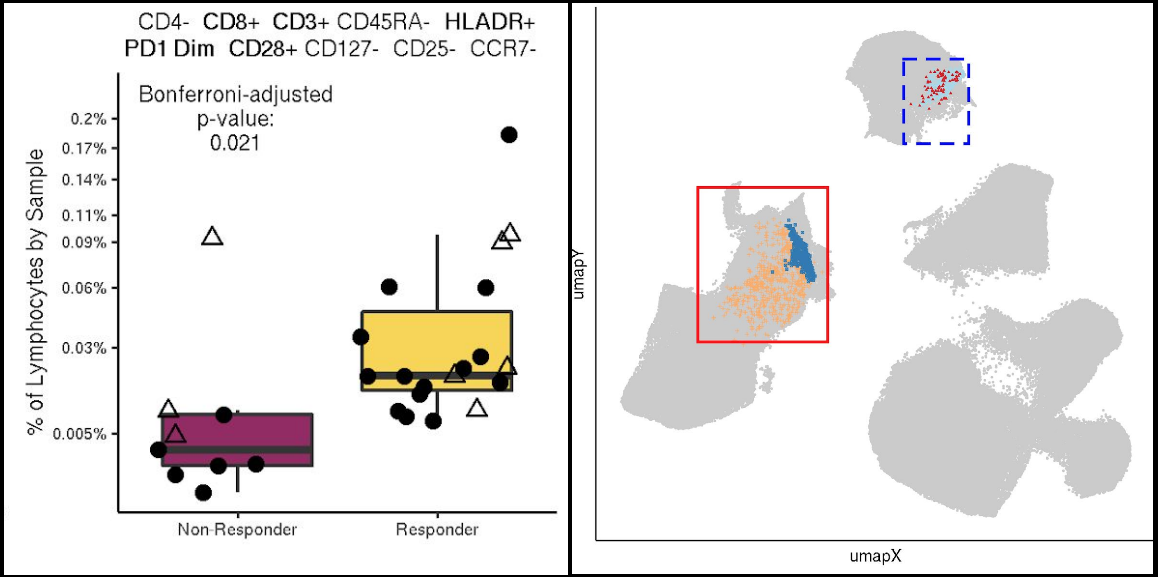
Benchmarking and Methods for Emerging Data

Greg Finak and Matt Ritchie

Fred Hutch Cancer Immunotherapy Program

Fresh Blood at baseline (pre-treatment)



Automated discovery and annotation of a novel biomarker of therapeutic response in MCC.

- High throughput (measure many cells) is critical to detect rare cell populations.
- Dimension reduction is just a visual aid, doesn't tell the full story.

THE FAUST ALGORITHM

An Interpretable machine learning approach

Unambiguously finds all cell populations in a data-driven manner

Complete phenotypic annotations and cell counts for biomarker screening, e.g. CD3+/CD4-/CD8+/PD1 Dim

Robust to biological and technological heterogeneity, diverse marker panels

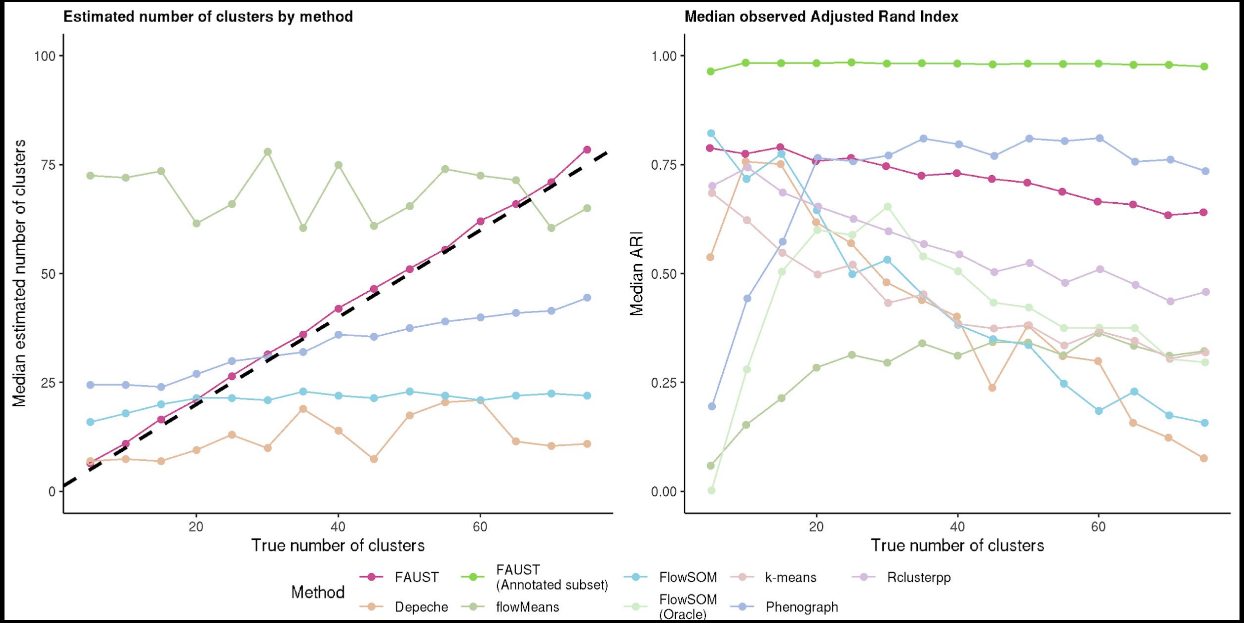
Flow and mass cytometry data can be used for biomarker discovery

Preprint available at
biorxiv.org/content/10.1101/702118v2

FAUST RETURNS A SPARSE SUMMARY OF DATA SETS

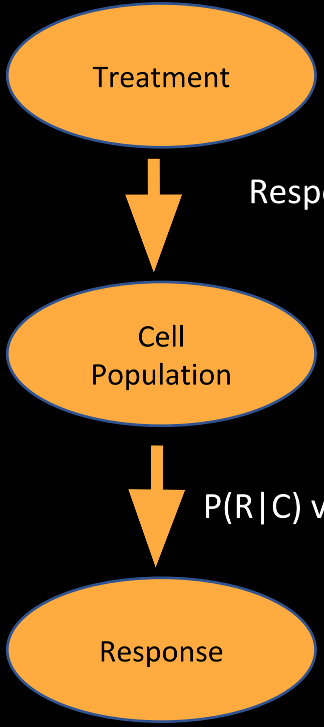
	Total number of samples in data set	Number of markers per data set	Number of Markers Selected by FAUST	Total possible number of phenotypes	Total number of discovered phenotypes	Discovery ratio
Data Set 1	~170	37	21	4,194,304	300	0.01%
Data Set 2	~190	27	19	1,769,472	663	0.04%
Data Set 3	~190	26	19	1,179,648	660	0.06%
Data Set 4	~190	22	18	884,736	909	0.10%
Data Set 5	~170	35	17	131,072	137	0.11%
Data Set 6	~75	18	18	262,144	275	0.11%
Data Set 7	~50	23	16	98,304	119	0.12%
Data Set 8	~190	21	17	131,072	558	0.43%
Data Set 9	~70	25	16	65,536	307	0.47%
Data Set 10	~1,160	16	13	8,192	76	0.93%
Data Set 11	~70	11	11	4,608	162	3.52%
Data Set 12	~360	11	10	3,456	165	4.77%
Data Set 13	~80	11	10	2,304	206	8.94%
Data Set 14	~30	11	9	512	82	16.02%

BENCHMARKING FAUST



- 10 simulated samples
- 10-dimensional data
- Use default settings.
- Estimate the number of cell populations
- Know ground truth.
- Non-gaussian and more realistic
- **How well do we estimate the true number of clusters?**
- **How well do we recover the true cluster structure?**

BENCHMARKING FAUST
SIMULATING AN IMPERFECT BIOMARKER



Treatment

Cell Population

Response

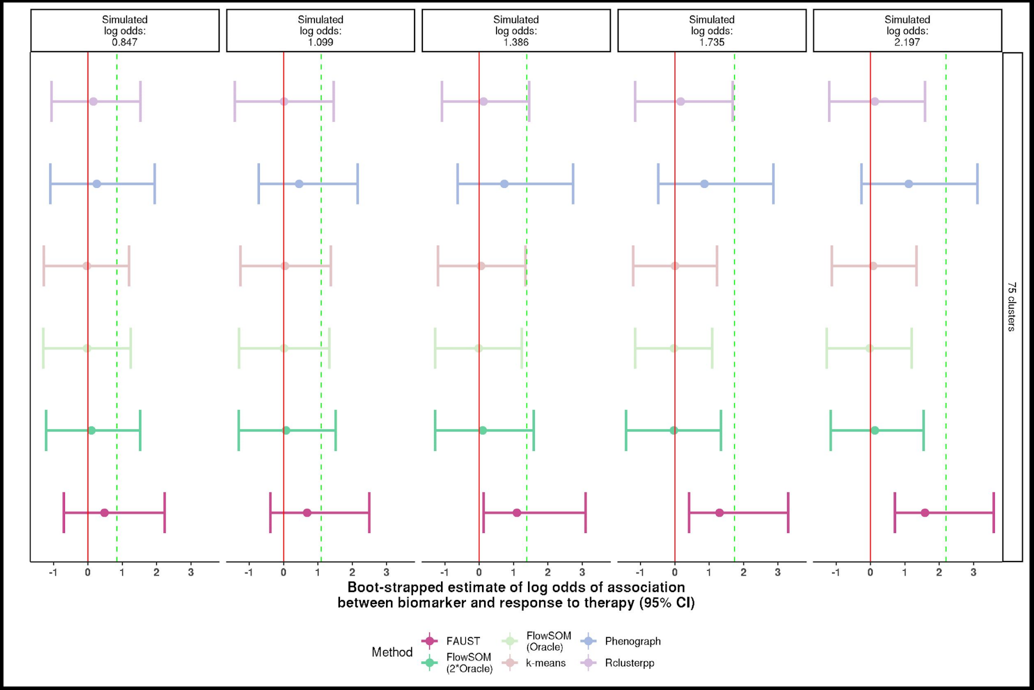
All subjects are treated

Response mediated through some cell population

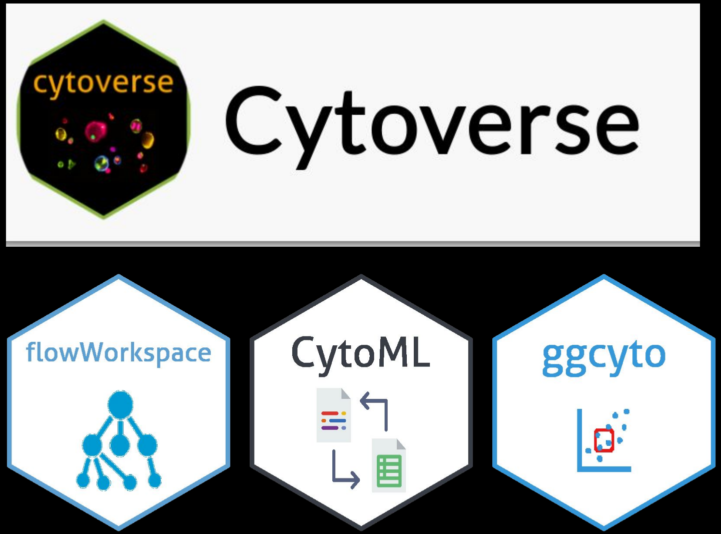
P(C)=0.5: prevalence

P(R|C) varies, biomarker is imperfect

N=100 samples
15,000 cell populations
Aiming to be fair to all methods



SINGLE-CELL SOFTWARE INFRASTRUCTURE



FAUST built on top of the Bioconductor "*cytoverse*" cytometry infrastructure.

- Mature tools with 10+ years of development.
- Disk-backed data storage (hdf5, tiledb).
- Optimized for millions of cells and hundreds of samples per cell.
- Hierarchical representations of cell populations and relationships
- Lots of "historical baggage" from the flow cytometry field.

Single cell RNA Seq

- core BioConductor infrastructure + other tools (Seurat).
  - New technology with ongoing development.
- Need adaptors between the flow cytometry and single-cell RNA seq worlds but it’s a moving target.

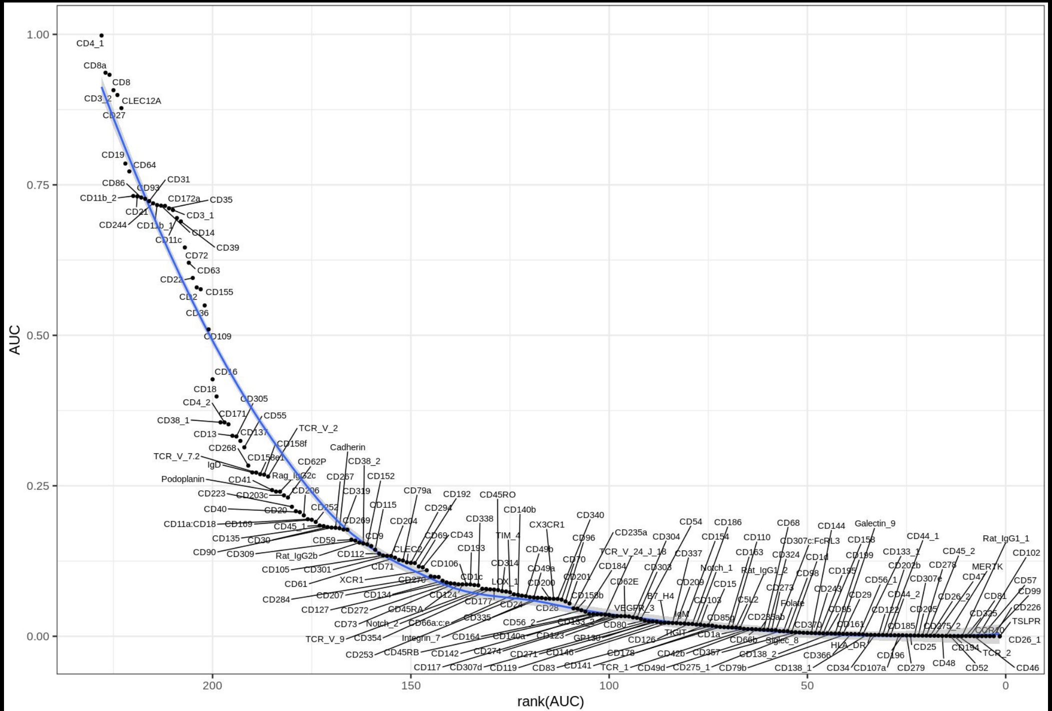
MULTIMODAL PROTEIN AND RNA SEQ TECHNOLOGIES

CITE-Seq, SCITOSeq are high-throughput RNAseq technology that enable simultaneous measurement of transcriptomic and cell surface protein data.

- Antibodies are conjugated with sequence tags.
- Cells encapsulated by drops.
- Each drop can contain zero, one, or more cells.
- Cells from multiple donors.
  - SNP information from sequencing used to identify different donors.
- Limitations
  - Not many large high throughput CITE-Seq data sets available until recently.
  - Small Vx data set (60k cells, 228 markers).
  - T cell data set (13k, 40+markers).
- SCITO-Seq: (Byungjin Hwang *et al.* 2020) : 100k-200k cells 28 markers
  - Each pool has its own sequence tags.
  - Each pool stained with a complete antibody cocktail.

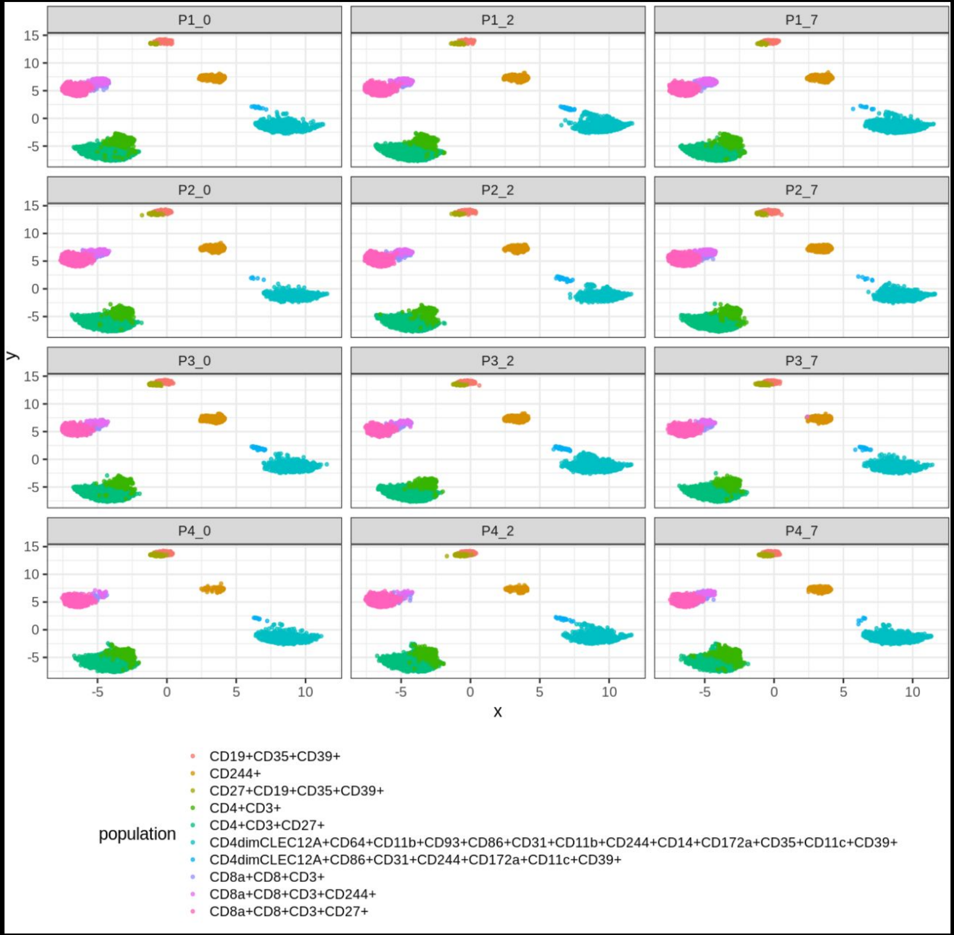
EX: 65K CELLS AND 228 MARKERS

- FAUST tells us how informative is each marker
  - i.e. how reliably a marker can be used to discriminate between cells of different phenotypes.
- Most markers are not very informative.
- Too few cells to reliably use these markers.



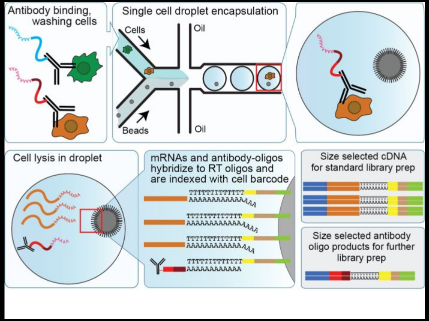
FAUST PHENOTYPIC ANNOTATIONS

- Automated selection of # of populations.
- Phenotypes defined using 19 markers.
- Broadly:
  - B-cells
  - CD4 and CD8 T cells
  - Activated T cells (CD4 and CD8)



Multimodal Data Measures Protein and Gene Expression

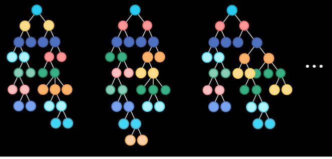
- FAUST resolves complex phenotypes.
- Protein provides much more reliable information.
- Two-stage analysis of protein -> RNA

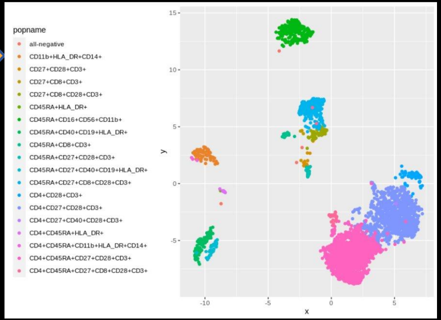


Antibody binding, washing cells
Single cell droplet encapsulation
Cells
Oil
Beads
Oil
Cell lysis in droplet
mRNAs and antibody-oligos hybridize to RT oligos and are indexed with cell barcode
TTTTTTTTTTTT
AAAAAAAAAAAA
Size selected cDNA for standard library prep
Size selected antibody oligo products for further library prep

Protein

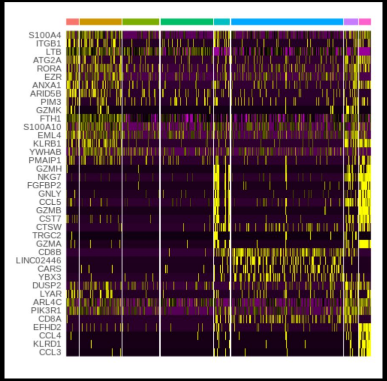
FAUST Cell Population Discovery & Annotation





RNA

MAST cell-type specific differential abundance between discovered phenotypes



Preprint: Greene *et al.* Biorxiv: https://doi.org/10.1101/702118
Code: http://github.com/RGLab/FAUST

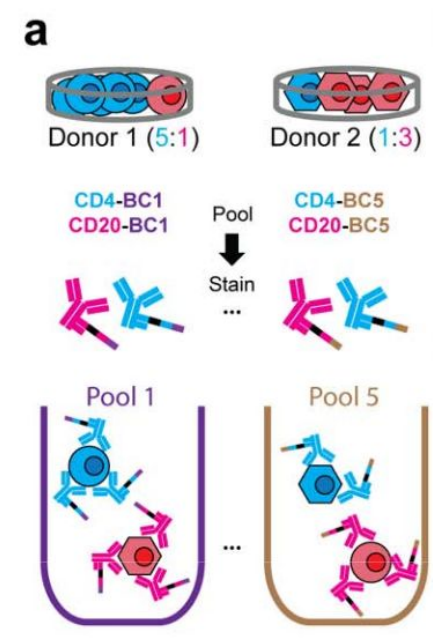
MAST: Finak et al. Genome Biol. **16**, 278 (2015)
Code: http://github.com/RGLab/MAST

Will show an application of this pipeline to data from a recent preprint by Byungjin Hwang *et al.* bioRXiv https://doi.org/10.1101/2020.03.27.012633.

FRED HUTCH
CURES START HERE®

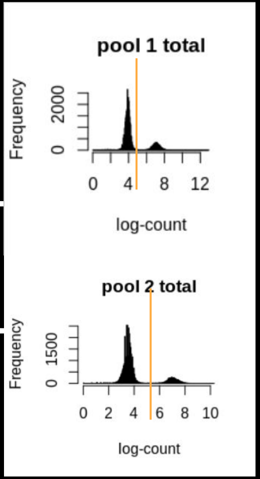
SCITOSEQ: USE INFORMATION FROM MULTIPLE POOLS

- RESOLVE DROPLETS WITH DOUBLETS, MULTIPLETS, ETC



Droplet id

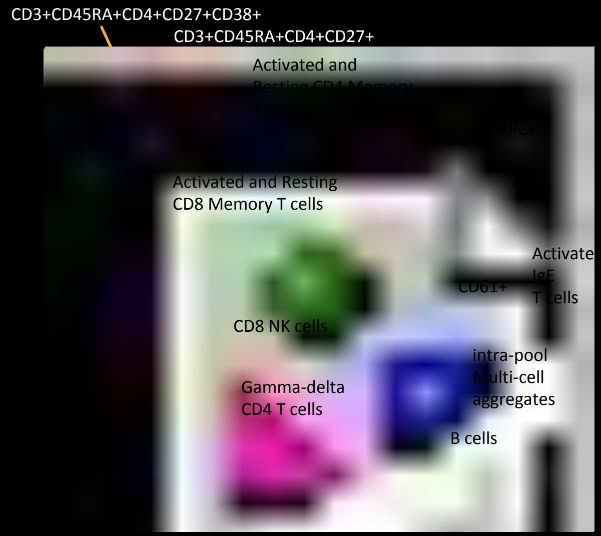
Protein x pool id	a	b	c	d	e
Protein 1 pool 1	200	500	12	211	20
Protein 2 pool 1	10	15	112	79	10
Protein 1 pool 2	5	432	171	135	210
Protein 1 pool 2	21	4	4	30	11
...					

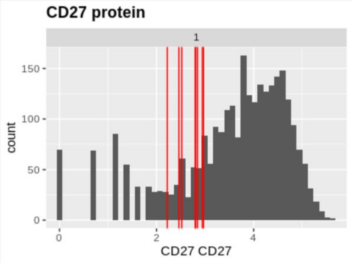


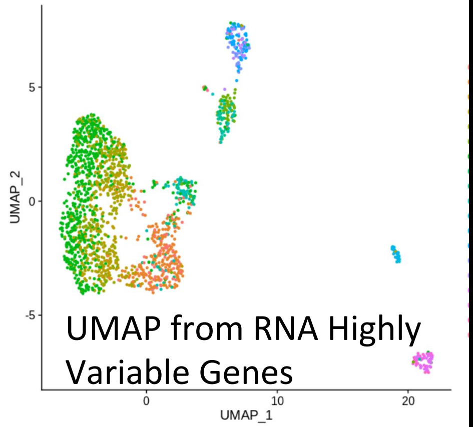
Byungjin Hwang, et al. 2020

Application of FAUST to Phenotyping Multiplexed CITE-Seq Data

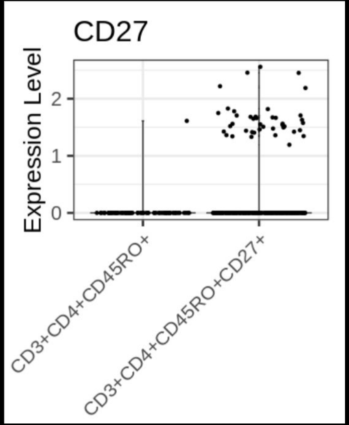
100k PBMCs & 28 markers, 10 pools
24 Cell Populations Annotated

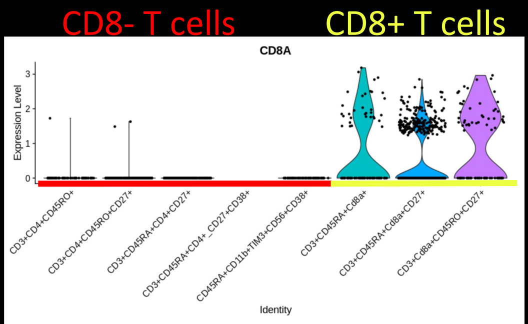






UMAP from RNA Highly Variable Genes





- FAUST resolves complex phenotypes not immediately obvious in dimension reduction.
- Protein provides much more reliable information than gene expression.
- Two-stage analysis of protein -> RNA

FRED HUTCH
CURES START HERE®

Accurate Cell Population Annotation will be Critical to Make the Most of Integrated Single-Cell Data

- Use methods and lessons learned from flow cytometry for automated phenotype assignment in multimodal single-cell CITE-seq data.
- Larger data sets will be critical for making the most of these technologies.
- Building infrastructure to integrate flow / mass cytometry single cell data and scRNASeq data.
    - Benchmark against manual / expert annotation

Acknowledgements

Evan Greene

Raphael Gottardo

Paul Nghiem

Candice Church

Leonard D’Amico

Steve Fling

Nirasha Ramachandran

Chan Zuckerberg Initiative

Human Cell Atlas

Cancer Immunotherapy Trials Network

FRED HUTCH
CURES START HERE®

NIH
National Institutes of Health

Characterization of full-length isoforms in single cells with Nanopore long-read sequencing and FLAMES

Matt Ritchie

FIAT LUX
Walter+Eliza Hall
Institute of Medical Research

Single cell RNA sequencing with nanopore sequencing technology

Obtain full-length cDNA during 10X library preparation and sequence on Oxford Nanopore Technologies (ONT) PromenthION platform

+

Advantage:

Full gene-body coverage

Isoform characterization at single cell level

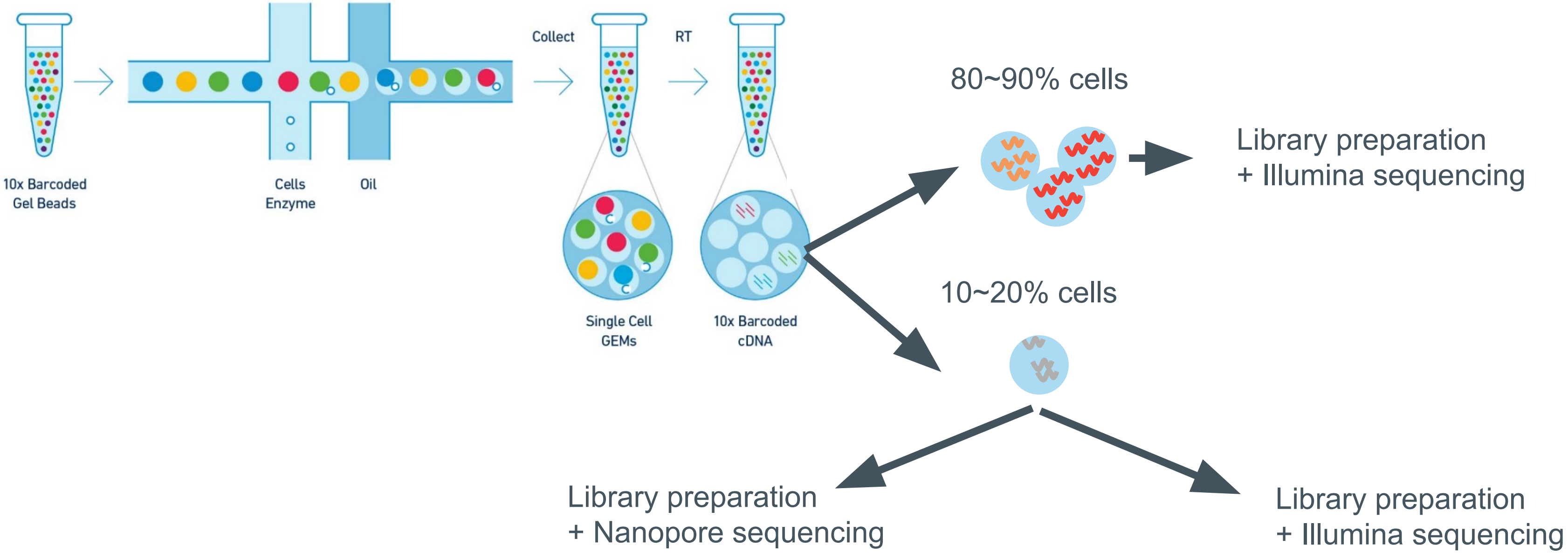
Challenge:

Limited throughput compared to short reads (~40M PromethION vs ~400M Nextseq)

Limited tools for data analysis

ScISOr-Seq Gupta *et al.* Nat Biotechnol 2018
RAGE-seq Singh *et al.* Nat Commun 2019
Lebrigand *et al.* bioRxiv 2019

*F*ull-*L*ength *T*ranscriptome *seq*uencing (*FLT-seq*)



80~90% cells

Library preparation
+ Illumina sequencing

10~20% cells

Library preparation
+ Nanopore sequencing

Library preparation
+ Illumina sequencing

Protocol by Jafar Jabbari, AGRF
Melbourne

FLT-seq can be used on different cell types and 10X scRNAseq kits

Cell lines (n=2)

*scmixology**: Equal mixture of cells from 5 cell lines

10X v2 and v3
scATAC-seq

Mouse stem cells (n=1)

control

injury

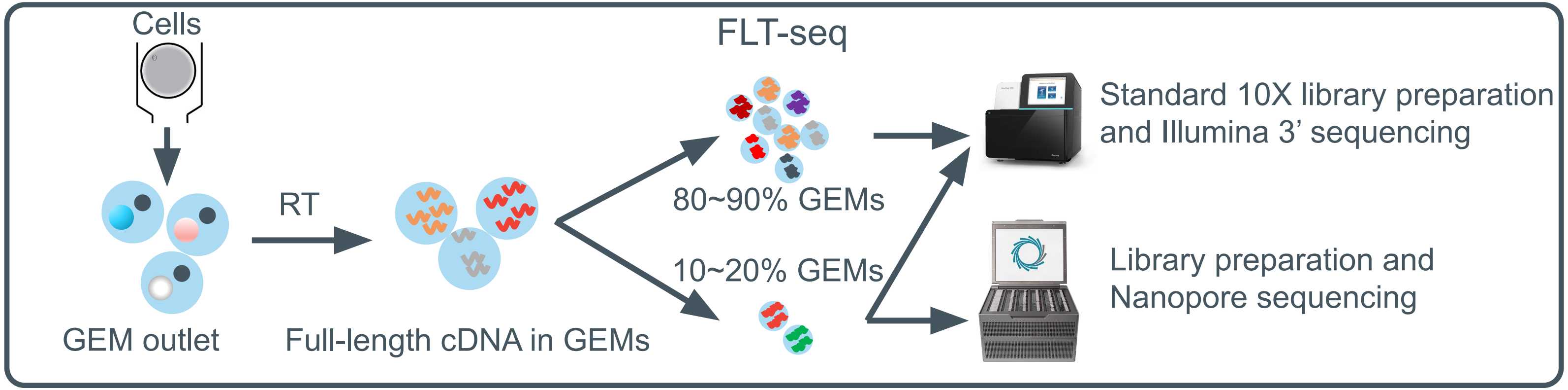
Muscle stem cells w/wo activation

10X v2

Patient samples (n=1)

PBMCs from relapsed CLL patients after Venetoclax treatment

10X CITE-seq



FLT-seq

Cells

GEM outlet

RT

Full-length cDNA in GEMs

80~90% GEMs

10~20% GEMs

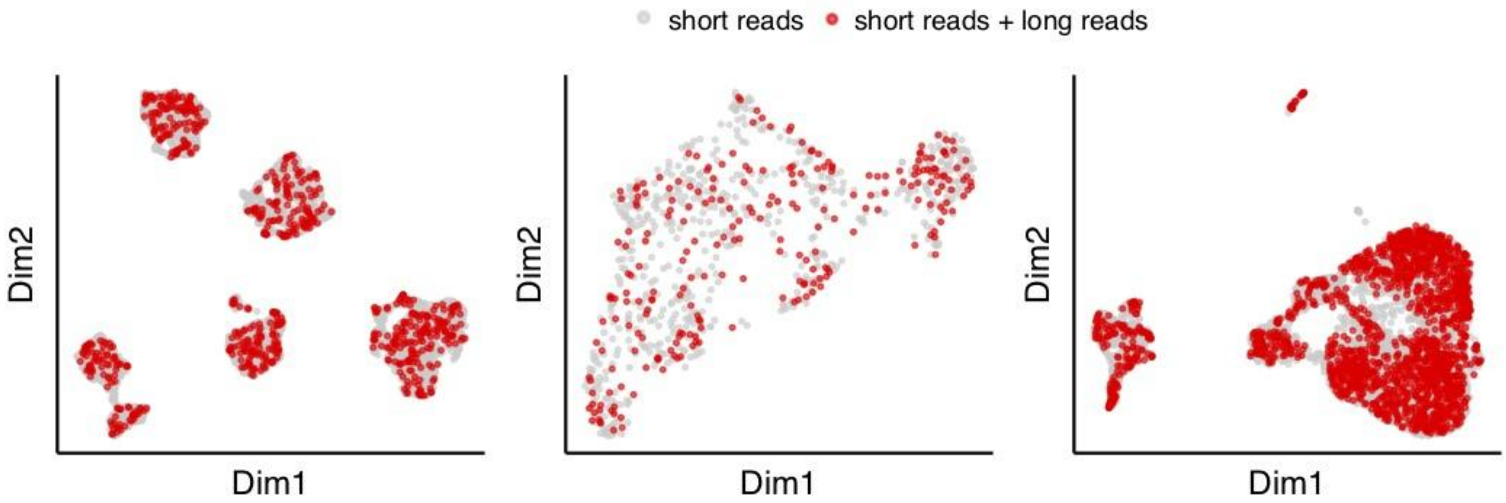
Standard 10X library preparation and Illumina 3’ sequencing

Library preparation and Nanopore sequencing

* Tian *et al*. Nat Methods 2019

UMAP visualization of representative samples shows uniform sampling of cells

In total, we profiled ~2,500 single cells using PromethION, together with ~16,000 cells with Illumina short-reads

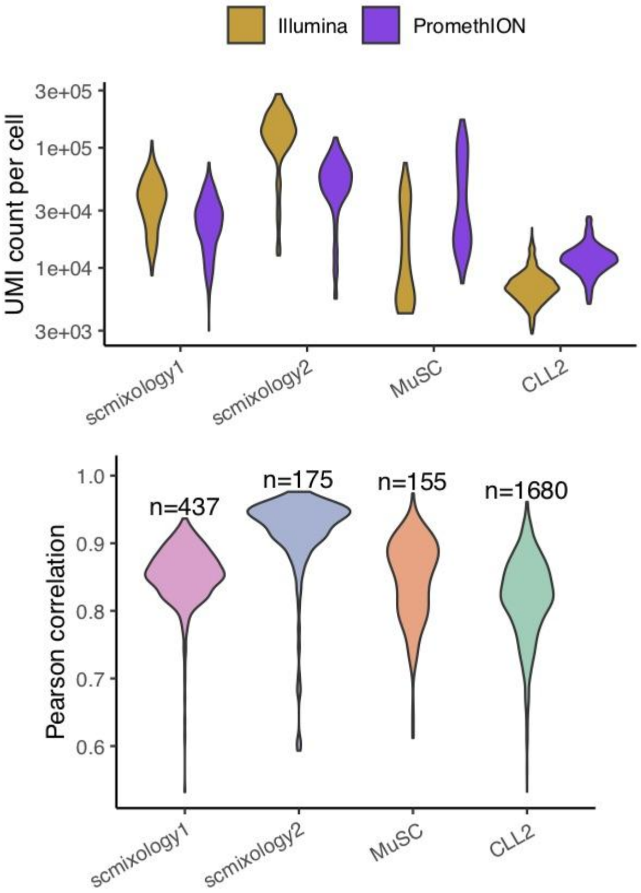


scmixology 1+2

Mouse muscle stem cells

CLL patient PBMCs (CLL2)

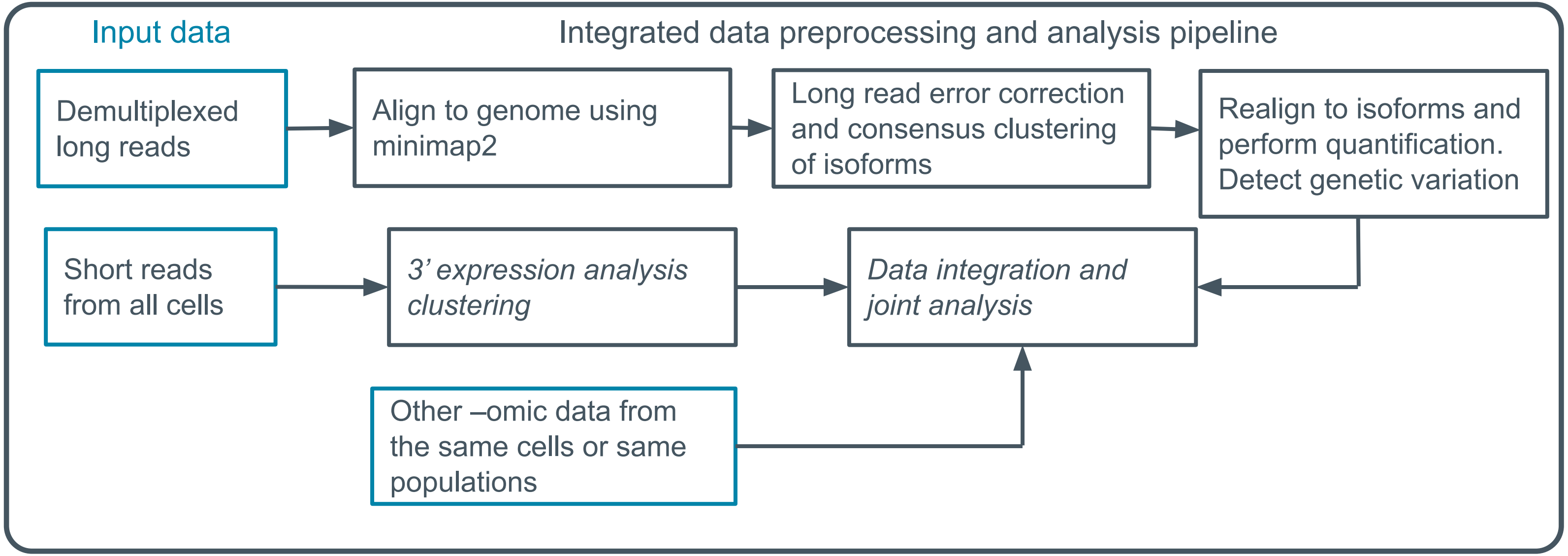
Basic QC, comparison to short read data



- 10X cell barcode could be detected in 40-60% of long-reads
- similar sequencing depth per cell achieved in long and short-read data
- high correlation in gene level quantification between long and short-read data

Data analysis pipeline

A new toolbox called ***FLAMES*** (***F***ull-***L***ength tr***A***nscript quantification, ***M***utation and ***S***plicing analysis for long-read data) was developed.



Input data

Integrated data preprocessing and analysis pipeline

Demultiplexed long reads

Align to genome using minimap2

Long read error correction and consensus clustering of isoforms

Realign to isoforms and perform quantification. Detect genetic variation

Short reads from all cells

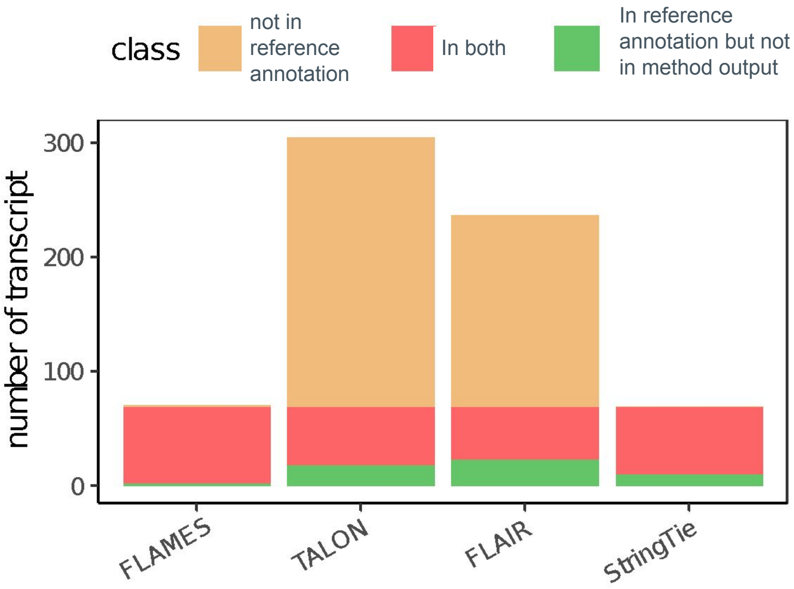
3' expression analysis clustering

Data integration and joint analysis

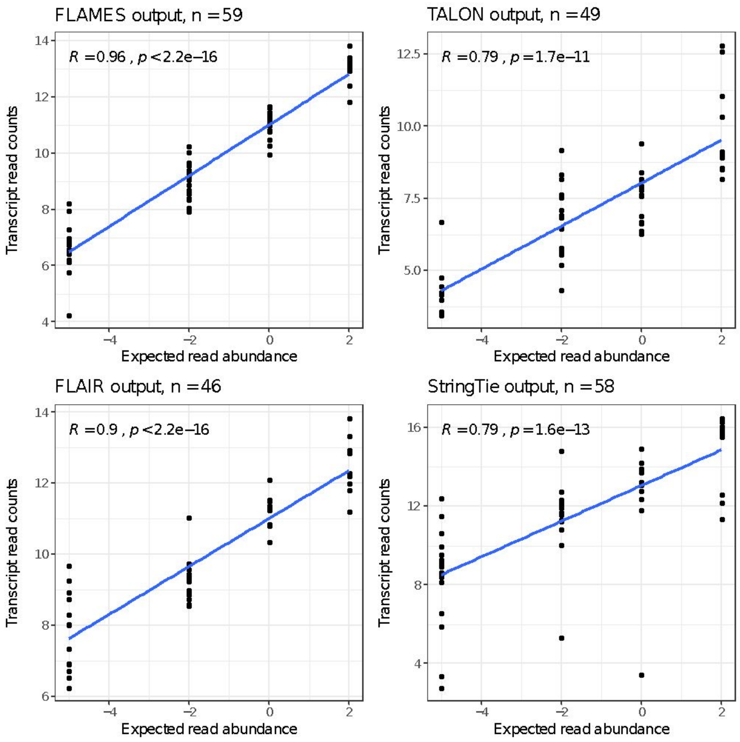
Other –omic data from the same cells or same populations

Benchmark isoform detection and quantification using SIRV spike-in dataset*

Isoform detection



Isoform quantification


FLAMES output, n = 59

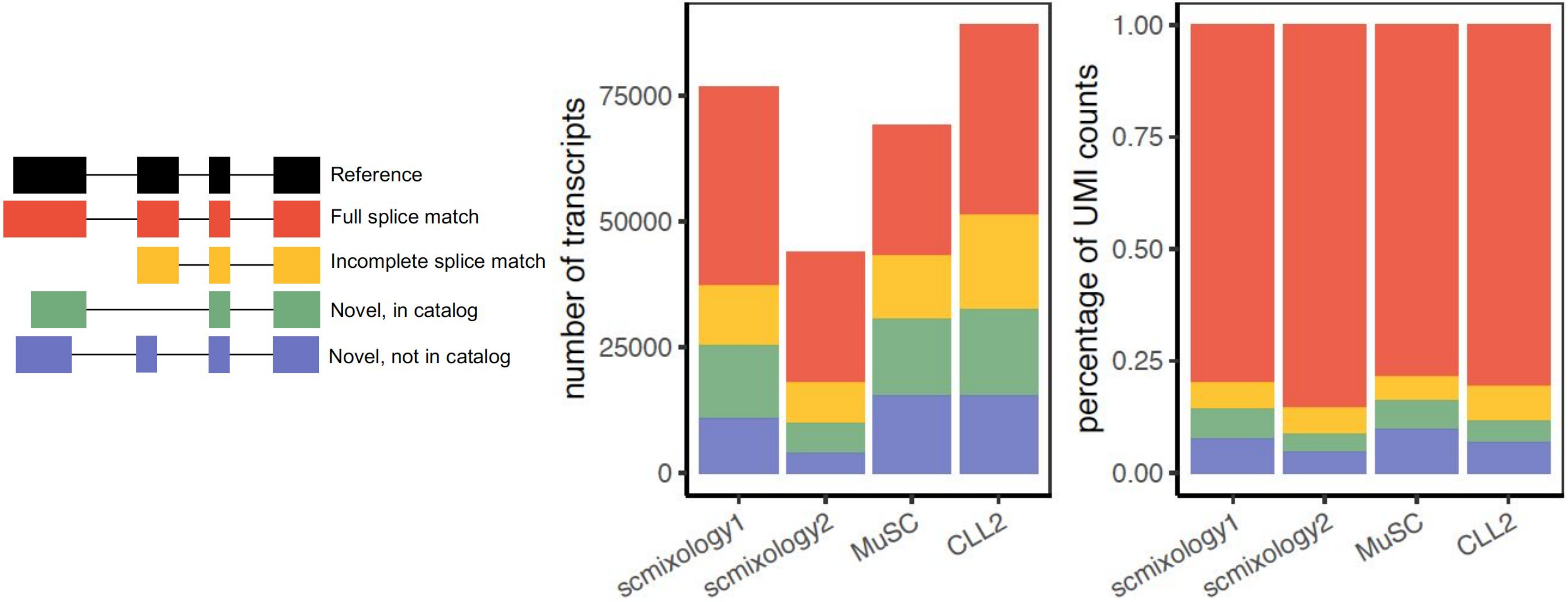
TALON output, n = 49

FLAIR output, n = 46

StringTie output, n = 58

* Garalde *et al.* Nat Methods 2018
Dong, Tian *et al.* bioRxiv 2020

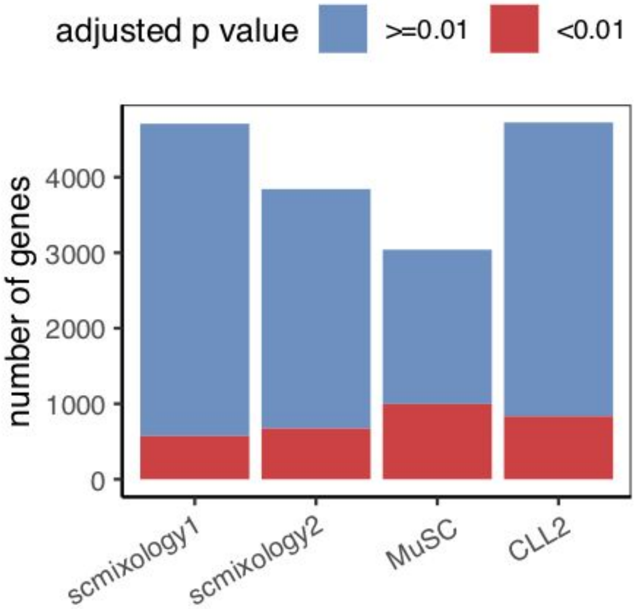
Characterization of isoforms



SQANTI2 (https://github.com/Magdoll/SQANTI2)
Tardaguila *et al.* Genome Research 2018

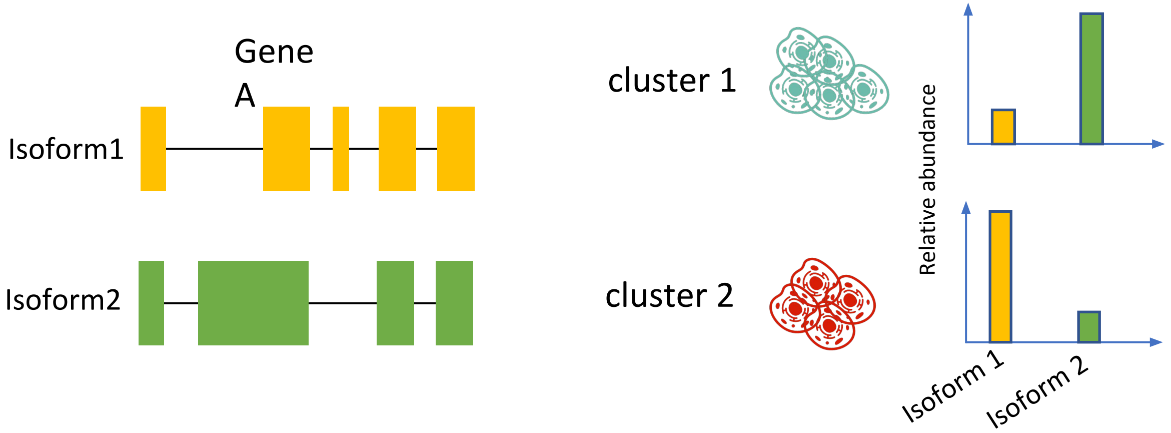
Differential transcript usage analysis

Find cluster/cell type specific transcripts

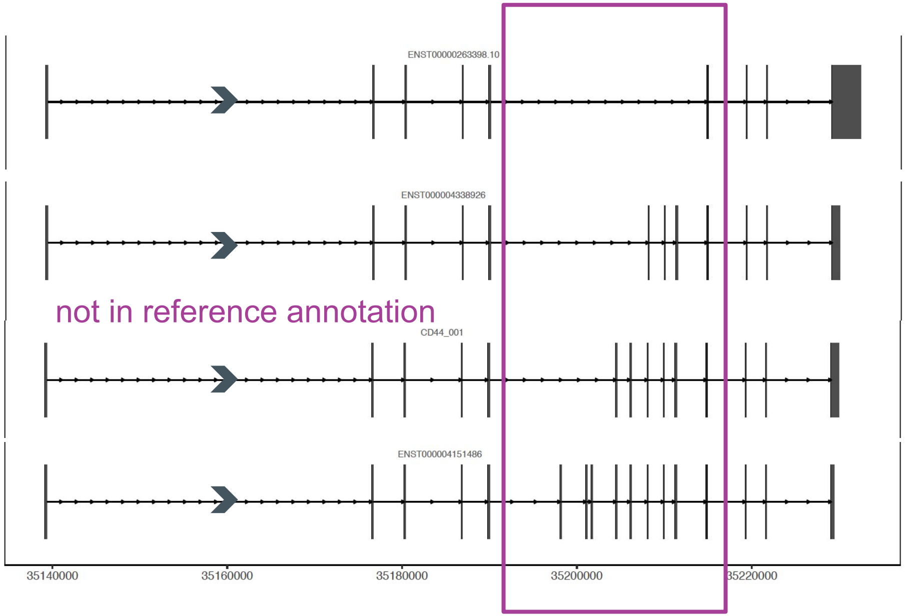


- After filtering by abundance, test for proportion differences between isoforms in different clusters for each gene
- 200 - 1,000 genes with *p*-value < 0.01

Find cluster/cell type specifc transcript

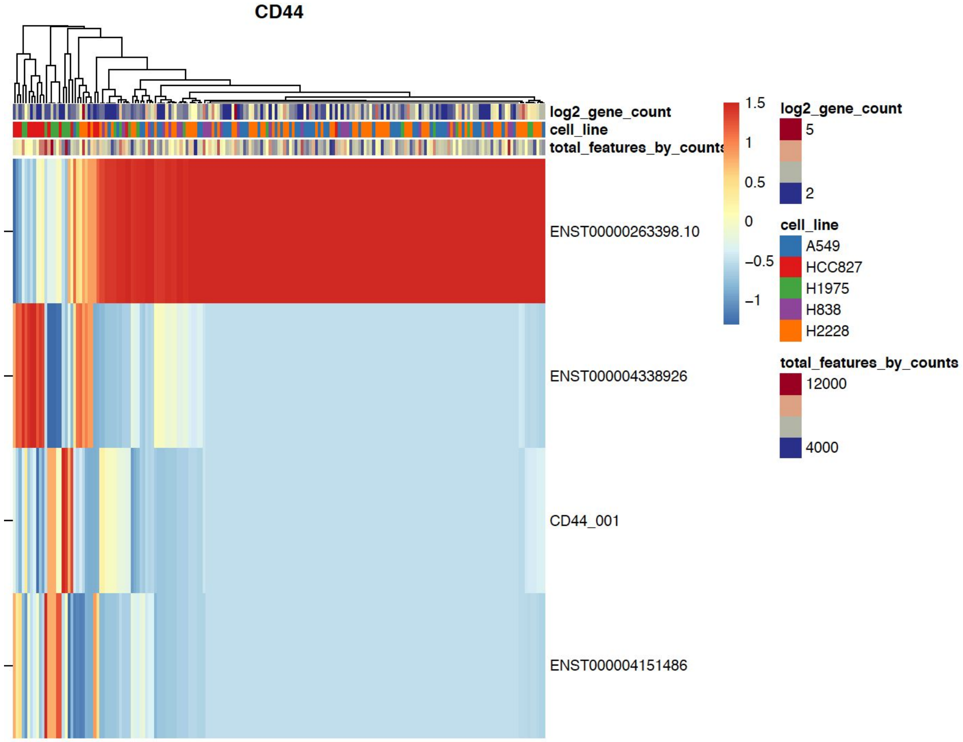


Expression of different CD44 isoforms in *scmixology* data

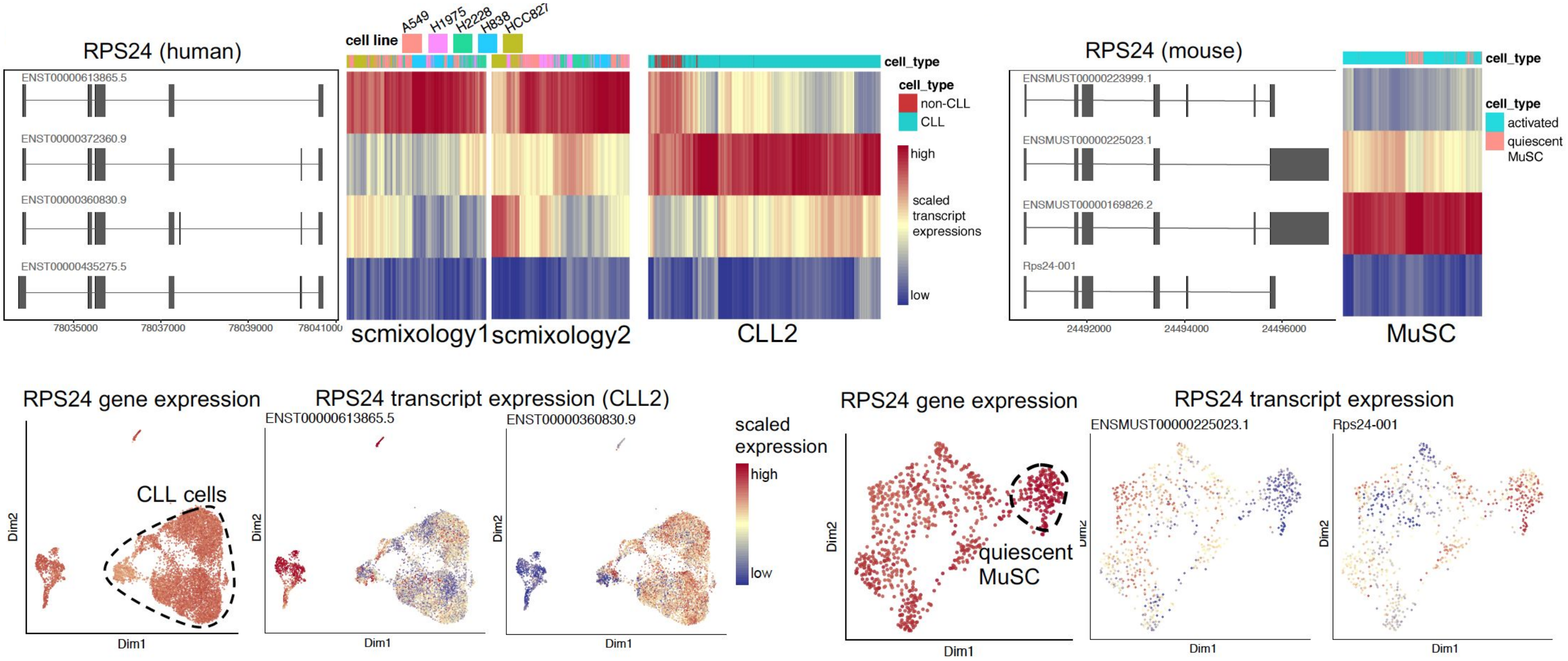


not in reference annotation

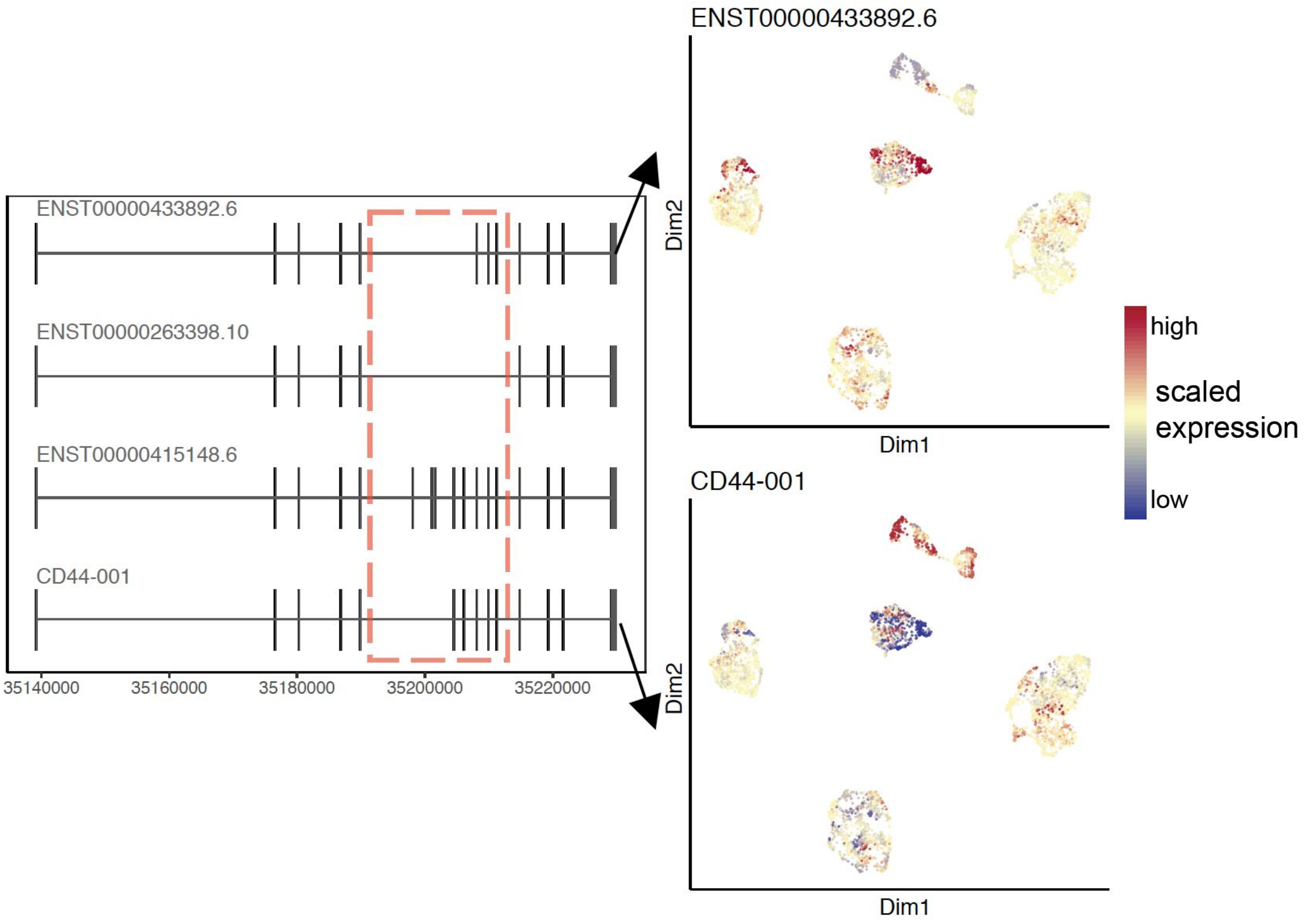
variable domain in protein



Expression of different RPS24 isoforms

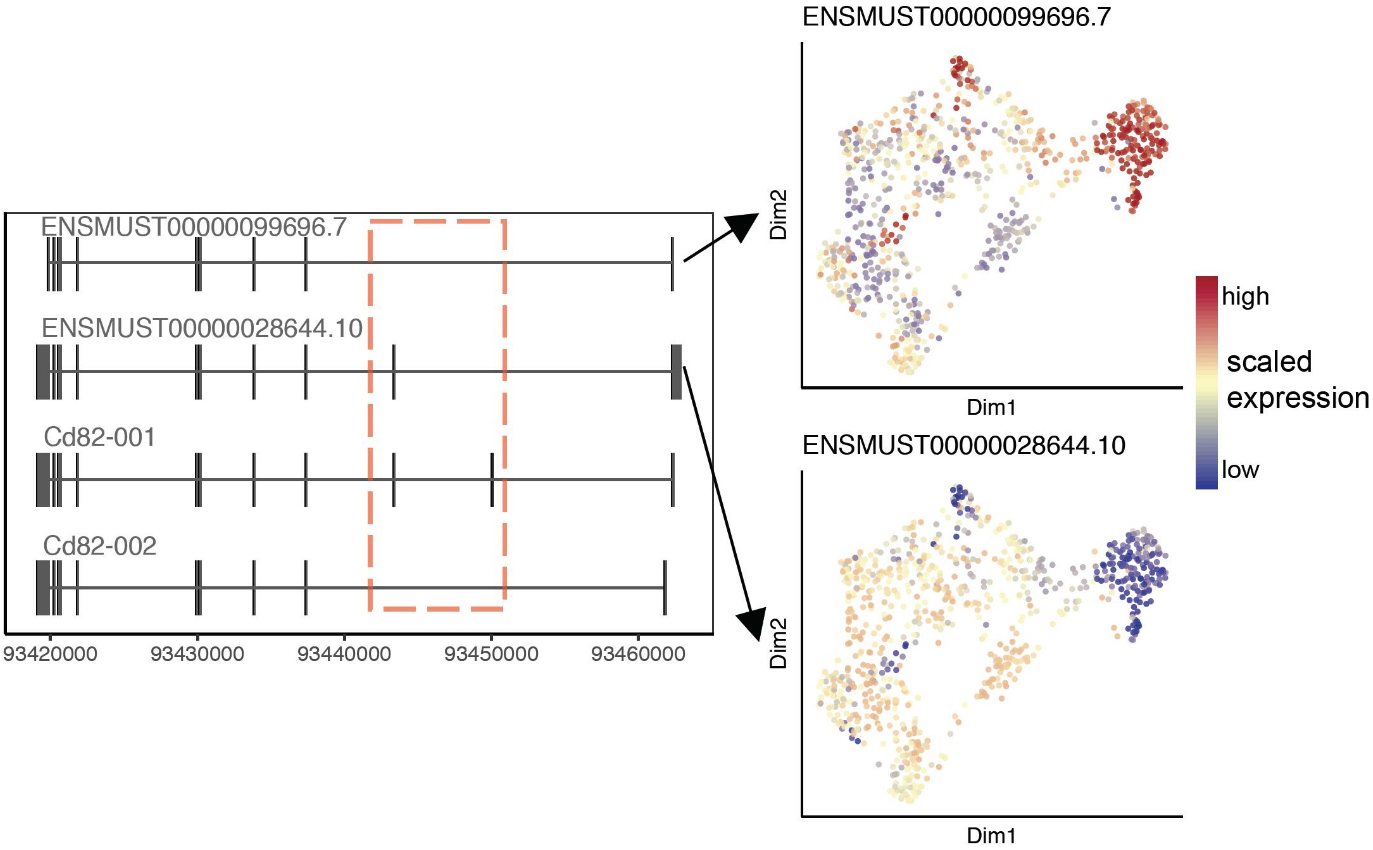


Expression of different CD44 isoforms in *scmixology* data

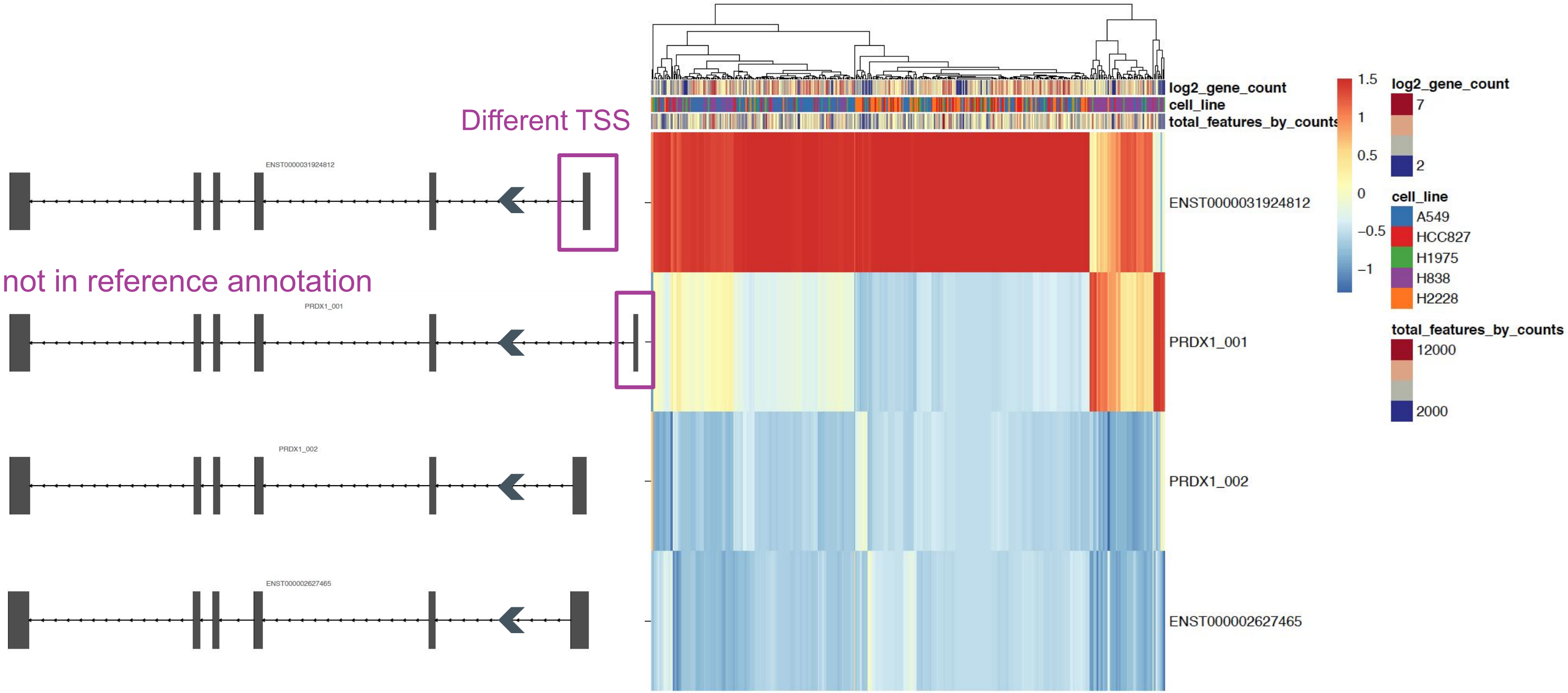


not in reference annotation

Expression of different CD82 isoforms in MuSC data



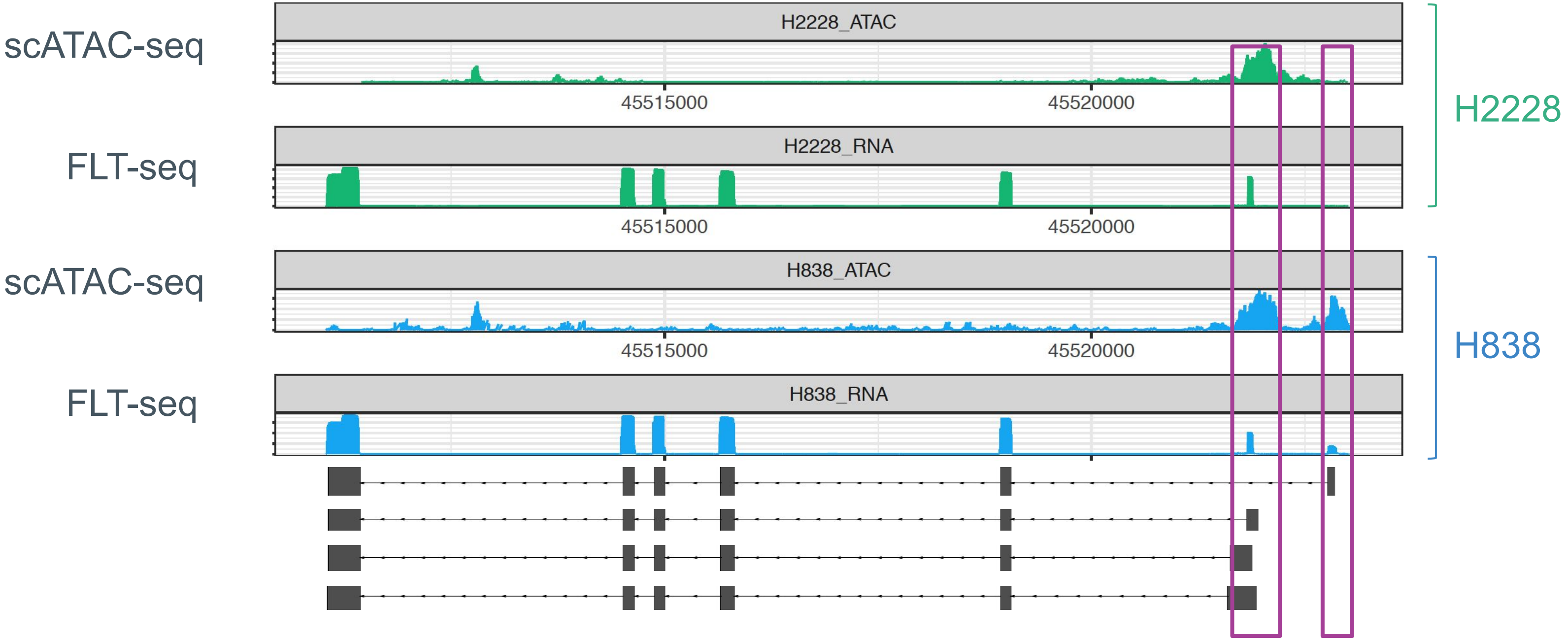
Expression of different PRDX1 isoforms in *scmixology* data



Different TSS

not in reference annotation

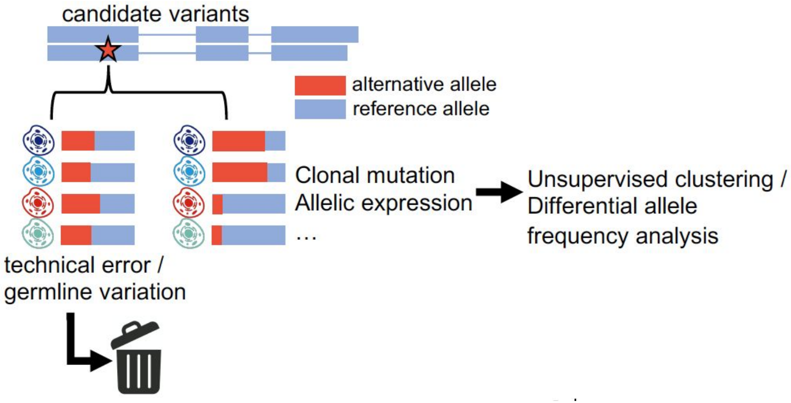
Different TSS correspond to different open chromatin regions after integration with scATAC-seq data

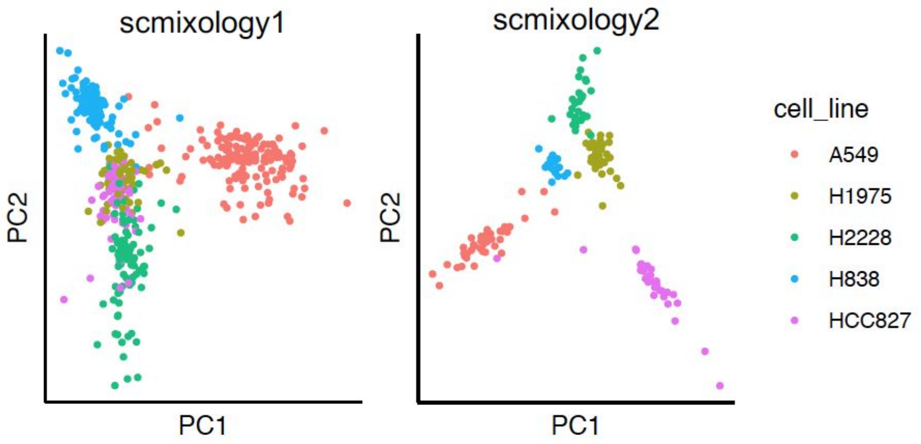


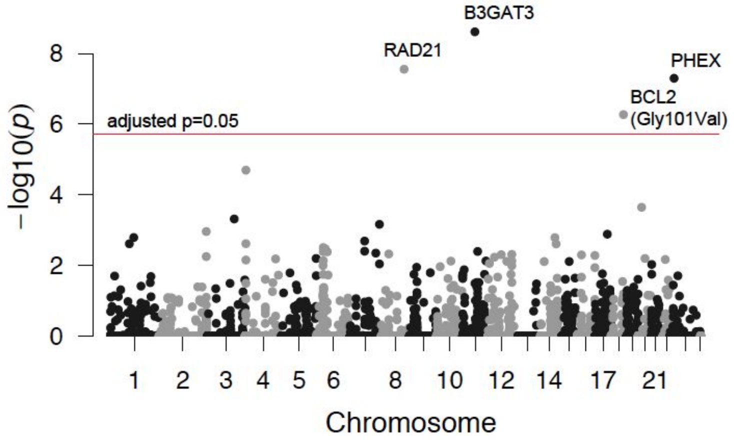
Signal aggregated per cell line in each track

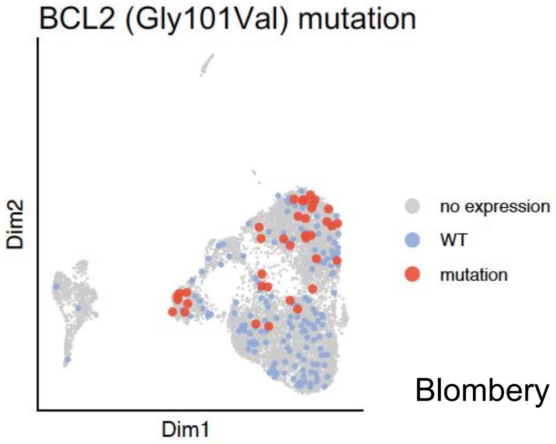
Shani Amarasinghe

FLAMES tests for differences in allele frequency between clusters









Blombery *et al.* Cancer Discovery 2019

Summary

- Developed FLT-seq that couples the popular 10X scRNA-seq protocol with ONT long-read sequencing platform
- Created new software (*FLAMES*) to detect and quantify isoforms in single cell (and bulk) RNA-seq data and also look for mutations
- Summarized and compared splicing across multiple samples from diverse cell types and tissues
- Current work: *FLAMES* -> Bioconductor
  Preprint -> bioRxiv
- Other projects underway:
  - Benchmarking of scRNA-seq preprocessing pipelines
  - Adapting *scPipe* to handle scATAC-seq data

FLT-seq method available through protocols.io:
dx.doi.org/10.17504/protocols.io.8d9hs96

***FLAMES* package available from GitHub:**

https://github.com/LuyiTian/FLAMES

Datasets submitted to GEO

(10th July 2020, awaiting accession numbers)

Acknowledgements

Luyi Tian
Shani Amarasinghe
Quentin Gouil
Has Kariyawasam
Charity Law
Xueyi Dong
Yue You
Shian Su
Kathleen Zeglinski

David Huang
Rachel Thijssen
Mary Ann Anderson
Andrew Roberts
Hongke Peng
Christoffer Flensburg

Jafar Jabbari
Alexis Lucattini
Azadeh Seidi

agrf
Empowering Australian Genomics

Mike Clark
James Ryall

THE UNIVERSITY OF MELBOURNE

FIAT LUX
Walter+Eliza Hall
Institute of Medical Research

Australian Government
National Health and Medical Research Council