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

<table>
<thead>
<tr>
<th></th>
<th>Total number of samples in data set</th>
<th>Number of markers per data set</th>
<th>Number of Markers Selected by FAUST</th>
<th>Total possible number of phenotypes</th>
<th>Total number of discovered phenotypes</th>
<th>Discovery ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set 1</td>
<td>~170</td>
<td>37</td>
<td>21</td>
<td>4,194,304</td>
<td>300</td>
<td>0.01%</td>
</tr>
<tr>
<td>Data Set 2</td>
<td>~190</td>
<td>27</td>
<td>19</td>
<td>1,769,472</td>
<td>663</td>
<td>0.04%</td>
</tr>
<tr>
<td>Data Set 3</td>
<td>~190</td>
<td>26</td>
<td>19</td>
<td>1,179,648</td>
<td>660</td>
<td>0.06%</td>
</tr>
<tr>
<td>Data Set 4</td>
<td>~190</td>
<td>22</td>
<td>18</td>
<td>884,736</td>
<td>909</td>
<td>0.10%</td>
</tr>
<tr>
<td>Data Set 5</td>
<td>~170</td>
<td>35</td>
<td>17</td>
<td>131,072</td>
<td>137</td>
<td>0.11%</td>
</tr>
<tr>
<td>Data Set 6</td>
<td>~75</td>
<td>18</td>
<td>18</td>
<td>262,144</td>
<td>275</td>
<td>0.11%</td>
</tr>
<tr>
<td>Data Set 7</td>
<td>~50</td>
<td>23</td>
<td>16</td>
<td>98,304</td>
<td>119</td>
<td>0.12%</td>
</tr>
<tr>
<td>Data Set 8</td>
<td>~190</td>
<td>21</td>
<td>17</td>
<td>131,072</td>
<td>558</td>
<td>0.43%</td>
</tr>
<tr>
<td>Data Set 9</td>
<td>~70</td>
<td>25</td>
<td>16</td>
<td>65,536</td>
<td>307</td>
<td>0.47%</td>
</tr>
<tr>
<td>Data Set 10</td>
<td>~1,160</td>
<td>16</td>
<td>13</td>
<td>8,192</td>
<td>76</td>
<td>0.93%</td>
</tr>
<tr>
<td>Data Set 11</td>
<td>~70</td>
<td>11</td>
<td>11</td>
<td>4,608</td>
<td>162</td>
<td>3.52%</td>
</tr>
<tr>
<td>Data Set 12</td>
<td>~360</td>
<td>11</td>
<td>10</td>
<td>3,456</td>
<td>165</td>
<td>4.77%</td>
</tr>
<tr>
<td>Data Set 13</td>
<td>~80</td>
<td>11</td>
<td>10</td>
<td>2,304</td>
<td>206</td>
<td>8.94%</td>
</tr>
<tr>
<td>Data Set 14</td>
<td>~30</td>
<td>11</td>
<td>9</td>
<td>512</td>
<td>82</td>
<td>16.02%</td>
</tr>
</tbody>
</table>
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
  - All subjects are treated
  - Response mediated through some cell population
    - \( P(C) = 0.5 \): prevalence
  - \( P(R|C) \) varies, biomarker is imperfect

- Cell Population

- Response

\[ \text{N=100 samples} \]
\[ 15,000 \text{ cell populations} \]
\[ \text{Aiming to be fair to all methods} \]
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

Preprint: Greene et al. BioRxiv: [https://doi.org/10.1101/702118](https://doi.org/10.1101/702118)
Code: [http://github.com/RGLab/FAUST](http://github.com/RGLab/FAUST)

Code: [http://github.com/RGLab/MAST](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](https://doi.org/10.1101/2020.03.27.012633).
SCITOSEQ: USE INFORMATION FROM MULTIPLE POOLS

- RESOLVE DROPLETS WITH DOUBLETS, MULTIPLETS, ETC

<table>
<thead>
<tr>
<th>Protein x pool id</th>
<th>Droplet id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein 1 pool 1</td>
<td>a 200</td>
</tr>
<tr>
<td>Protein 2 pool 1</td>
<td>b 500</td>
</tr>
<tr>
<td>Protein 1 pool 2</td>
<td>c 12</td>
</tr>
<tr>
<td>Protein 2 pool 2</td>
<td>d 211</td>
</tr>
<tr>
<td>Protein 1 pool 2</td>
<td>e 20</td>
</tr>
<tr>
<td></td>
<td>10 15</td>
</tr>
<tr>
<td></td>
<td>112 79</td>
</tr>
<tr>
<td></td>
<td>10 10</td>
</tr>
<tr>
<td></td>
<td>5 432</td>
</tr>
<tr>
<td></td>
<td>171 135</td>
</tr>
<tr>
<td></td>
<td>210 210</td>
</tr>
<tr>
<td></td>
<td>21 4</td>
</tr>
<tr>
<td></td>
<td>4 30</td>
</tr>
<tr>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>

Byungjin Hwang, et al. 2020
Application of FAUST to Phenotyping Multiplexed CITE-Seq Data

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

- 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
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

Human Cell Atlas
Cancer Immunotherapy Trials Network
FRED HUTCH
NIH
Characterization of full-length isoforms in single cells with Nanopore long-read sequencing and FLAMES

Matt Ritchie

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) PromethION 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
Full-Length Transcriptome sequencing (FLT-seq)

Library preparation + Nanopore sequencing

80~90% cells

Library preparation + Illumina sequencing

10~20% cells

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**

- GEM outlet
- Full-length cDNA in GEMs
- 80~90% GEMs
- 10~20% GEMs

* 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
A new toolbox called **FLAMES** (Full-Length trAnscript quantification, Mutation and Splicing analysis for long-read data) was developed.
Benchmark isoform detection and quantification using SIRV spike-in dataset*

**Isoform detection**

- **class**
  - not in reference annotation
  - In both
  - In reference annotation but not in method output

**Isoform quantification**

- **FLAMES output, n = 59**
  - $R = 0.96$, $p < 2.2e^{-16}$

- **TALON output, n = 49**
  - $R = 0.79$, $p = 1.7e^{-11}$

- **FLAIR output, n = 46**
  - $R = 0.9$, $p < 2.2e^{-16}$

- **StringTie output, n = 58**
  - $R = 0.79$, $p = 1.6e^{-13}$

* 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 specific transcript

Gene A

Cluster 1

Cluster 2

Relative abundance

Isoform 1

Isoform 2
Expression of different CD44 isoforms in *scmixology* data

not in reference annotation

variable domain in protein
Expression of different RPS24 isoforms

RPS24 (human)

RPS24 (mouse)

RPS24 gene expression

RPS24 transcript expression (CLL2)

scaled expression

high

low

CLL cells

MuSC

quiescent MuSC
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

Unsupervised clustering / Differential allele frequency analysis

BCL2 (Gly101Val) mutation

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

Mike Clark
James Ryall

Walter+Eliza Hall
Institute of Medical Research