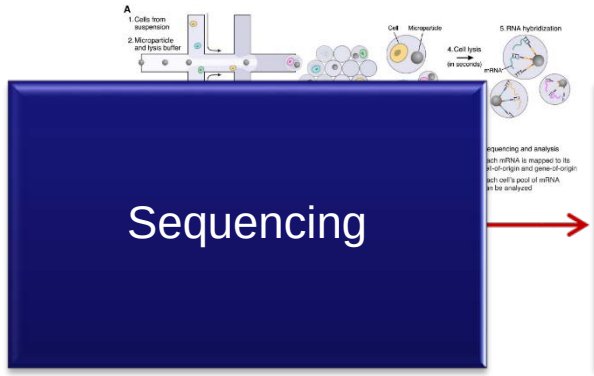
STARsolo:
ultra-fast mapping and quantification of single-cell RNA-seq

Alexander Dobin

Cold Spring Harbor Laboratory

Single-cell RNA-seq processing

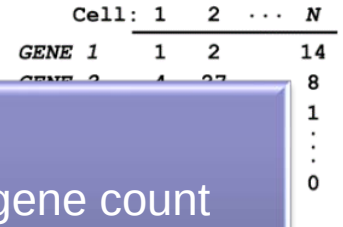
STARsolo



Sequencing

Raw FASTQ reads

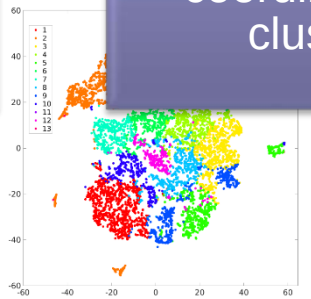
Mapping and Quantification



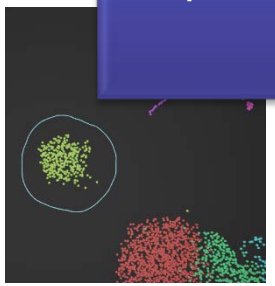
Cell-gene count matrix

Statistical Analysis

Seurat,
Scanpy,
…



Filtered/corrected counts, t-SNE/UMAP coordinates and clustering



Visualization and Exploratory Analysis

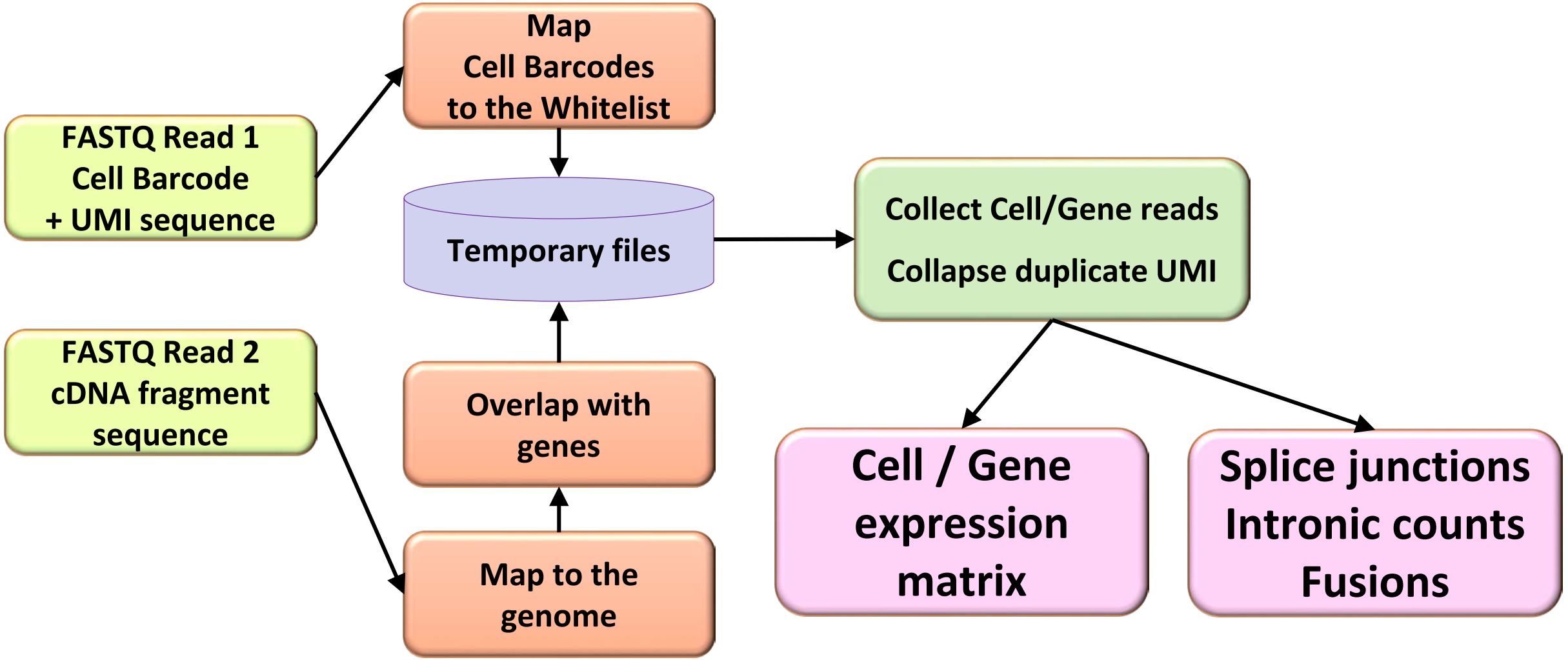
Cell type labeling, differential expression, discoveries!

Mercury

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
- CellRanger main output is gene/cell count matrix
  other transcriptomic features are hidden in the BAM file
- CellRanger cannot be used for other scRNA-seq protocols
  Drop-seq, SeqWell, Split-seq, sci-RNA-seq
  Smart-seq
- Hard to change STAR (or other) parameters
- Not open source, not well documented

STARsolo is integrated into STAR

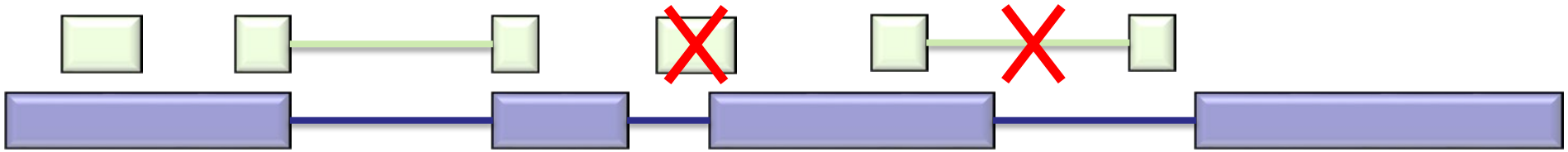


STARsolo follows CellRanger rules for assigining reads to genes, demultiplexing cell barcodes and deduplicating UMIs

Read mapping and assignment to genes

- Reads are mapped to the genome using standard STAR spliced alignment algorithm
  STAR mapping parameters can be modified as in standard STAR run

- Read alignments concordant with transcripts are assigned to genes



- Reads concordant with multiple genes are discarded
  Genomic multimappers that map to paralogs
  Unique mappers that map to overlapping genes

- A genomic multimapper concordant with only one gene is assigned to that gene

CBs and UMIs are represented by integers

2 bits per base

A C G T = { 0 0 } { 0 1 } { 1 0 } { 1 1 }

1 Cell Barcode (16 Bases) = 32 Bits = 1 (32 Bit) Integer

TAGGCCCGAGCCCAGC = 11001010010101100010010101001001 = 3394643273

Cell barcode demultiplexing

- 10X cell barcode whitelist
  Length = 16b
  V2: 700k barcodes
  V3: 6.8M barcodes

- Exact match of the read barcode to the whitelist
  Binary search in the sorted list of barcode integers

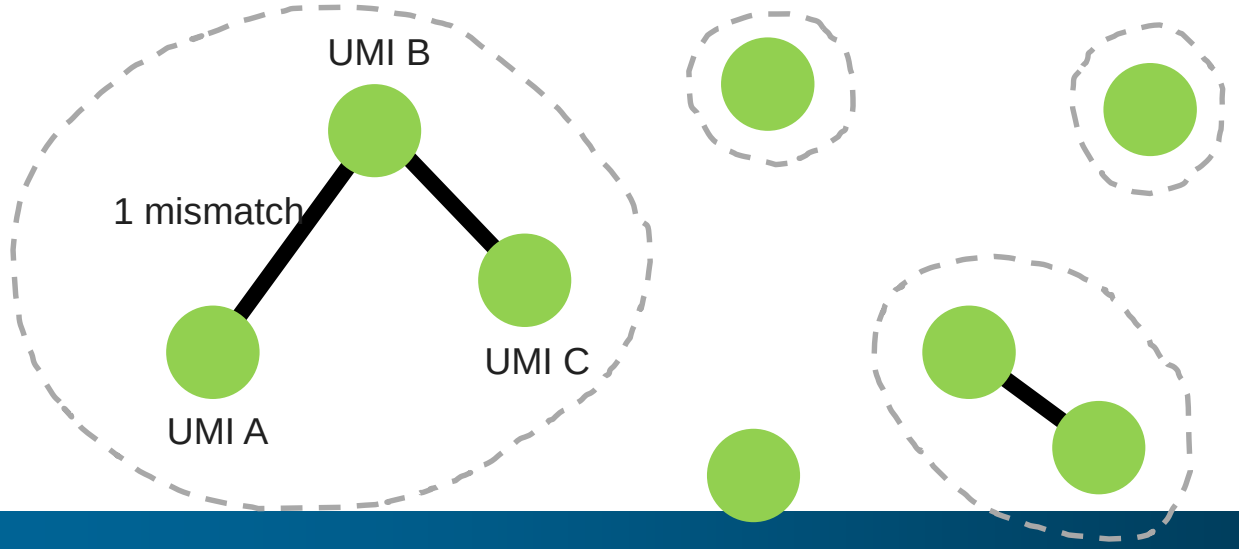
- Read barcode matching whitelist with one mismatch (i.e. Hamming distance = 1)
  If more than 1 hit in the whitelist, calculate posterior probability of each match

$$P_w \propto 10^{-Q_w/10} \cdot N_w$$

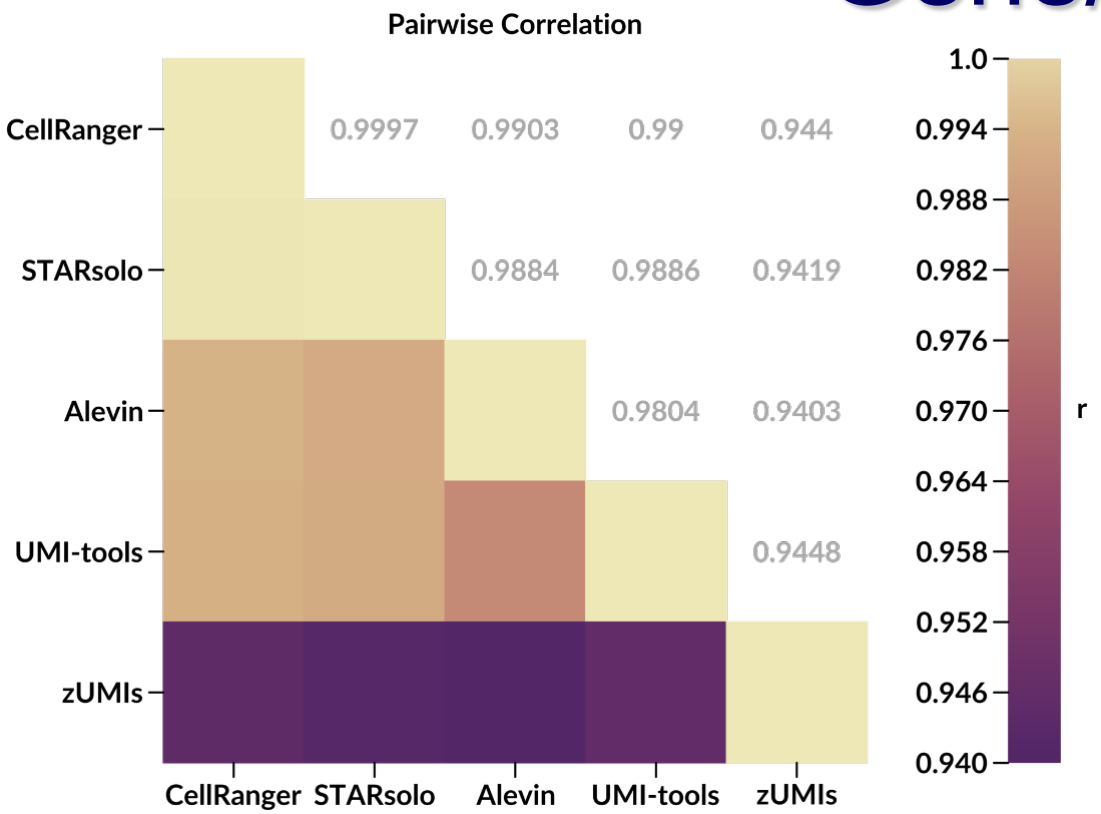
  Assign read to whitelist barcode if probability is > 95%

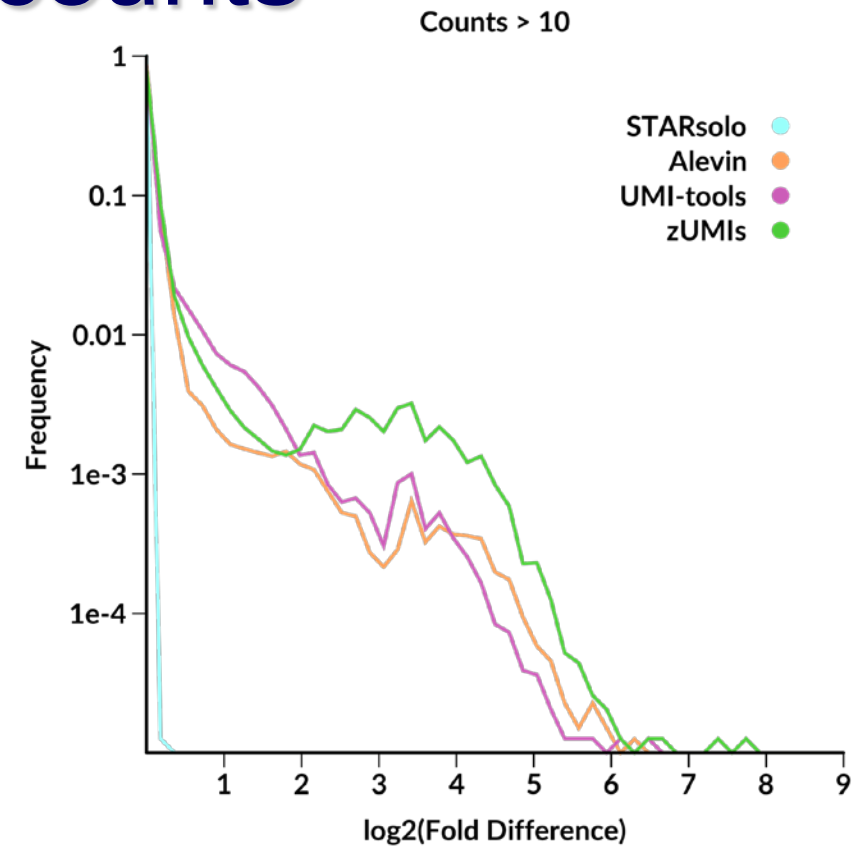
Deduplicating (collapsing) UMIs

- Collapse UMIs within set of reads for each cell/gene
- Collapse identical UMIs
- Collapse UMIs that are within Hamming distance = 1 (i.e. 1 mismatch)
- Graph:
  UMIs are nodes
  UMIs with Hamming D=1 are connected
  Find the number of connected components in the graph.



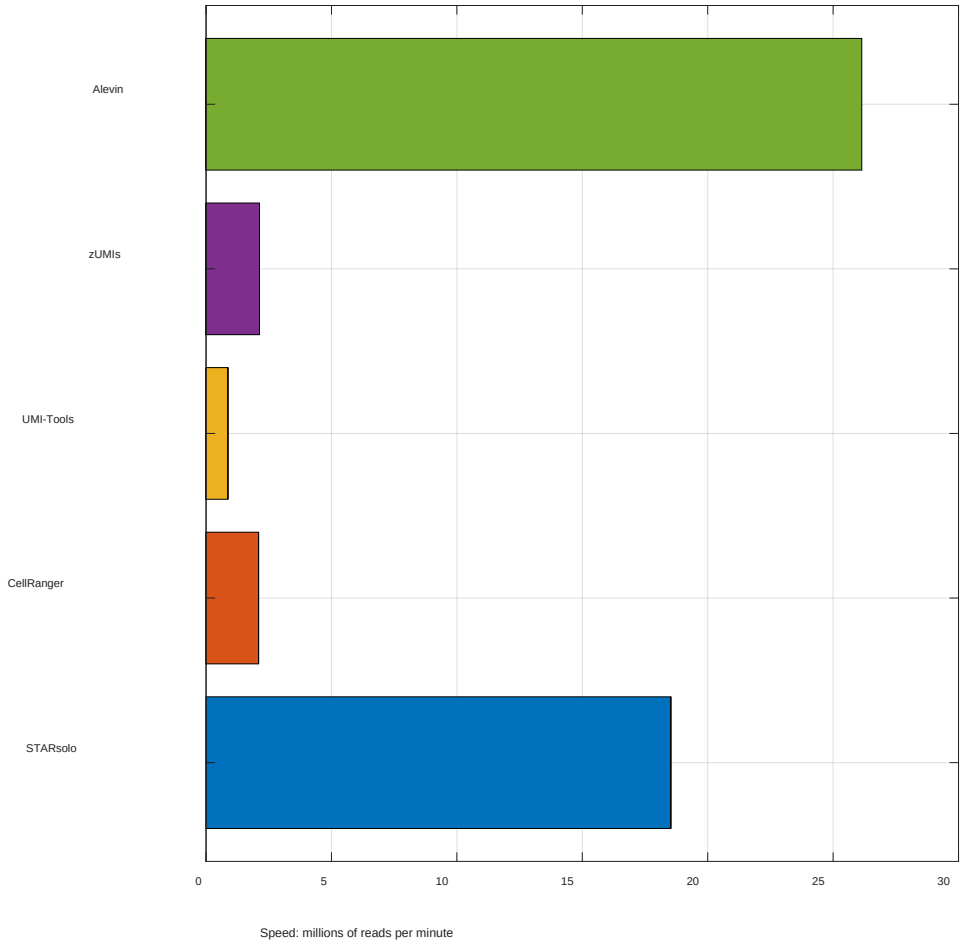
Gene/cell counts





- **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

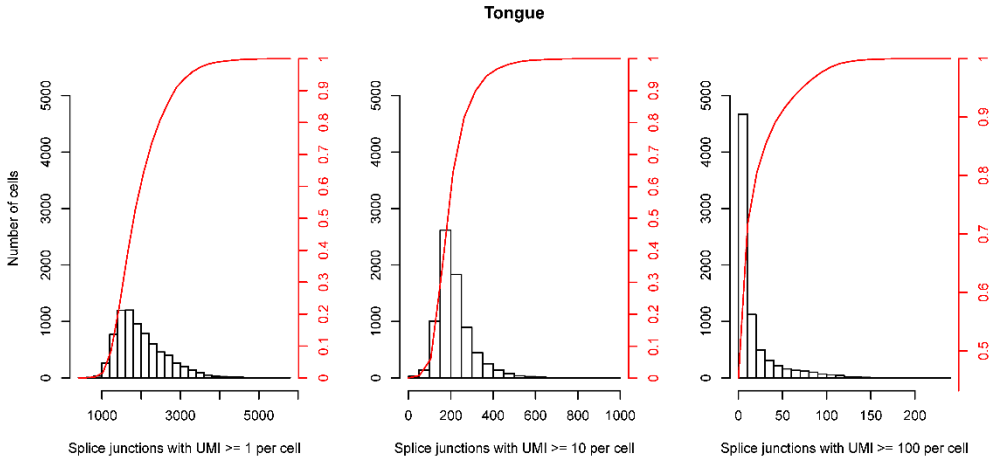
- Alevin (and Kallisto) map to transcriptome – not genome
  intronic reads are abundant
  require mapping to the genome

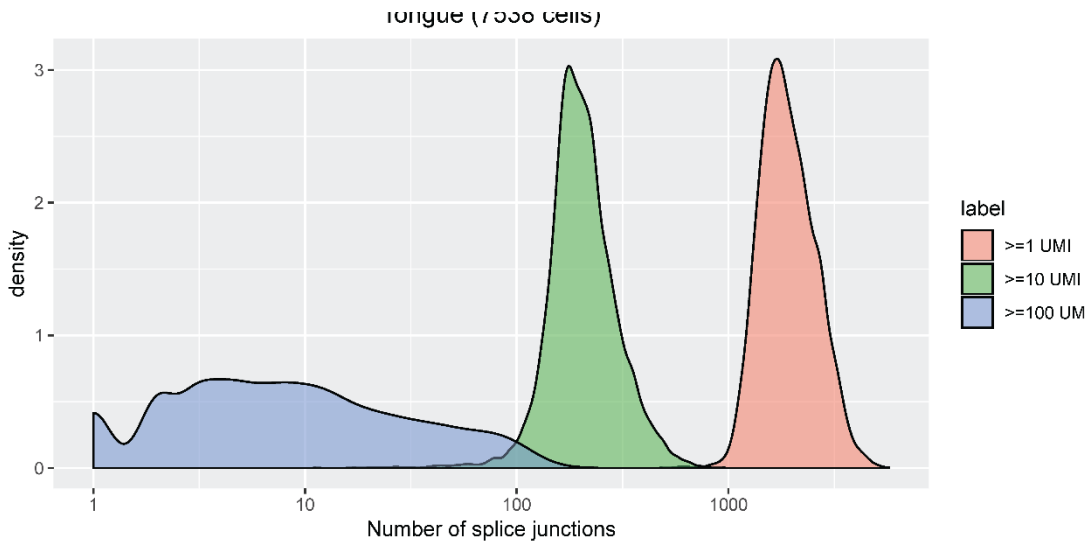
STARsolo advantages

- Gene/cell counts are nearly identical to CellRanger's
  - drop-in replacement for unfiltered gene/cell count matrix
- `STAR … --soloType Droplet --soloCBwhitelist 737K-august-2016.txt`
- 10x faster than CellRanger
  - only 10% overhead over mapping to genome
- Map to the genome
  - intronic reads, novel isoforms, etc
- Support other protocols
  - Drop-seq, SeqWell, etc: no whitelist
  - Split-seq, sci-RNA-seq (coming soon)
- Open source (GPL)
  - open to implementing other features
  - manuscript in preparation
- Output other transcriptomic features
  - splice junctions: annotated and novel
  - intronic reads, fusions
  - alternative polyA, promoters (5' protocol), isoforms (coming soon)

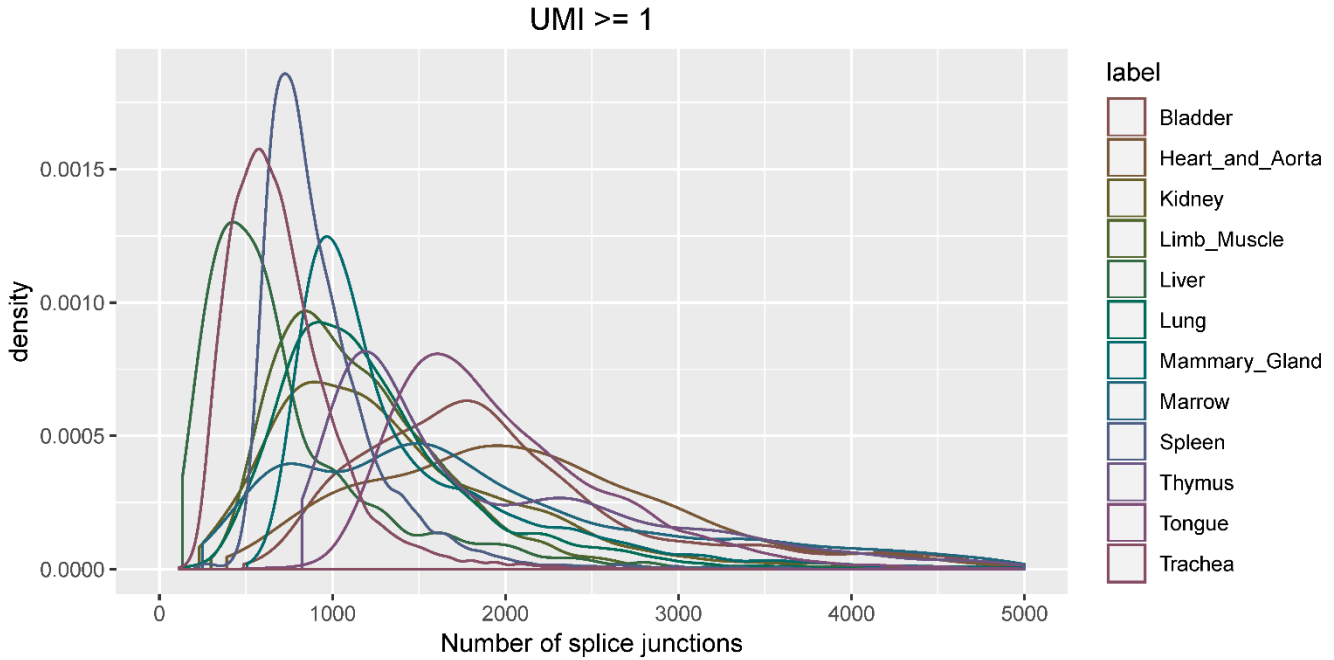
Splice junctions in single-cell RNA-seq

HCA Tabula Muris, 10X dataset





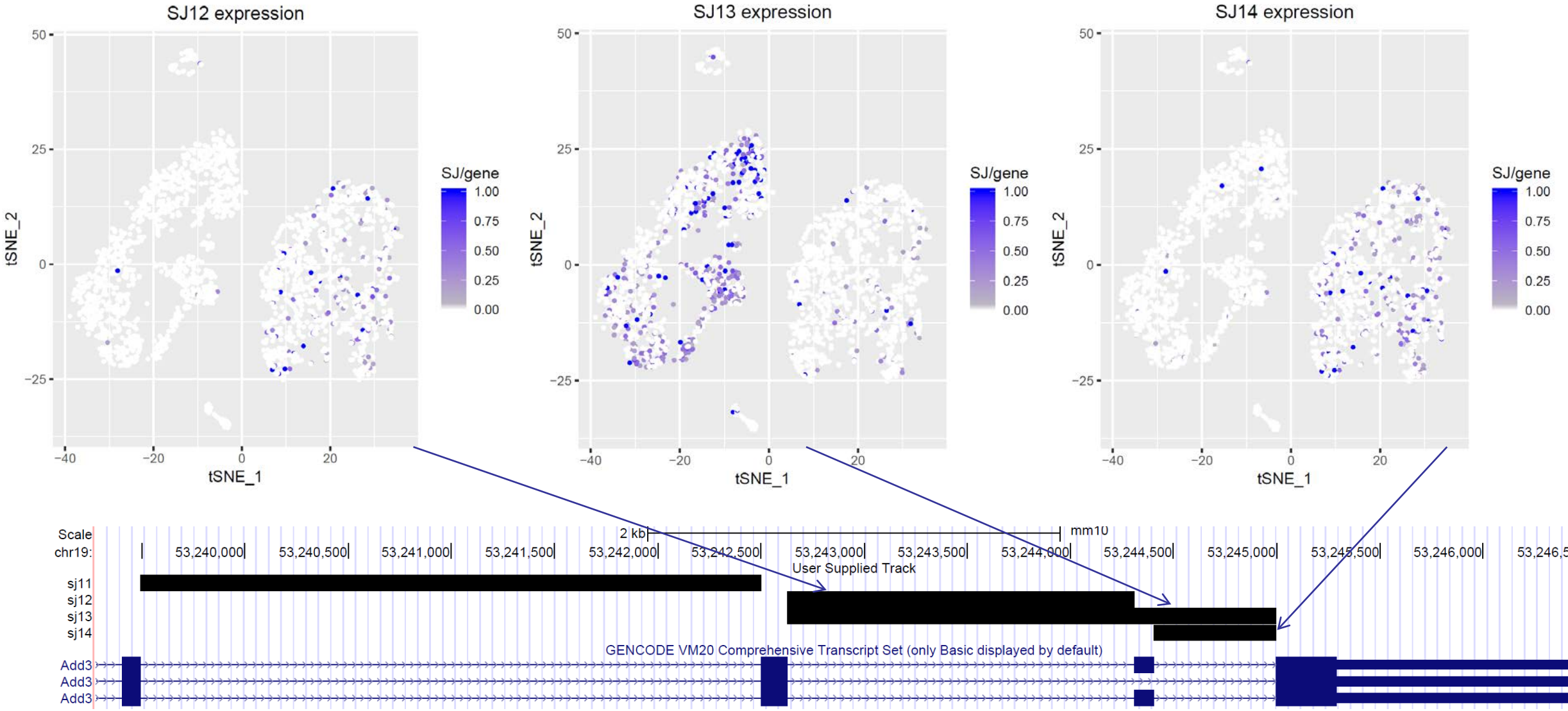
Despite 3' cloning bias, 20-40% of reads are spliced



Differential splicing example

Add3: adducing, 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



Status and plans

- STARsolo introduced in 2.7.0a: Jan 2019
- Current version: 2.7.1a: https://github.com/alexdobin/STAR/releases/tag/2.7.1a
- 2.7.1a

  No cell barcode whitelist operation (for Drop-seq, SeqWell, etc)
  Gene/cell counts for intronic reads (pre-mRNA - single-nucleus)
  Fusions: CB and UMI output to Chimeric.out.junction
  >16b Cell Barcodes

- Future plans

  Support for Smart-seq; sci-RNA-seq, Split-seq, etc
  Counts for alternative polyA sites and transcription start sites
  Alternative isoforms quantification
  Allele-specific counts

  Mercury: 3D cell browser

Acknowledgments

Dobin lab
Ash Blibaum
Jonathan Werner

Preall lab
Jonathan Preall

Gillis lab
Jesse Gillis
Sara Ballouz
Megan Crow

10X Genomics
Patrick Marks

Funding:
NHGRI