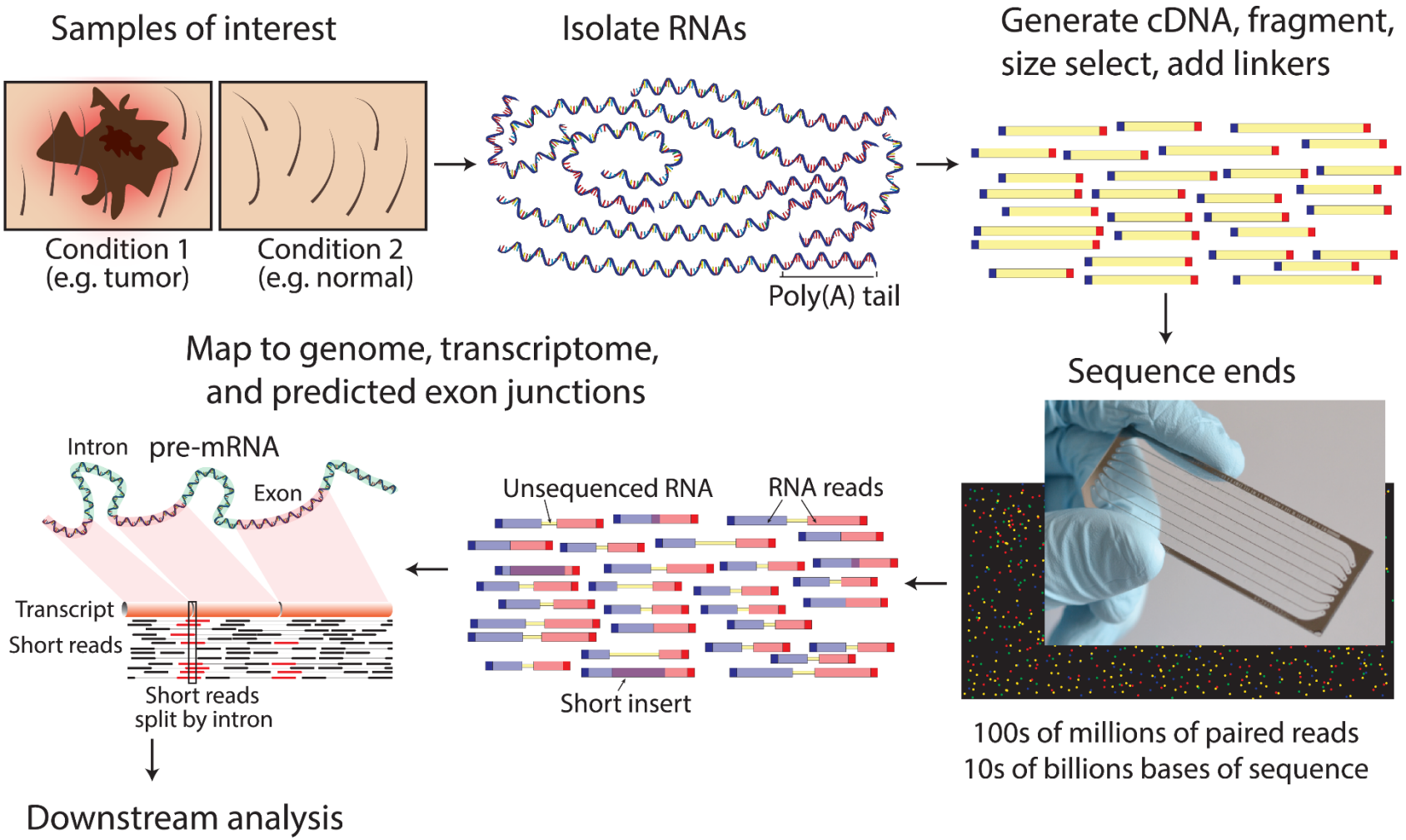
Alexander Dobin
CSHL

Outline

- Introduction: RNA-seq technology, analyses, pipelines
- Mapping of RNA-seq reads to the genome
- Post-mapping analyses
- Processing Single-Cell RNA-seq

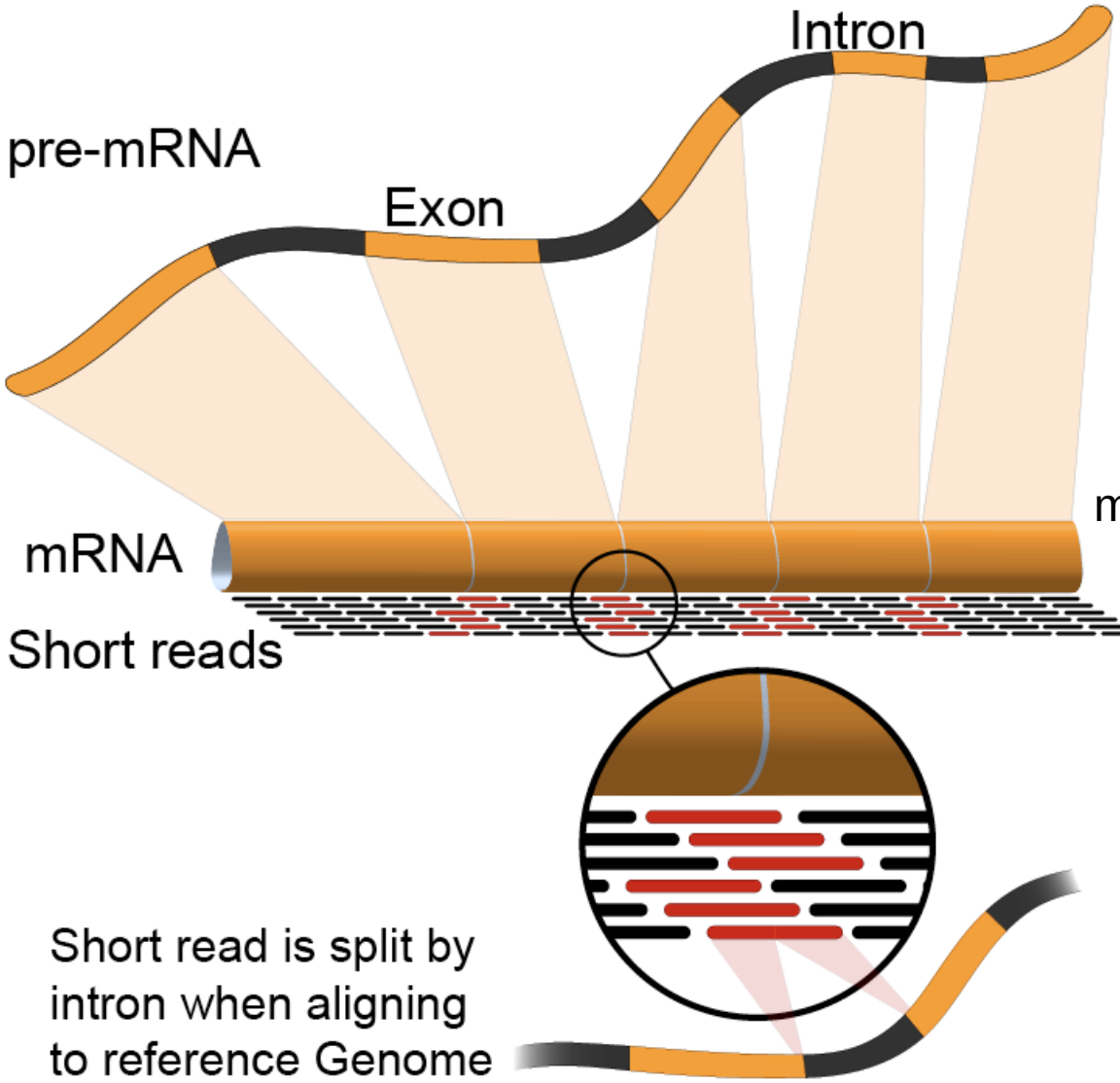
Introduction: RNA-seq technology, analyses, pipelines

RNA-seq



Griffith M, et al. PLOS Computational Biology 11(8): e1004393. https://doi.org/10.1371/journal.pcbi.1004393

RNA-seq



mRNA:0.3-10kb, median=1.4kb

short reads: 50-300b

https://en.wikipedia.org/wiki/RNA-Seq#/media/File:RNA-Seq-alignment.png

RNA-seq analysis challenges

Same as for DNA resequencing

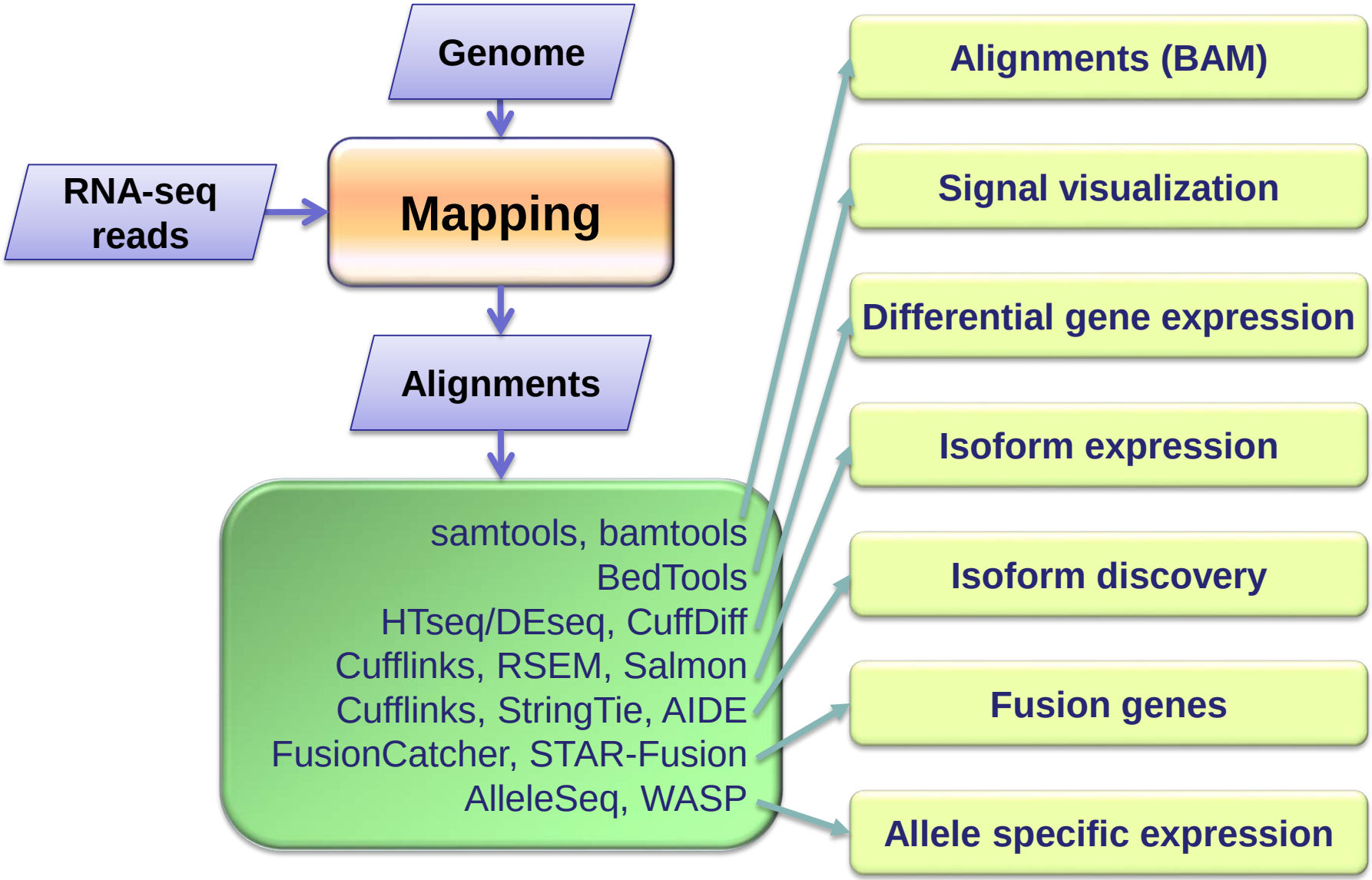
- Sequencing errors
- Genomic variations: single nucleotide, insertions/deletions, structural

Specific for RNA-seq

- Multitude of tasks and objectives in transcriptome analysis
- Mapping spliced reads to the genome

- Multi-mappers are important (expression of repeats, paralogs, etc)
- Highly expressed loci create mapping artifacts
- Allele-specific epxression
- Post-transcriptional RNA editing

Multitude of tasks and pipelines



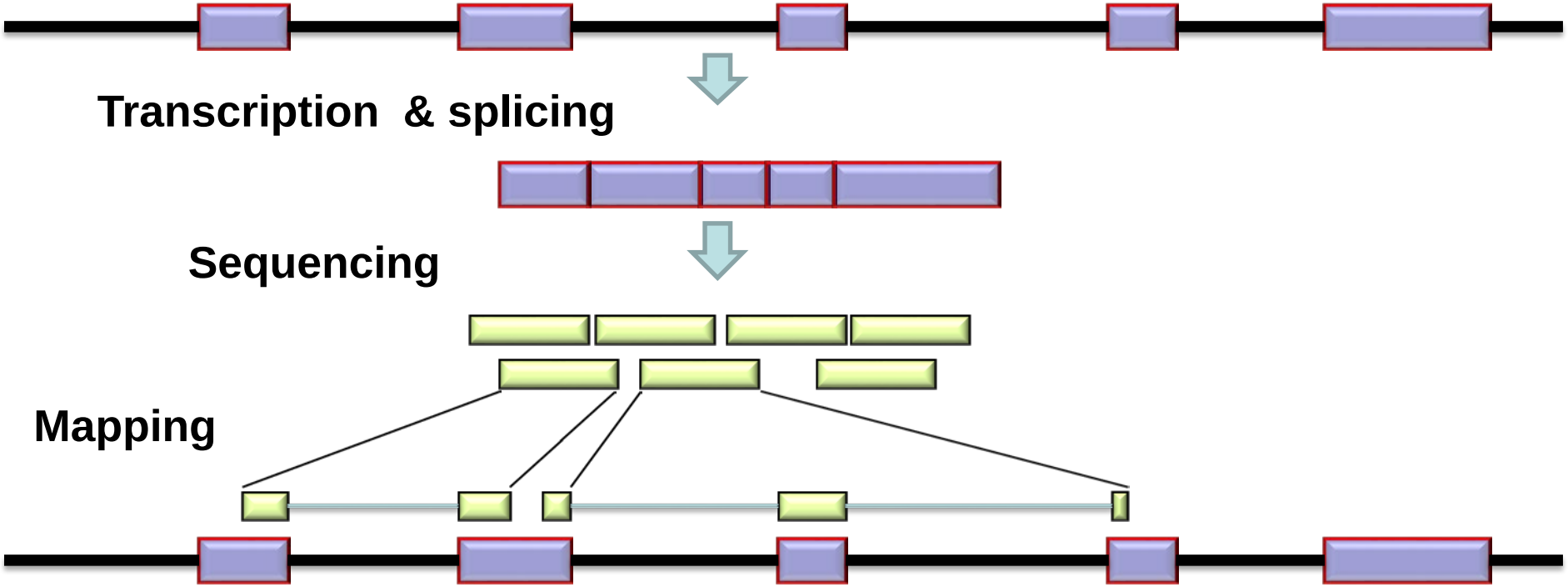
Mapping RNA-seq reads to the genome

Short reads aligners

DNA	
BWA	
Bowtie(2)	

RNA	
TopHat(2)	Slow
STAR	Fast, many features
HISAT	Fast, low RAM
GSNAP	Slow, accurate
Salmon	Pseudo-Alignment to Transcriptome Very fast, low RAM
Kallisto	

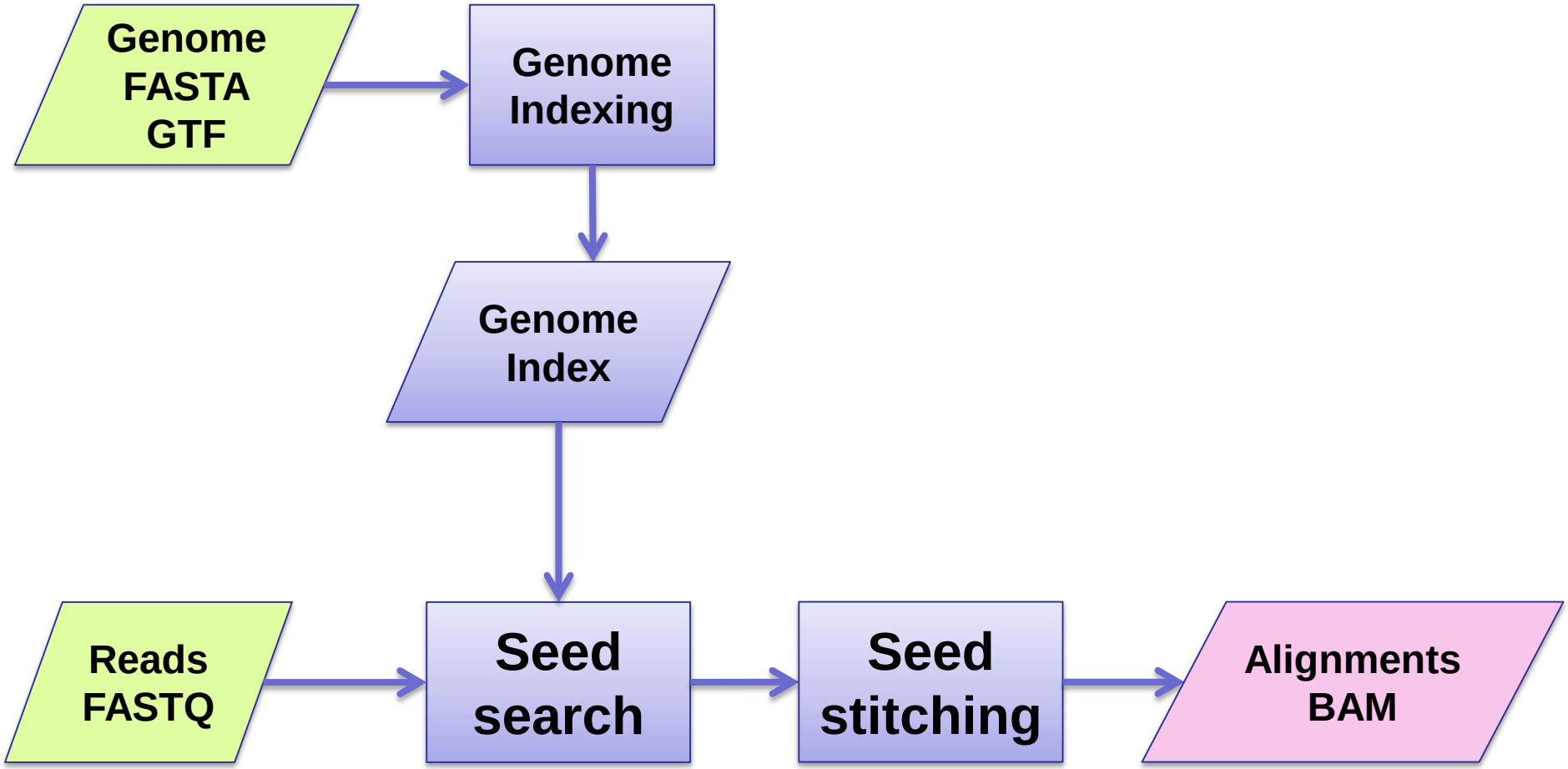
Mapping spliced reads to the genome



- **Most long RNAs are spliced**

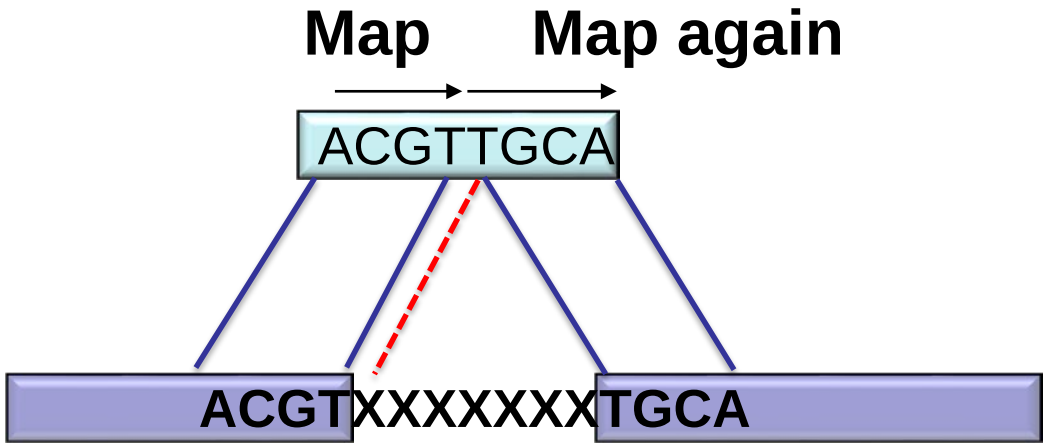
- Short reads map non-contiguously, may contain >1 splice junction
- Large introns: ~0.1-1,000 kb in mammals

STAR mapping workflow



Seed search

- Consecutive maximal mappable prefix search:



- Natural way to find splice junction

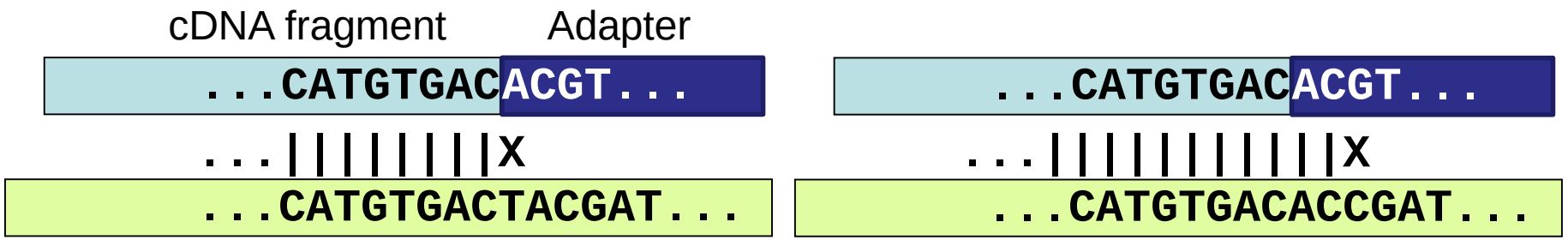
  also finds mismatches and unmappable tails

- Fast binary search in suffix arrays

- Uncompressed suffix arrays:

  faster than compressed, but require more RAM

Adapter trimming

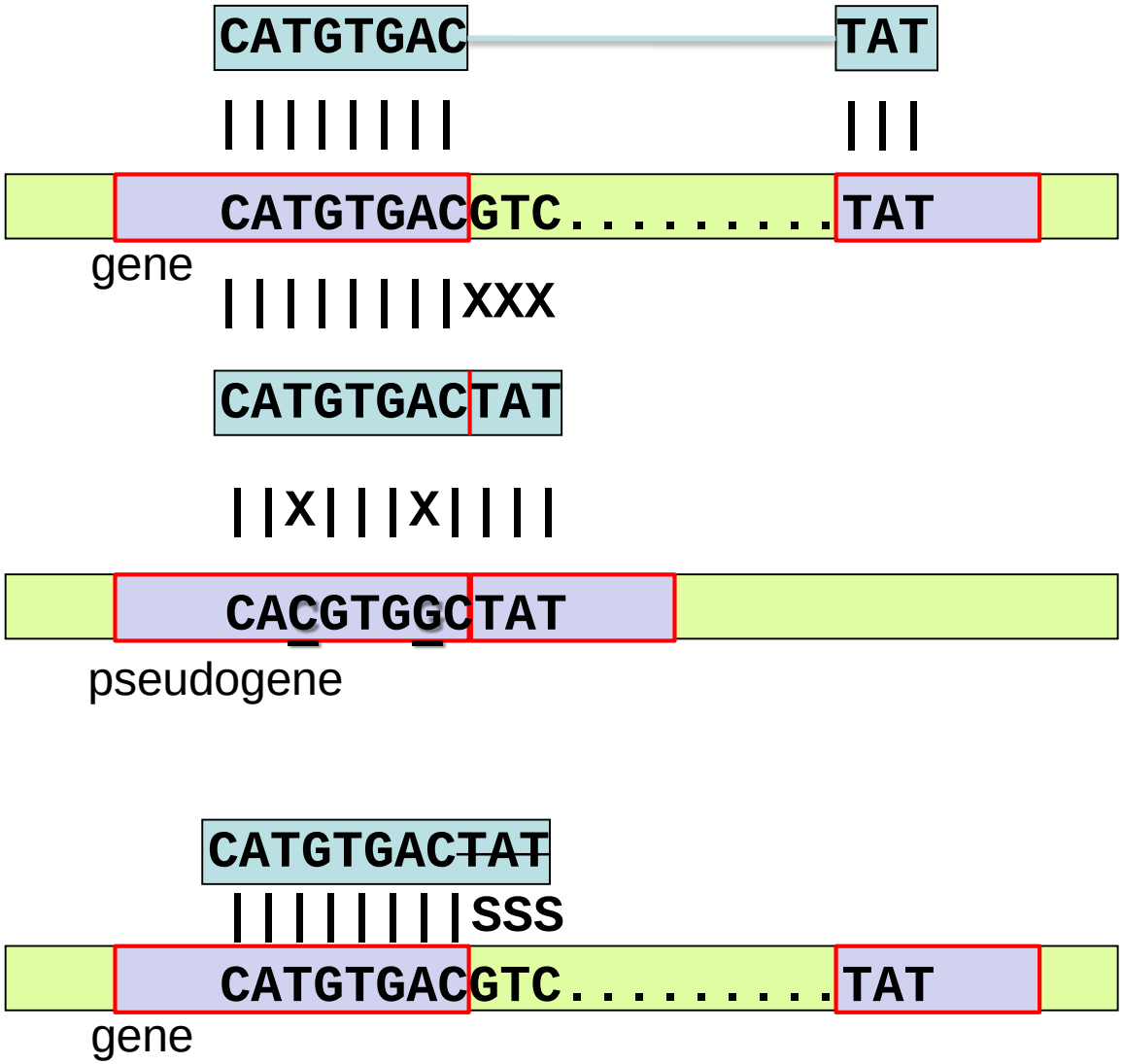


Locus 1:
only the cDNA sequence
maps to the genome

Locus 2:
cDNA sequence + 2 adapter bases
map to the genome

- Adapter at 3' of the read sequence if
  fragment ("insert") length < sequence length
- Untrimmed adapter can turn multi-mappers into unique mappers
- Trimming software: Cutadapt, Trimmomatic, FASTX, etc.
  take care not to mess up the read order for paired-end reads
- Basic aggressive 3' adapter trimming in STAR with
  `--clip3pAdapterSeq <sequence>  --clip3pAdapterMMp 0.1`

Soft-clipping vs. End-to-End alignment



True alignment: spliced

Hard to find because of short overhang: 3b will randomly occur every 64b in the genome

End-to-end alignment: 3MM

May be filtered out

End-to-End alignment to the pseudogene: 2MM

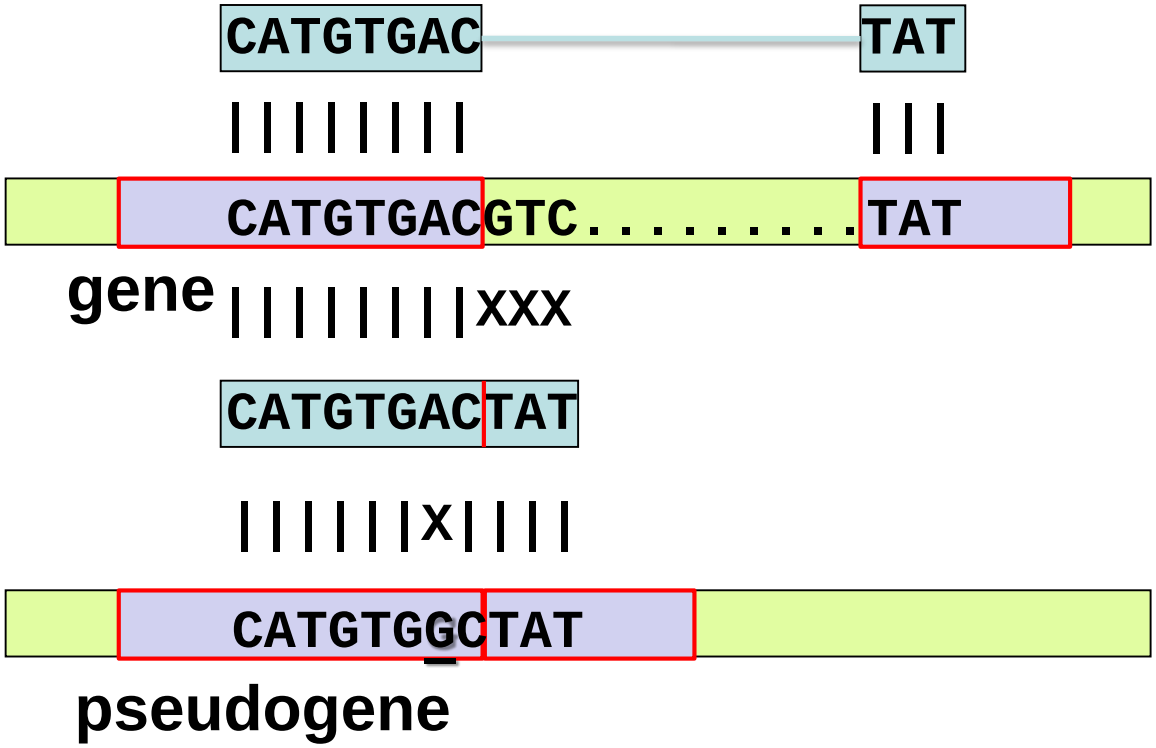
False alignment!

Soft-clip unmatched bases

Soft-clipping penalty < MM penalty

Also helps to catch A-tails, end modifications, adapters, poor quality tails

Mapping short splice overhangs



True alignment: spliced

Short splice overhang:

3b will randomly occur every 64b in the genome

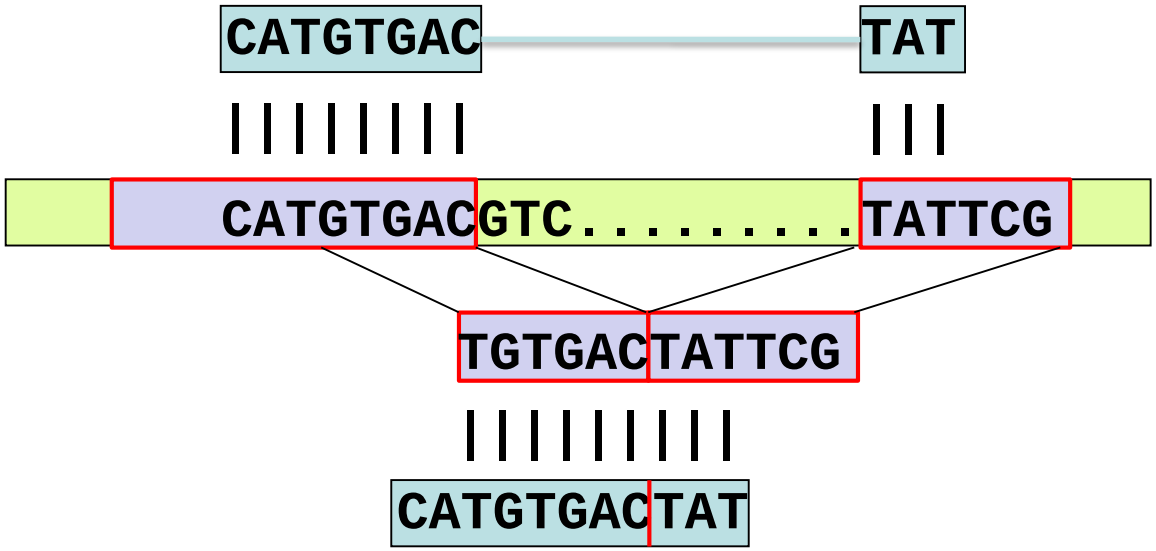
3 mismatches or 3 soft-clipped bases

Alignment to the "processed" pseudogene:

1 mismatch

False alignment!

Using annotations



Splice overhangs shorter than ~5-10 bases are hard to position

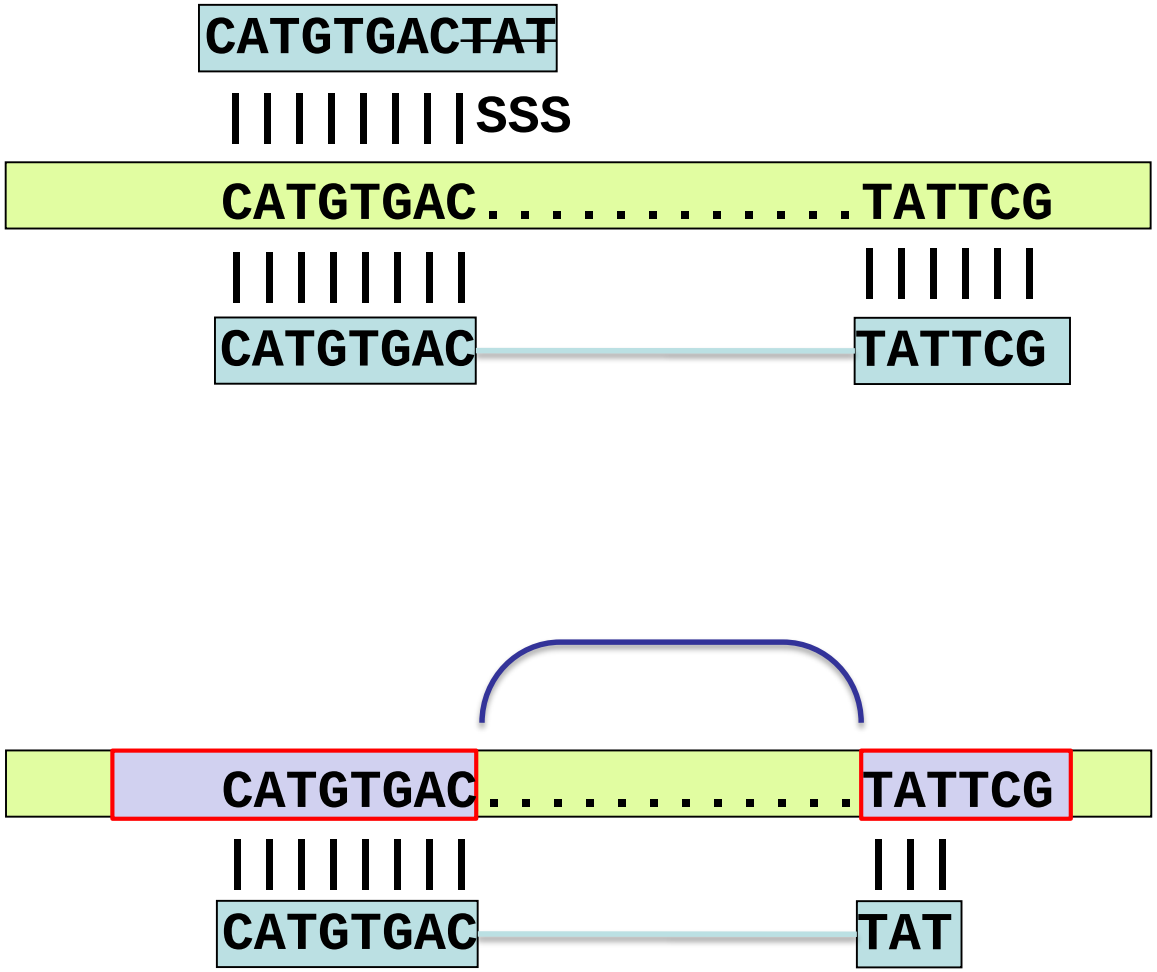
Solution:

use prior information:

“annotated” junctions

	BLAT	TopHat	STAR	STAR +annotations
False alignments	2.9%	5.4%	2.0%	0.1%
Mapped to pseudogenes	0.7%	4.5%	1.8%	0.1%

2-pass mapping



1st pass:
Reads with short overhangs are soft-clipped

Read with long overhangs identify novel junctions

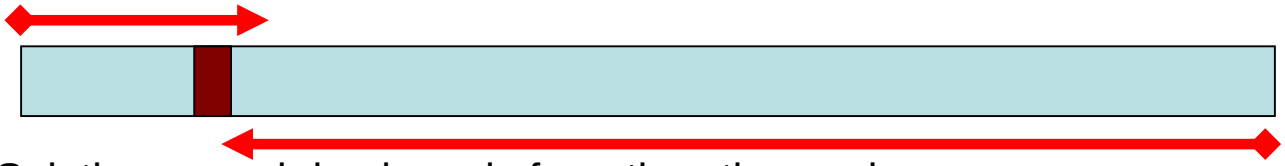
2nd pass:
Junctions from the 1st pass are added to the search space

Read with short overhangs map spliced to novel junctions

Increasing seed search sensitivity

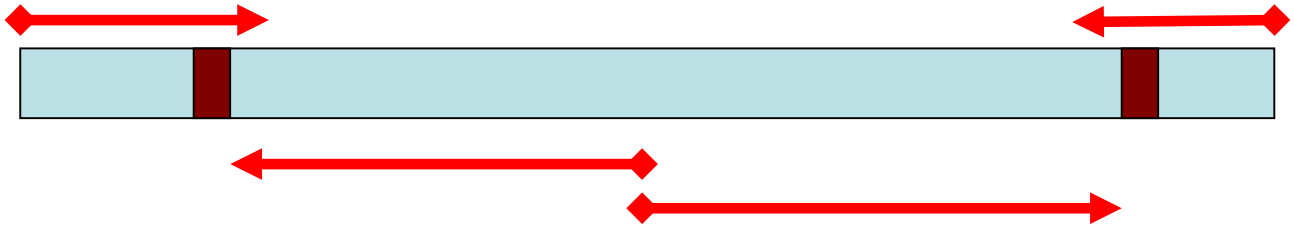
- One mismatch near one of the ends

  Seed is too short – max exact search does not stop at the mismatch and maps to a wrong locus



  Solution: search backwards from the other end

- Two mismatches near the ends



  Solution: start search from the middle of the read

- **`--seedSearchStartLmax <N>`**

  user defined parameter to start search as often as needed

  – Reducing N will increase sensitivity, but reduce mapping speed

  – Default N= 50b, works well for Illumina reads

Anchor seeds and windows

- **`--seedMultimapNmax <N>`**

  All seeds that map <N times are recorded: =10,000 by default

  10-mers map 10,000 on average in human genome

- **`--winAnchorMultimapNmax  <N>`**

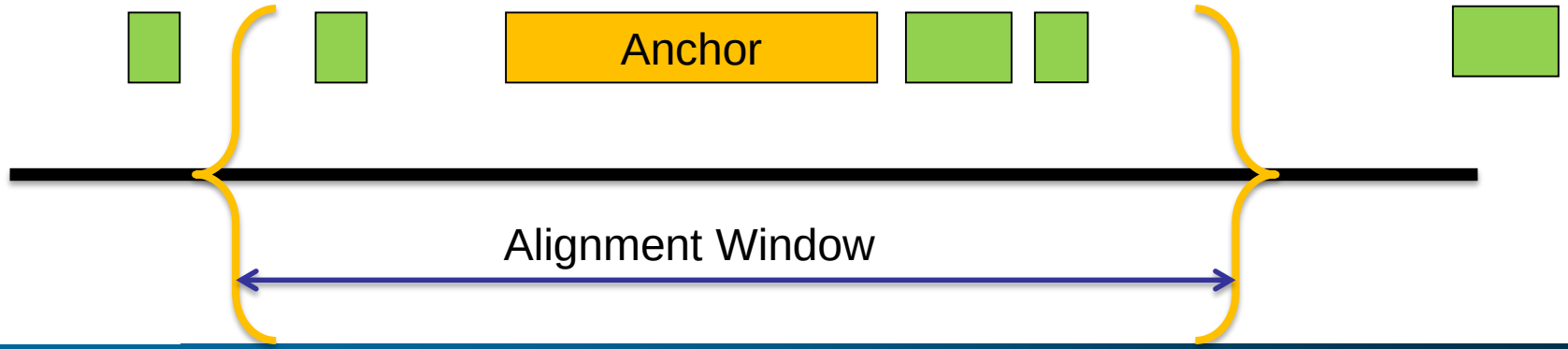
  “Anchors”: seeds that map <N times: =50 by default

- “Alignment windows”: genome regions around anchors

  All seeds inside alignment windows are stitched together

  Size of the window ~ maximum intron size, ~1Mb for human

  `--alignIntronMax <N>, --alignMatesGapMax <N>`



Understanding mapping results

STAR's Log.final.out file:

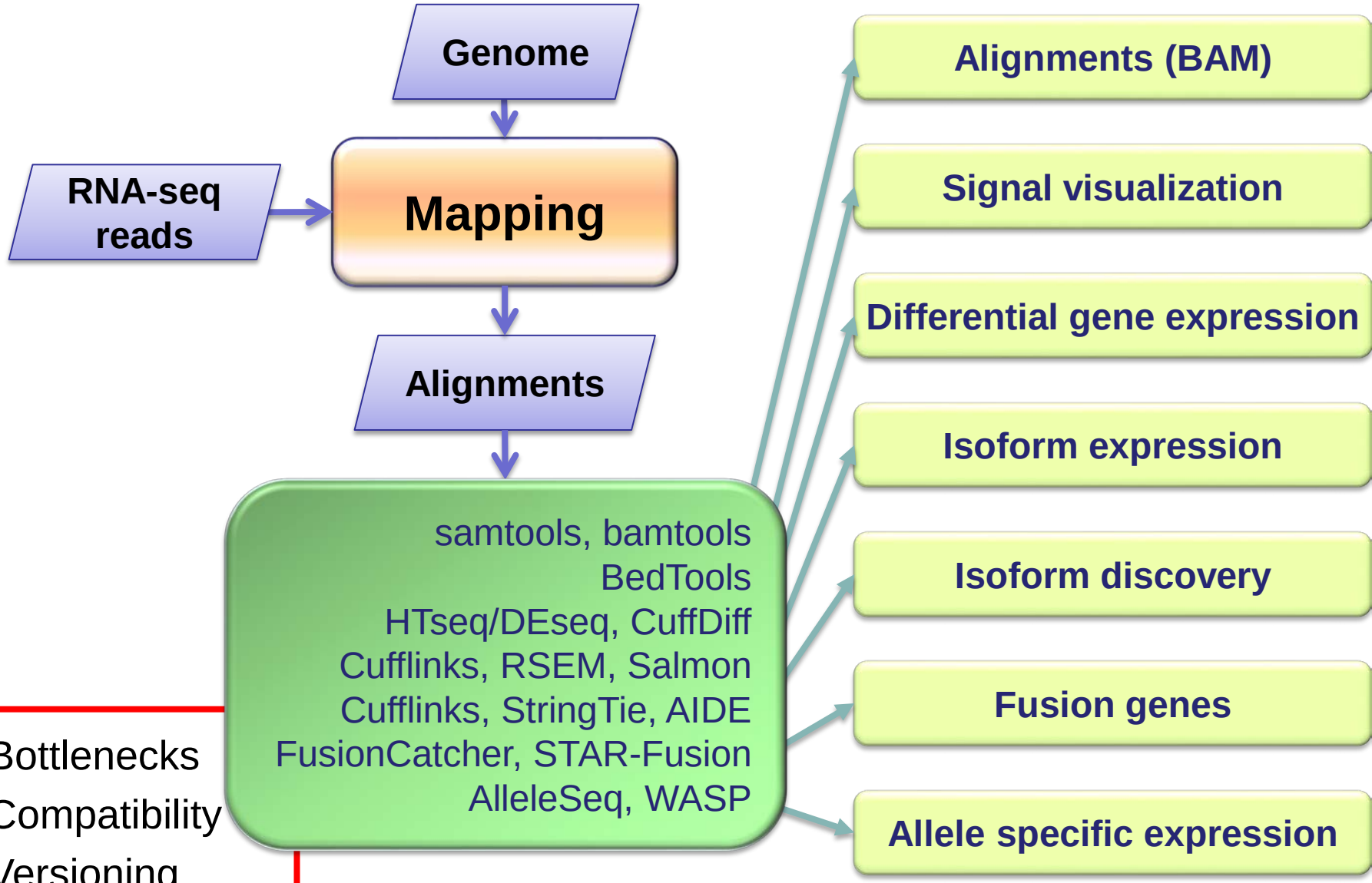
```
                          Average input read length |       202
                                       UNIQUE READS:
                         Uniquely mapped reads % |       90.08%
                           Average mapped length |       201.98
                        Mismatch rate per base, % |       0.30%
                           Deletion rate per base |       0.02%
                          Insertion rate per base |       0.01%
                                       MULTI-MAPPING READS:
              % of reads mapped to multiple loci |       3.55%
              % of reads mapped to too many loci |       0.02%
                                       UNMAPPED READS:
        % of reads unmapped: too many mismatches |       2.82%
                  % of reads unmapped: too short |       3.44%
                      % of reads unmapped: other |       0.08%
```

Why my mapping rate is low?

Possible problem	Checks/Solutions
File formatting mix-up:	Ensure the same order of read1/2
read1/read2 order broken	Map read1/2 separately
Poor quality of sequencing	Plot quality scores vs read length
Tails	
Poor quality	Trim by quality
Adapter - short insert	Trim adapter
rRNA insufficient depletion	Include rRNA sequences in the reference
Contamination with other species	BLAST unmapped reads

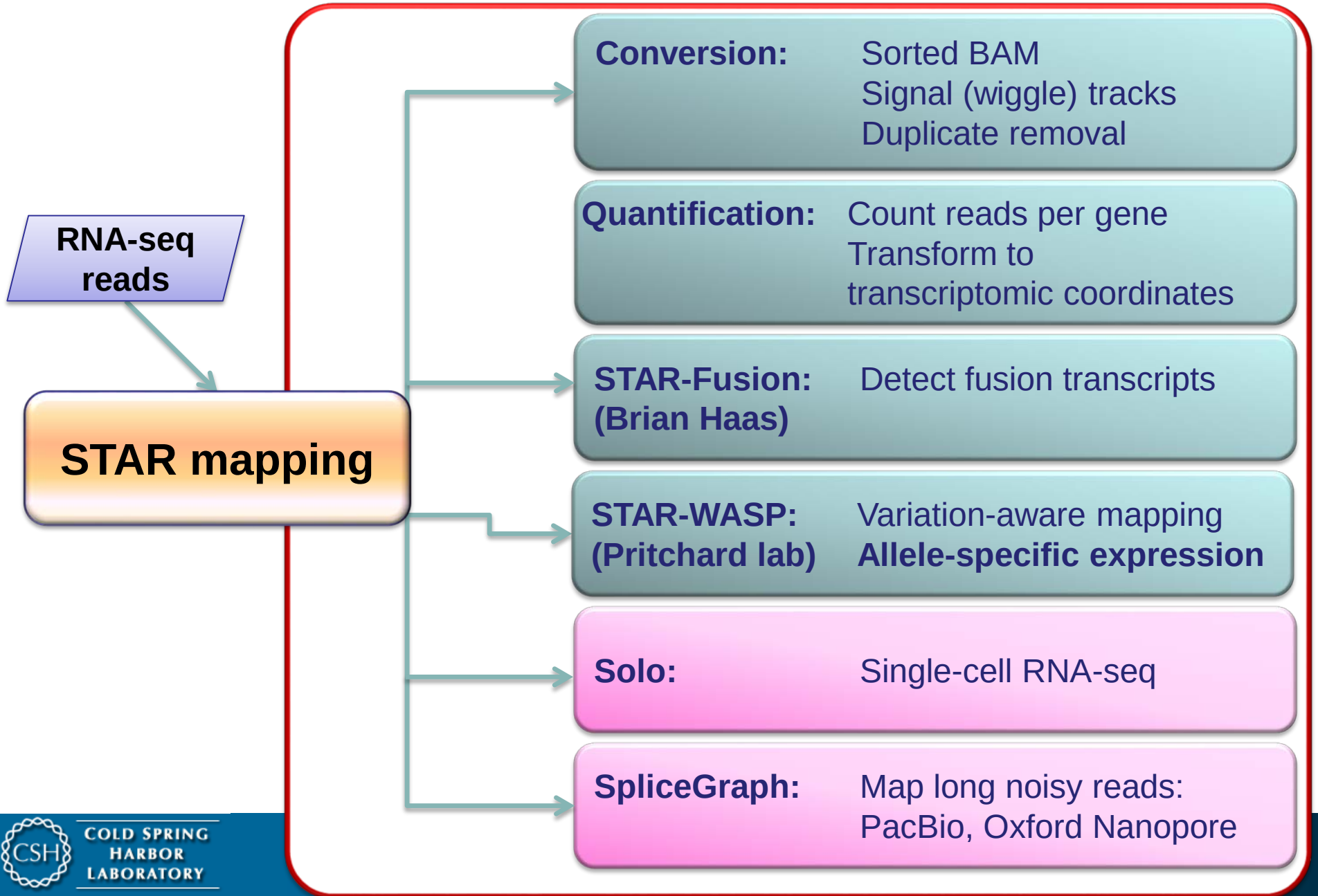
Post-mapping analyses

Multitude of tasks and pipelines



- Bottlenecks
- Compatibility
- Versioning
- Reproducibility

STARtools


RNA-seq reads
STAR mapping
Conversion:
Sorted BAM
Signal (wiggle) tracks
Duplicate removal
Quantification:
Count reads per gene
Transform to transcriptomic coordinates
STAR-Fusion: (Brian Haas)
Detect fusion transcripts
STAR-WASP: (Pritchard lab)
Variation-aware mapping
Allele-specific expression
Solo:
Single-cell RNA-seq
SpliceGraph:
Map long noisy reads: PacBio, Oxford Nanopore

Quantification tools

- Count how many reads overlap any of the isoforms of each gene
- Used in differential gene expression analysis
- HTseq: produces counts from BAM alignments
- 109M reads, 2x101b:

  STAR map: 22 min

  HTseq count: 250 min

  STAR map+count: 22 min

  `--quantMode GeneCounts`

- eXpress, RSEM …
- maximum likelihood estimation of isoform expression
- need alignments in “transcriptomic” coordinates
- mapping to transcriptome with BWA, Bowtie…
- STAR maps to the genome,

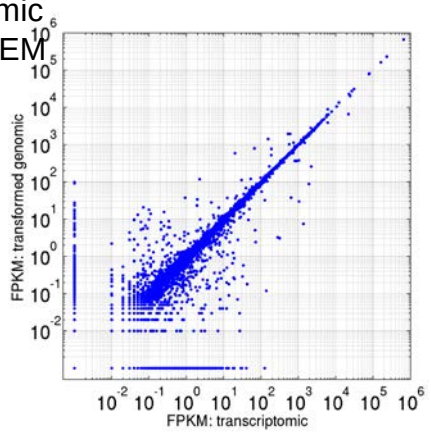
  and at the same time

  converts genomic alignments to transcriptomic

  no extra computational time required

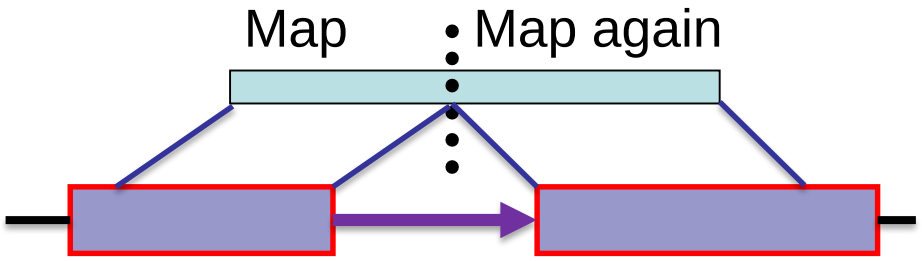
  `--quantMode TranscriptomeSAM`

Expression genomic => transcriptomic STAR / RSEM

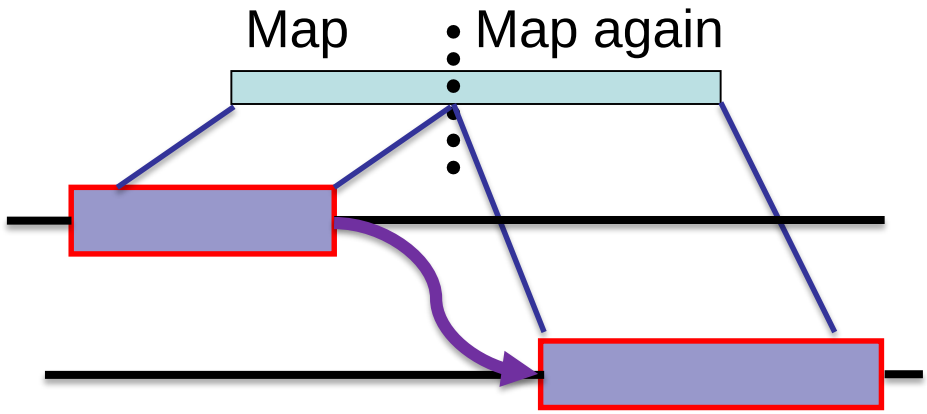


Expression using transcriptomic alignments Bowtie / RSEM

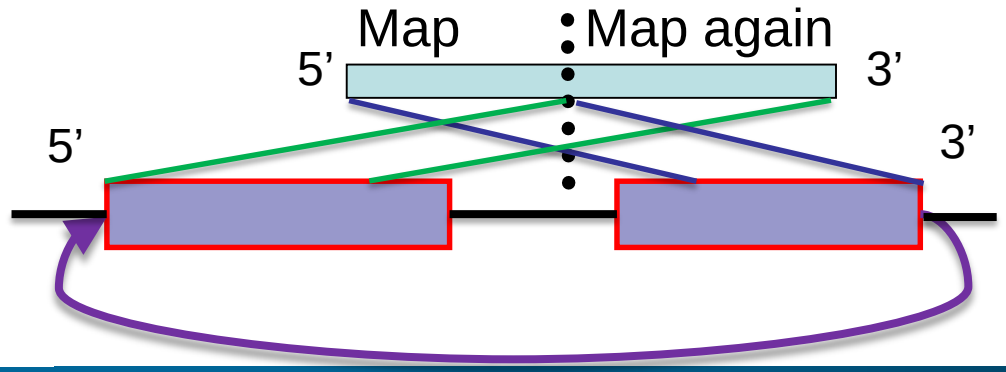
Chimeric and circular junctions



Linear junction

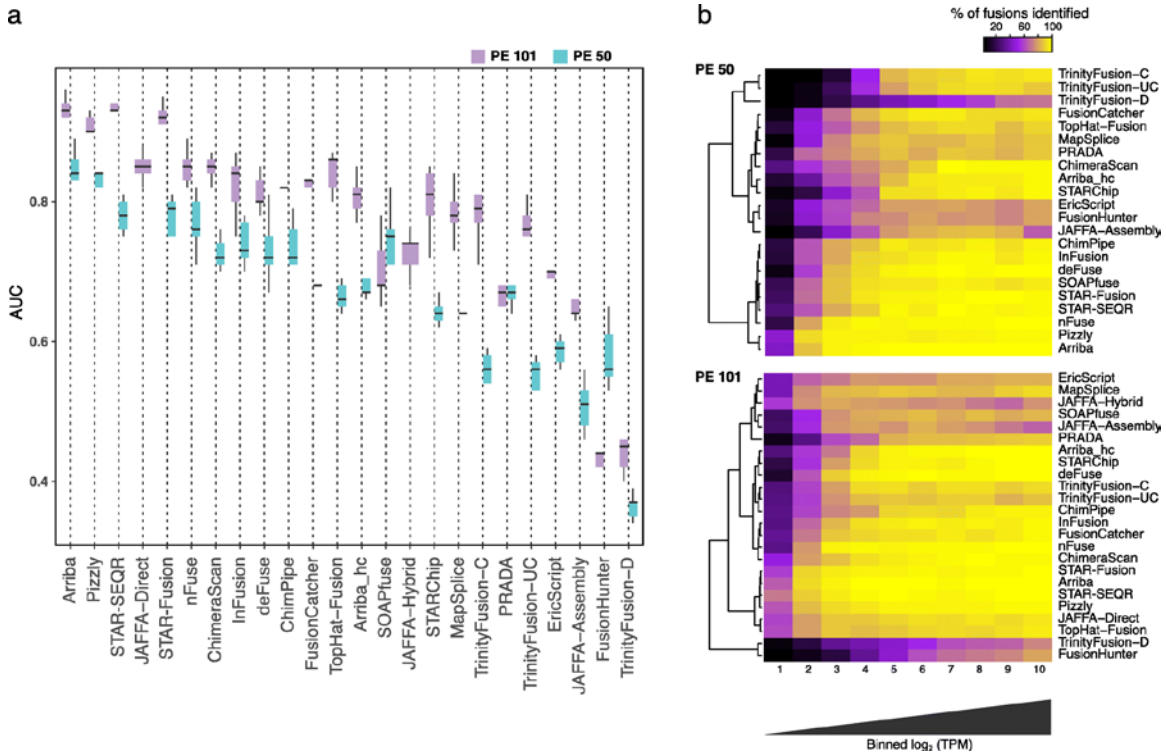


Chimeric junction



Circular junction

STAR-Fusion



STAR-Fusion
FusionInspector

developed by **Brian Haas** (Broad Institute)

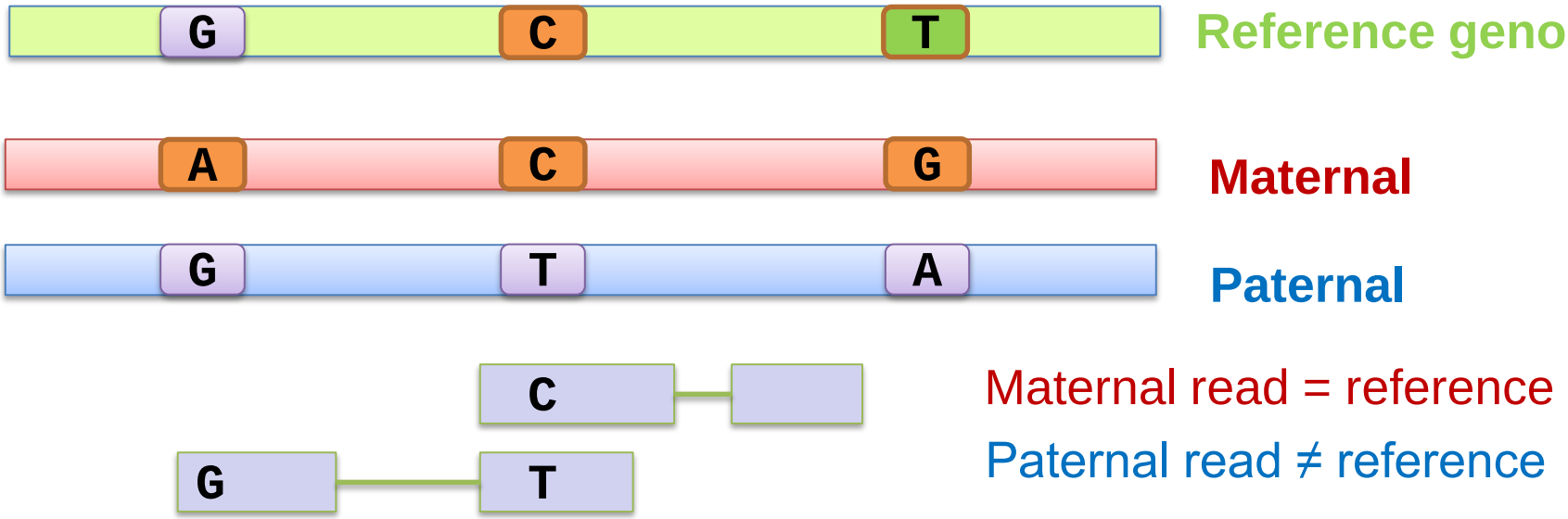
Analyzes STAR chimeric alignments to detect fusion transcripts in RNA-seq data

https://github.com/STAR-Fusion/STAR-Fusion

Haas, Dobin, Li, Stransky, Pochet, Regev: Genome Biology, 2019

Personal genomics

- Precision Medicine: personal diploid genome for everyone
- Analyze RNA-seq data using personal genomes.



van de Geijn, McVicker, Gilad, Pritchard: **WASP: allele-specific software for robust molecular quantitative trait locus discovery,** Nat. Meth. 2015

Now implemented as **STAR-WASP**

Long reads

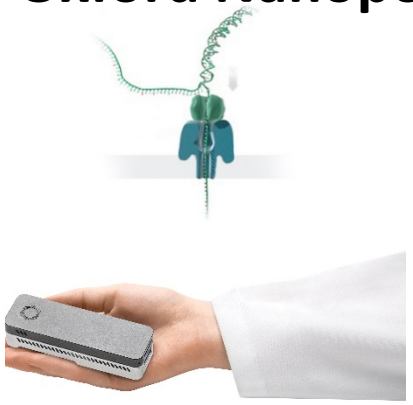
- 3rd generation sequencing technology
- Long reads: ~10 kilobases
- High error rate:

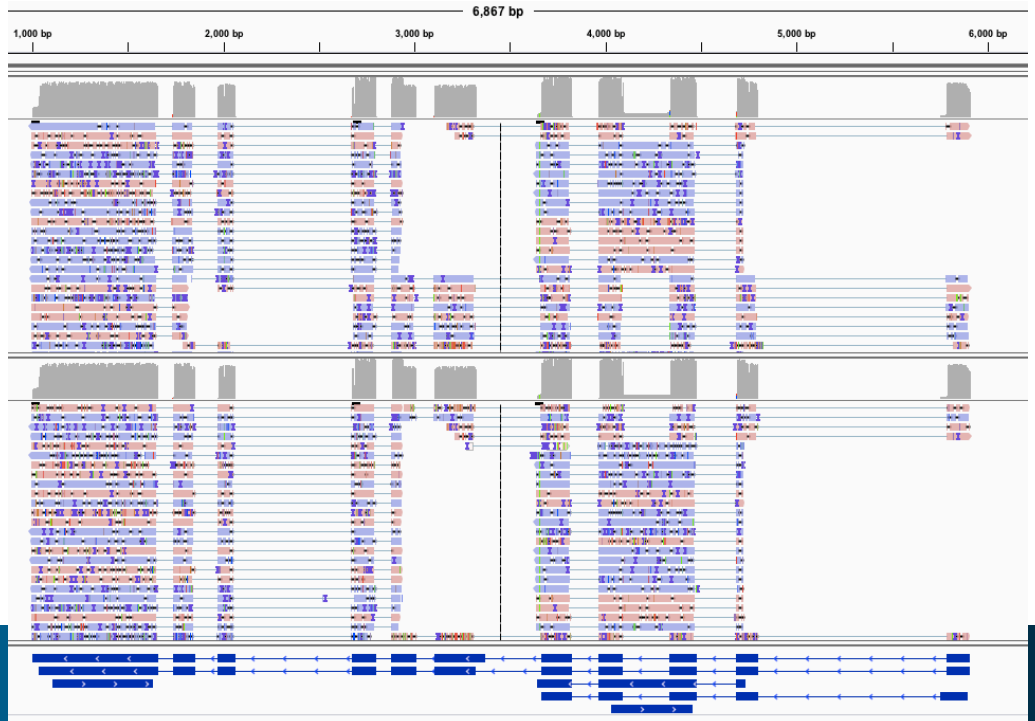
  10-15%

  (vs Illumina's <0.5%)
- Error every 6-10 bases

Pacific Biosciences

Oxford Nanopore

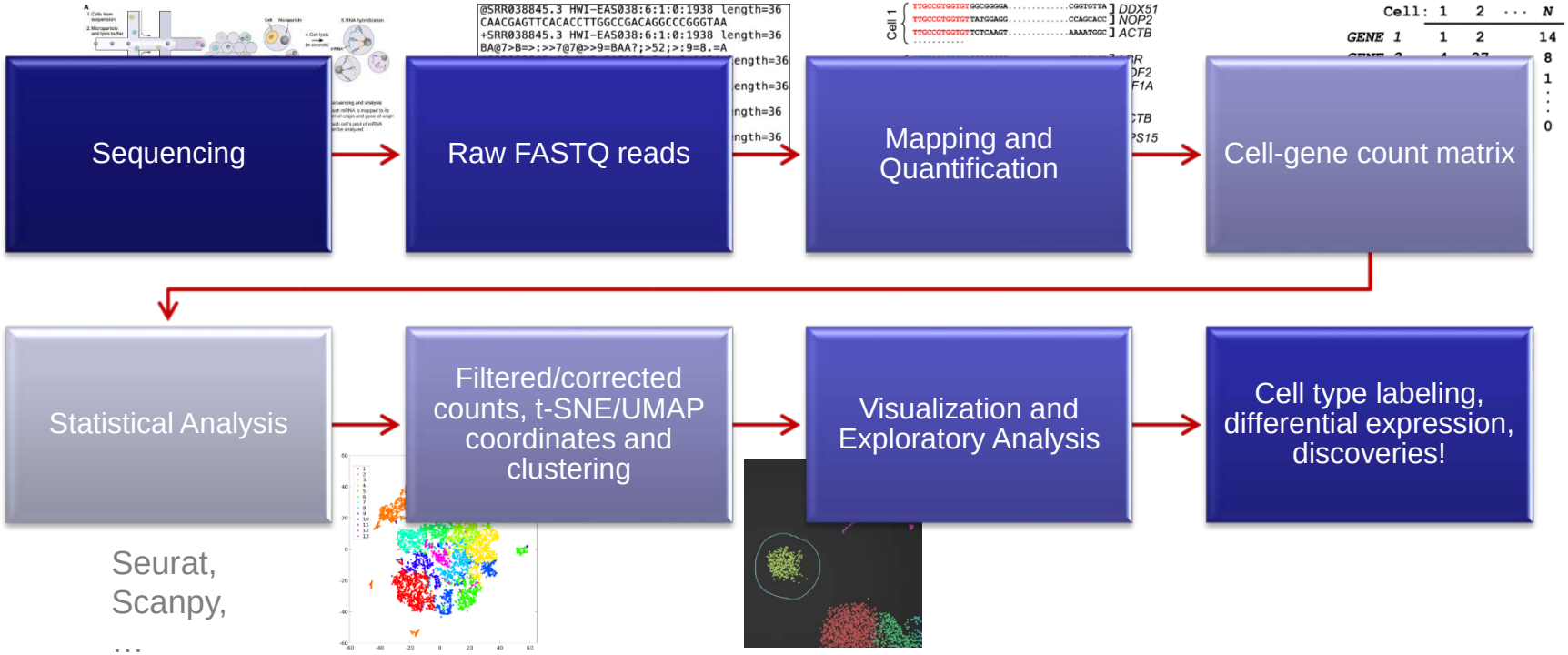




Single-cell RNA-seq

Single-cell RNA-seq processing

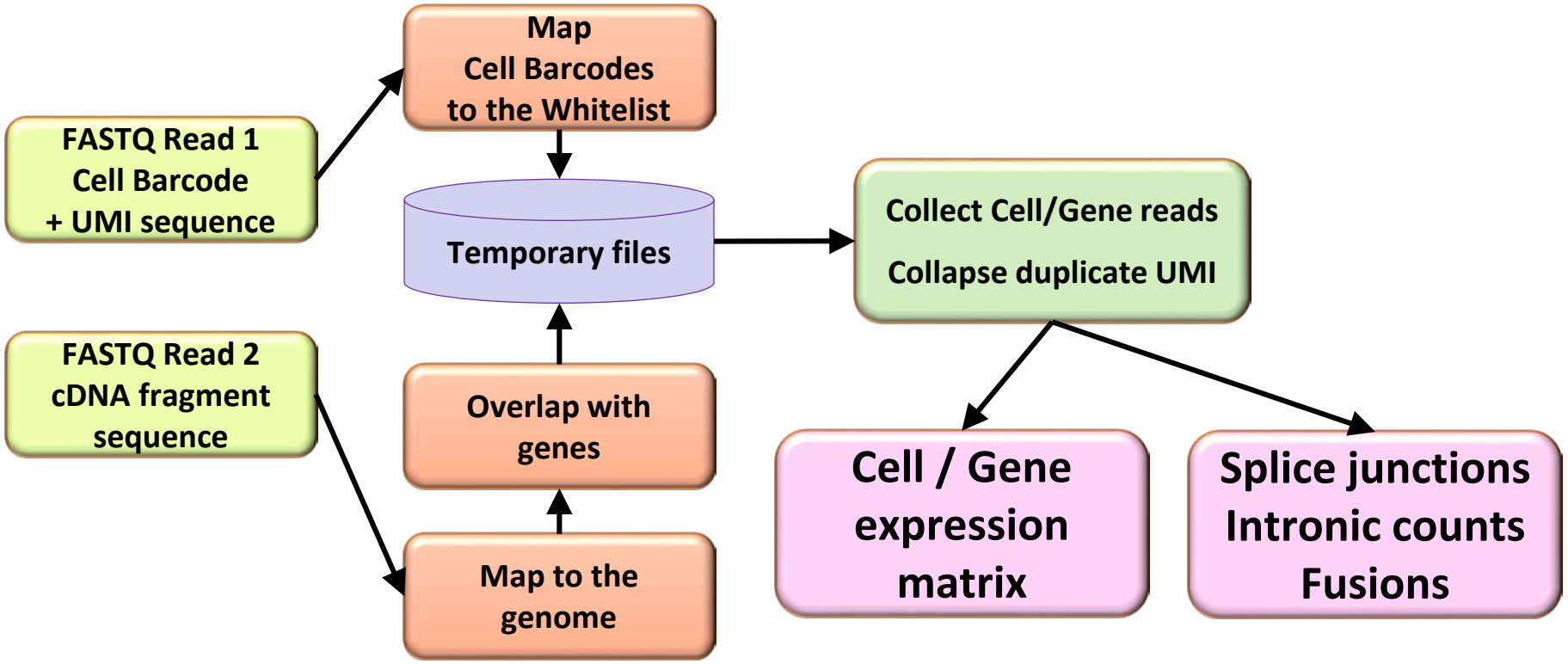
STARsolo



Motivation

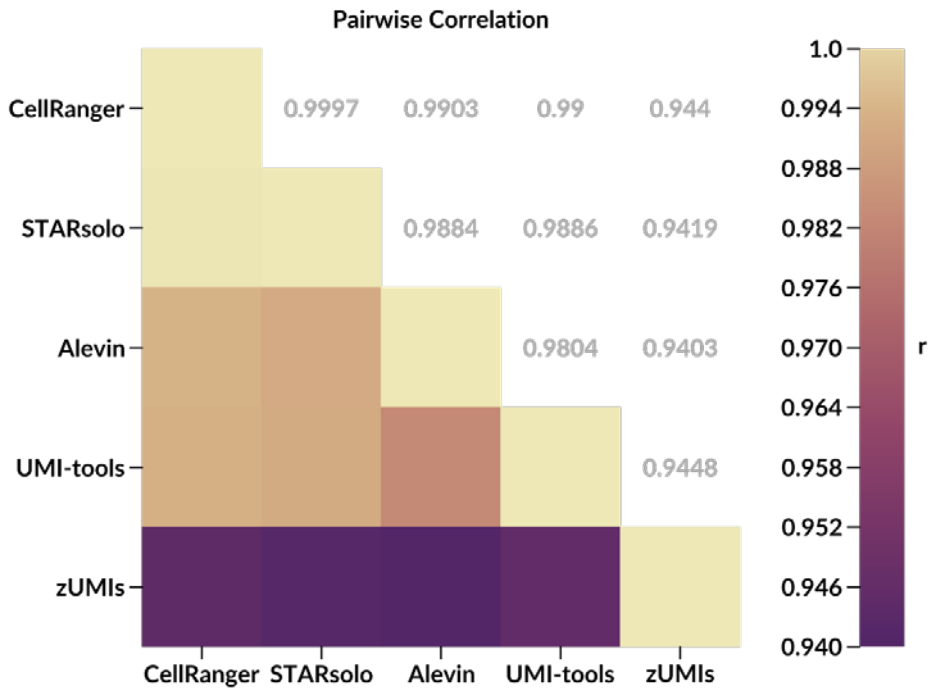
- 10X CellRanger uses STAR for mapping read to the genome
- CellRanger processing time is 10x of STAR mapping time
  i.e. CellRanger spends 90% of time on demultiplexing and UMI collapsing
  which are easier tasks than mapping to the genome
- Hard to change STAR (or other) parameters
- Not open source, not well documented

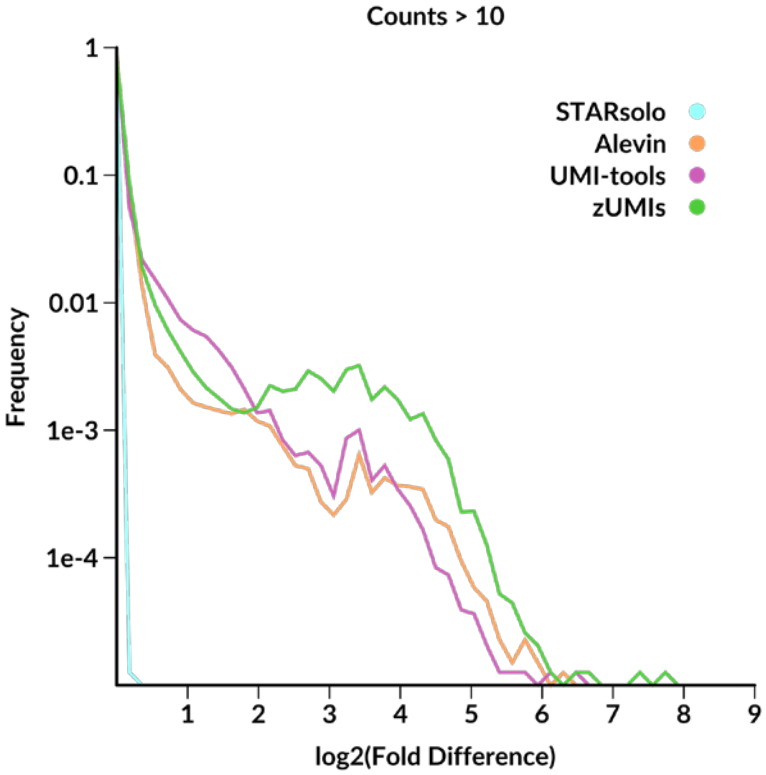
STARsolo is integrated into STAR



STARsolo follows CellRanger rules for assigining reads to genes, demultiplexing cell barcodes and deduplicating UMIs

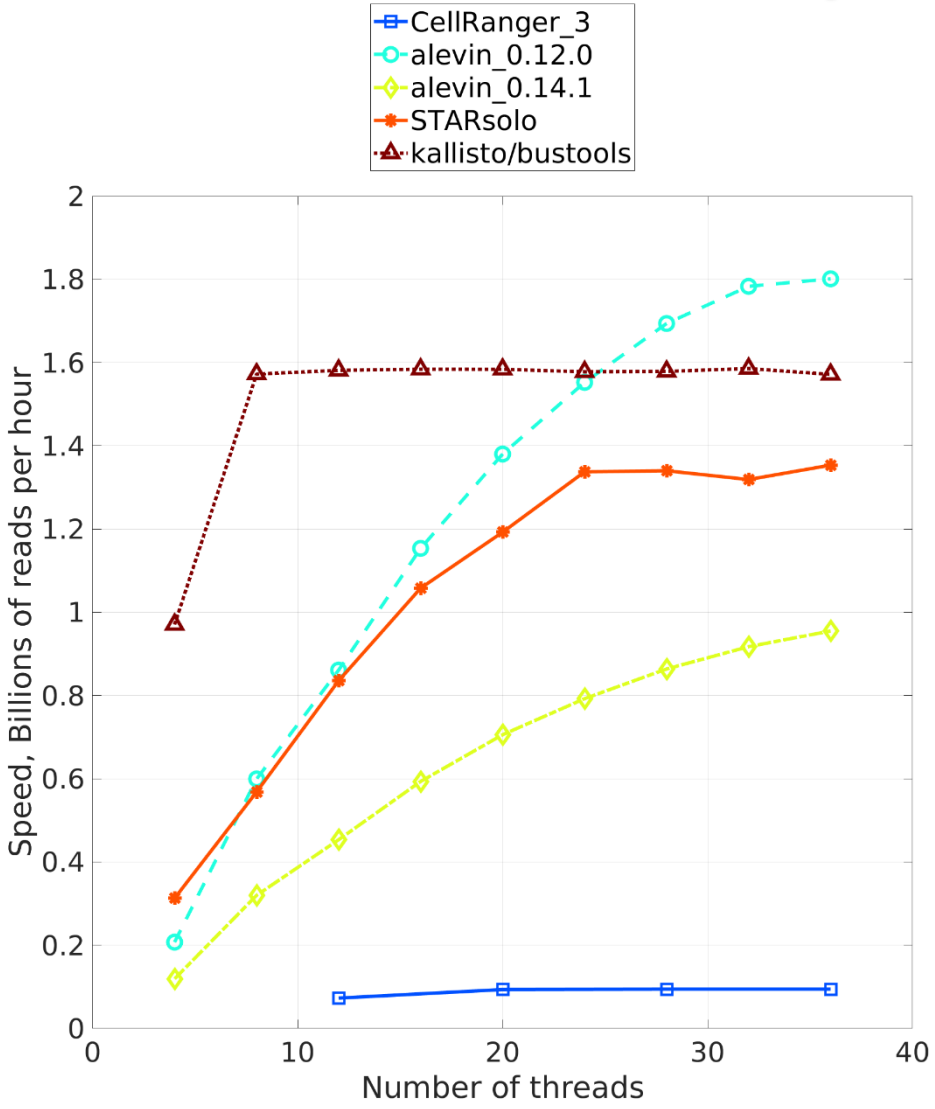
Gene/cell counts





- **STARsolo gene counts are nearly identical to CellRanger's**
- STARsolo gene count matrices can be used as drop-in replacement for CellRanger's gene count matrices
- Very small differences between STAR and CellRanger are caused by inconsistent UMI collapsing in CellRanger

Speed



- STARsolo: time spent on cell barcodes demultiplexing and UMI collapsing: only 10% of the mapping time

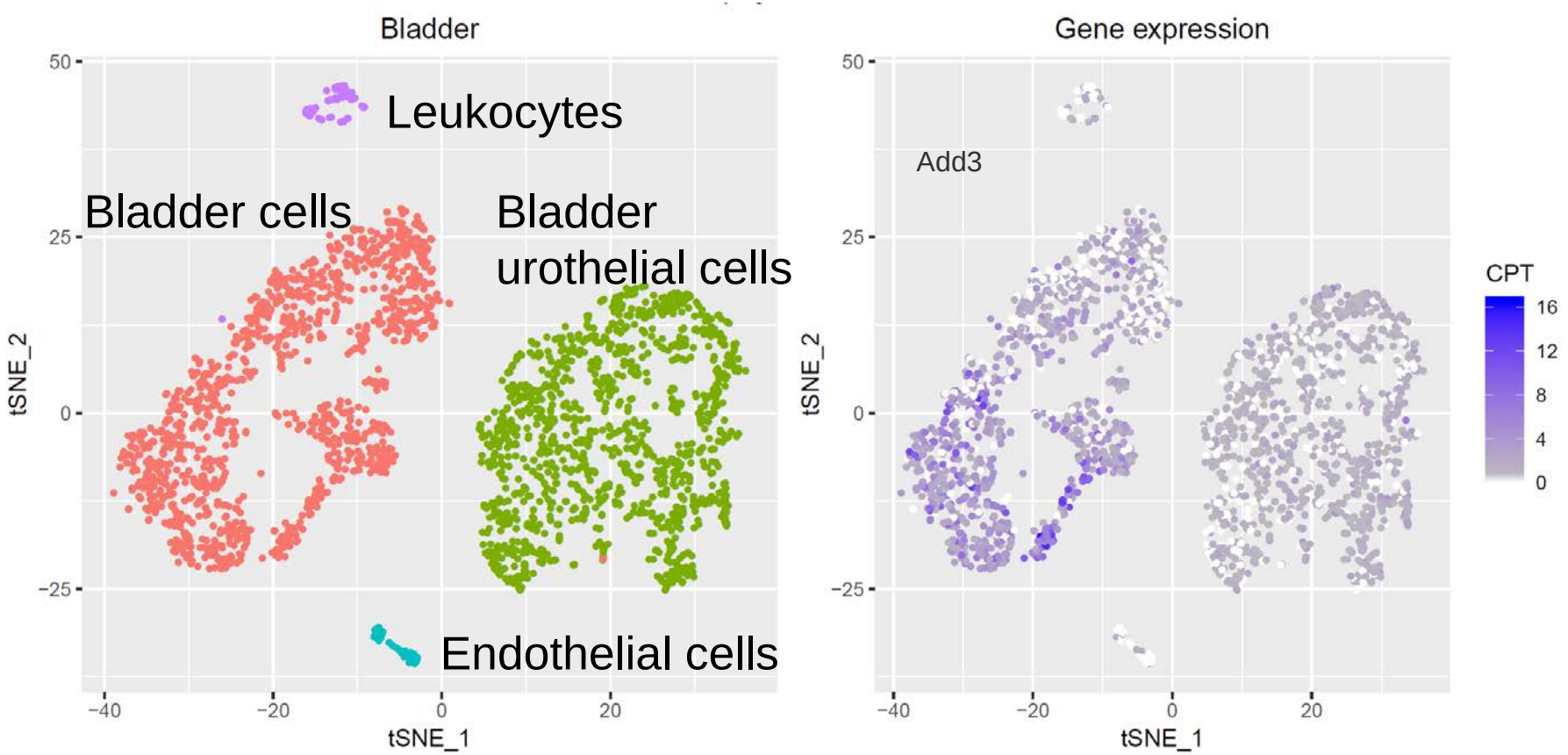
- **STARsolo is much faster than CellRanger**, e.g.:
  10X dataset Pan T cells:
  4.5k cells
  335M reads, 20 threads
  CellRanger: 160 min
  STARsolo: 18 min

- **Alevin and Kallisto map to transcriptome – not genome**
  intronic reads are abundant
  require mapping to the genome

STARsolo adavantages

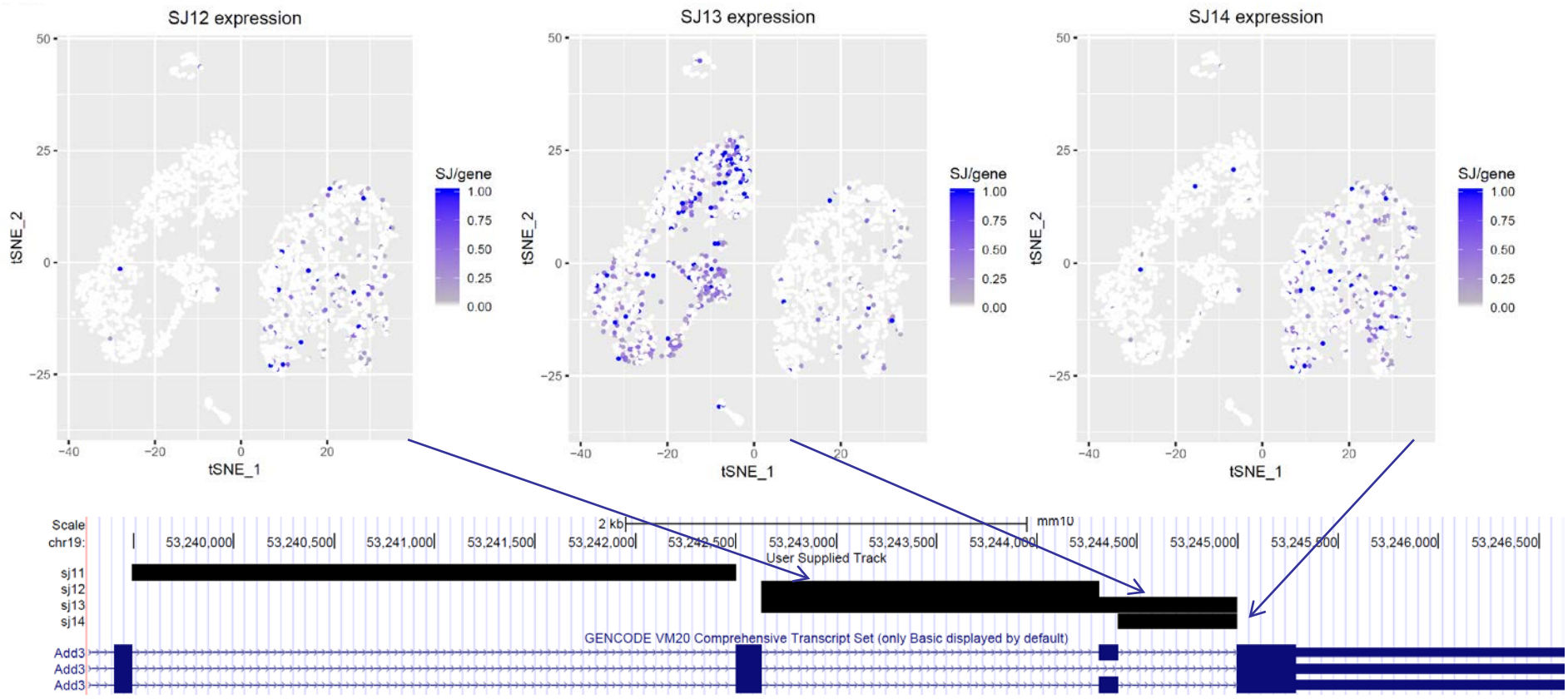
- Gene/cell counts are nearly identical to CellRanger's
  drop-in replacement for unfiltered gene/cell count matrix
- 10x faster than CellRanger
  only 10% overhead over mapping to genome
- `STAR … --soloType Droplet --soloCBwhitelist 10XwhiteList.txt`
- Maps to the genome
  intronic reads, novel isoforms, etc
- Support other protocols
  Drop-seq, SeqWell: no whitelist
  inDrop, Microwell-seq: complex barcodes
  SPLiT-seq, sci-RNA-seq [coming soon]
- Output other transcriptomic features
  splice junctions: annotated and novel
  intronic reads (single-nucleus RNA-seq, velocyto)
  fusions
  alternative polyA, promoters (5' protocol), isoforms [coming soon]

Differential splicing example: HCA Tabula Muris



Add3: gamma-adducin, belongs to a family of membrane skeletal proteins involved in the assembly of spectrin-actin network in erythrocytes and at sites of cell-cell contact in epithelial tissues.

Differential splicing example



Summary

- RNA-seq pipelines
- RNA-seq alignment challenges
- Tweaking mapping parameters
- Post-mapping: STARtools
- single-cell RNA-seq: STARsolo

https://github.com/alexdobin/STAR

https://groups.google.com/forum/#!forum/rna-star