Introduction to reproducible research and exploratory data analysis

Gabi Fragiadakis

Assistant Professor, Department of Medicine, Division of Rheumatology, ImmunoX

Director of the Data Science CoLab

BMS 225A Workshop

December 5th, 2019

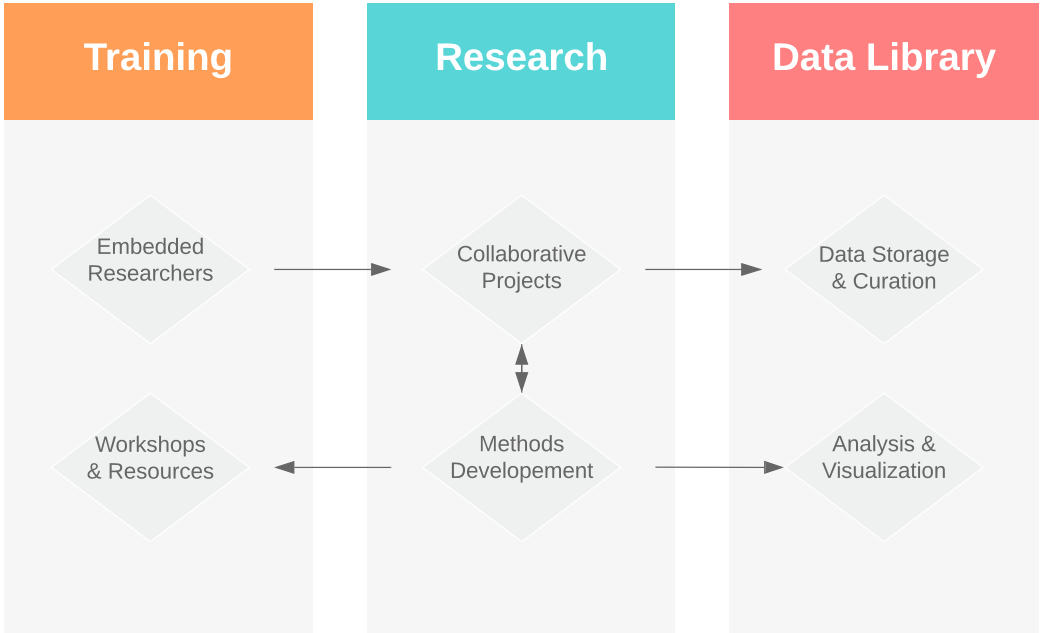
BAKAR IMMUNO

Today's Workshop

- Introduction to the DSCoLab
- Principles of reproducible research
- Version control
- Exploratory data analysis
- Resources for getting started in CyTOF and scRNAseq

The Data Science CoLab

A collaboration-based research group

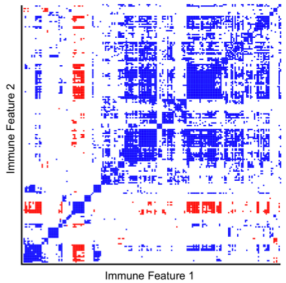


Research

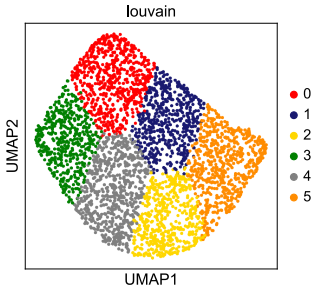
Using CoProjects and Immunoprofiler to learn larger principles of immune state across cohorts

Healthy (Ye, Spitzer)

Chronic viral infection (Baron)



Autoimmunity (Krummel)



Neurodegeneration (Huang)

Cancer (Krummel, Turnbaugh, Spitzer, Roose)

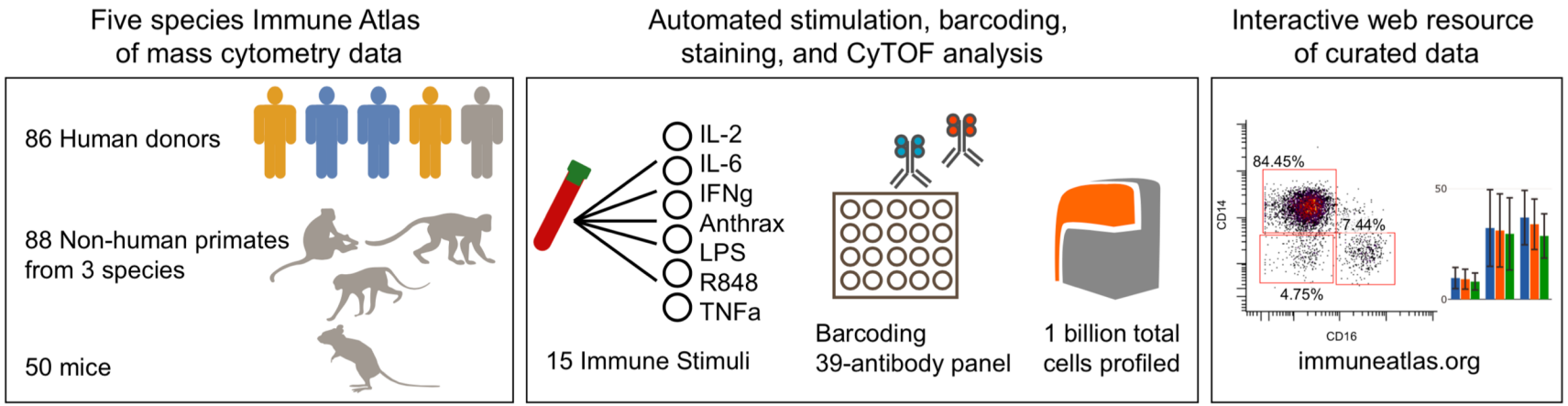
Bushra Samad

Arjun Rao

Lenny Lupin

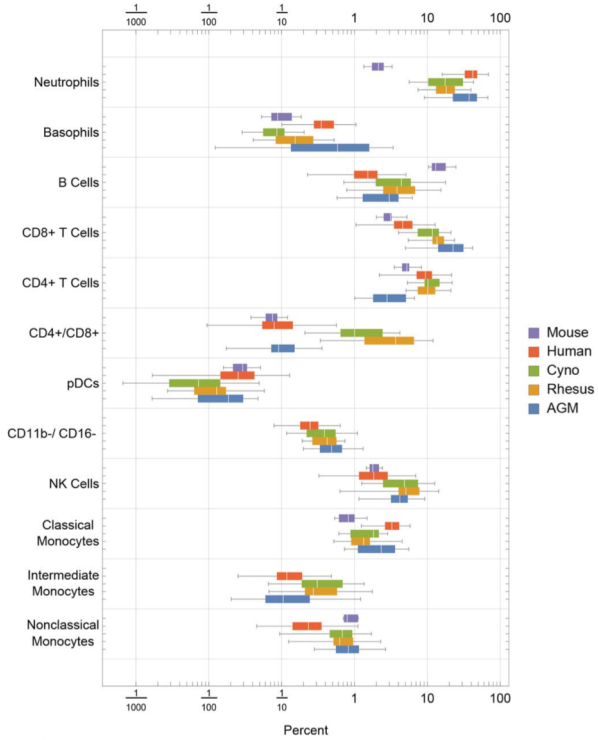
UCSF

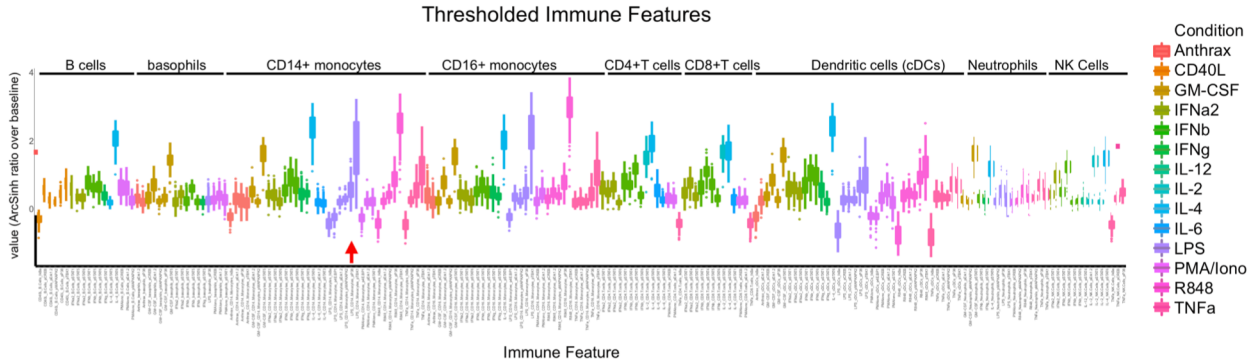
Building a cross-species immune reference map of mass cytometry data



With Zach Bjornson, Matthew Spitzer, Garry Nolan
Fragiadakis et. al, *Biorxiv*, 2019
Immuneatlas.org

Reference ranges for population abundances and immune features across species

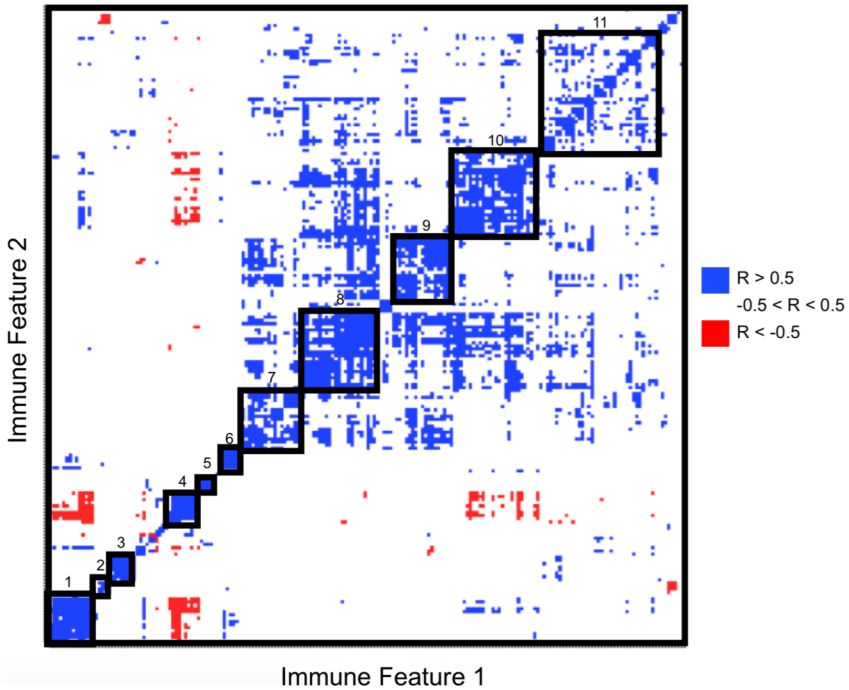




Bjornson et. al, *Biorxiv*, 2019
Fragiadakis et. al, *Biorxiv*, 2019

Inferring immune structure from data

Immune modules defined by signaling protein across cell types and stimulation

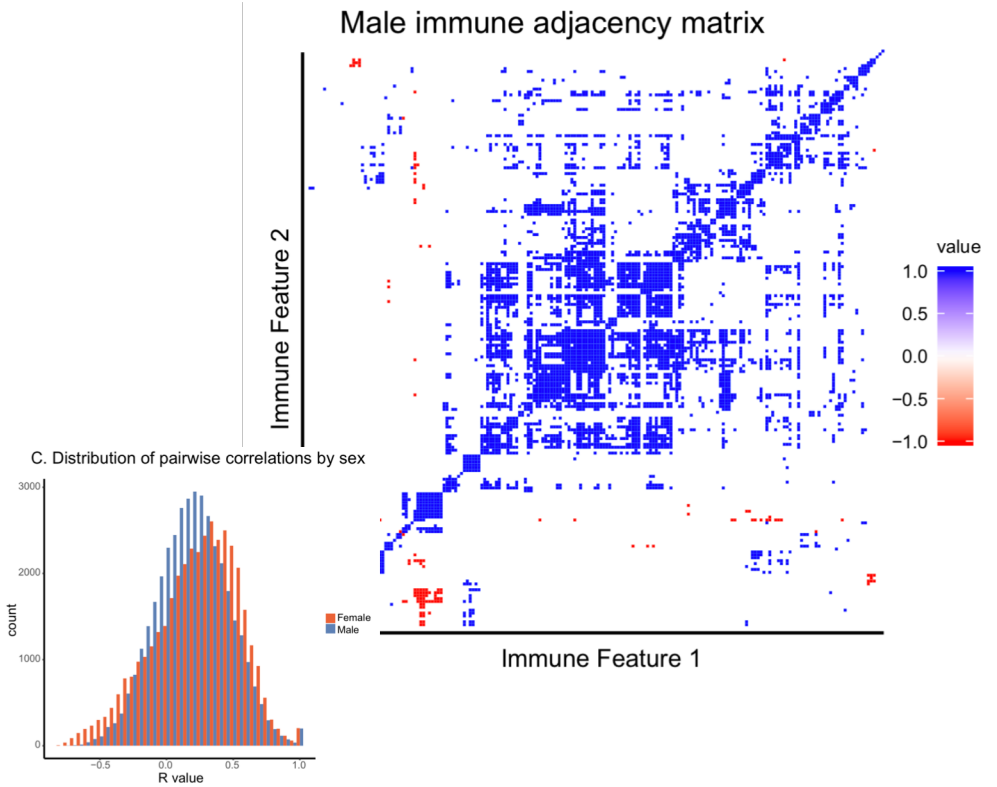


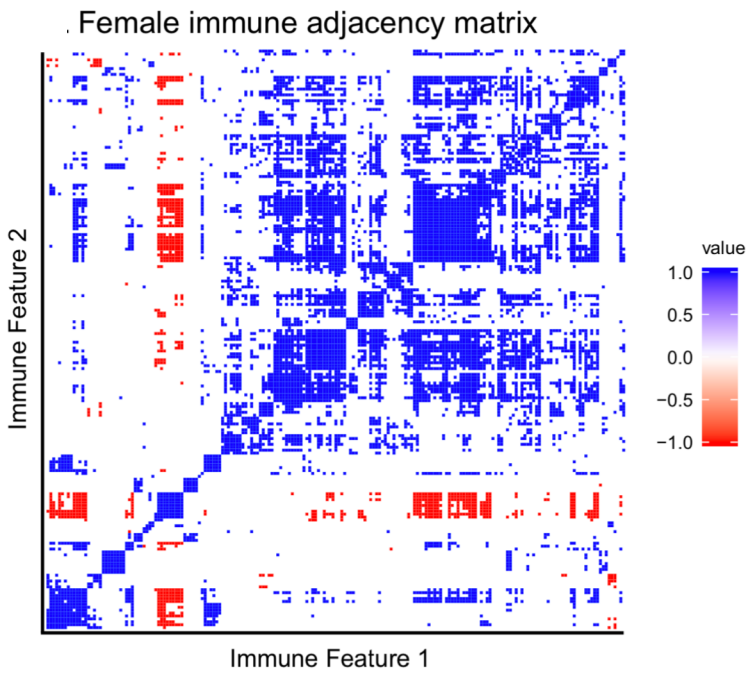
Module	Proteins	Cell types	Conditions
1	pSTAT1	CD8 T cells, CD4 T cells, DCs, NK cells, B cells	IL-6, IFNa, IFNb, IFNg
2	pSTAT1	Basophils, monocytes	PMA/iono, IFNa, IFNb
3	pERK1/2, pCREB, pMAPKAPK2	CD4 T cells, CD8 T cells, B cells	PMA/iono
4	IkB	B cells, NK cells, CD4, CD8, monocytes, DCs	CD40L, TNFa, LPS, R848
5	pP38, pERK1/2	Monocytes, basophils	Anthrax
6	pSTAT1	CD14 monocytes, CD16 monocytes	IFNa, IFNb, IFNg
7	pCREB, pP38, pERK1/2	Monocytes, neutrophils, basophils	GMCSF, IL-6, R848, LPS, PMA/iono
8	pSTAT5, pSTAT6	Monocytes, neutrophils, DCs, T cells	IL-2, GMCSF, IFNa, IFNb
9	pTBK1, pCREB, pMAPKAPK2, pP38, pERK1/2	DCs, NK cells, monocytes	TNF
10	pSTAT4, pSTAT5, pSTAT6	CD8, CD4, DCs, NK cells, neutrophils, monocytes	IFNa, IFNb, IL-4, IL-6
11	pTBK1, pCREB, pMAPKAPK2, pP38, pERK1/2	Monocytes, DCs, neutrophils	PMA/iono, LPS, R848, GMCSF

Signaling proteins:	**Conditions:**	**Proposed panel:**
p-STAT1	Unstimulated control	CD45
p-STAT5	TNFa	CD3
IkB	IFNa	CD4
p-P38	LPS	CD14
		p-STAT1
Cell types:		p-STAT5
CD4+ T cells		IkB
CD14+ Monocytes		p-P38

Fragiadakis et. al, *Biorxiv*, 2019

Immune structure differs across sex and age

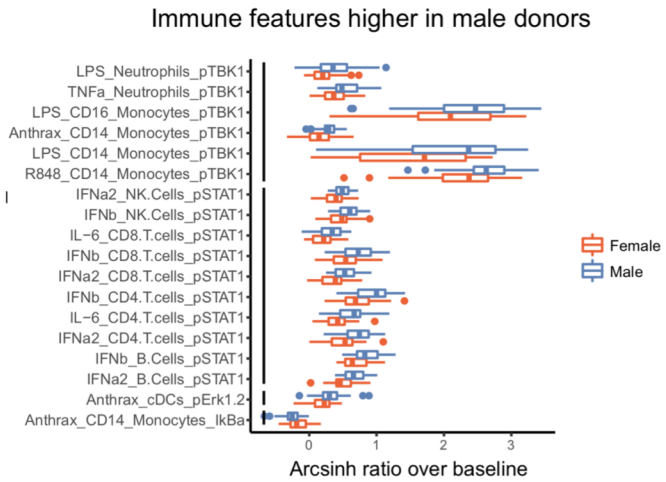


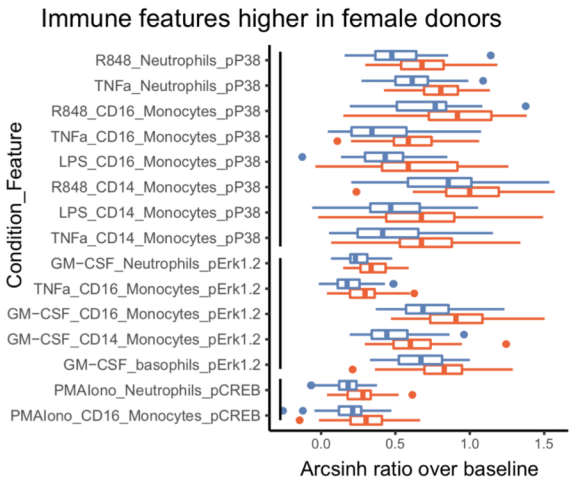


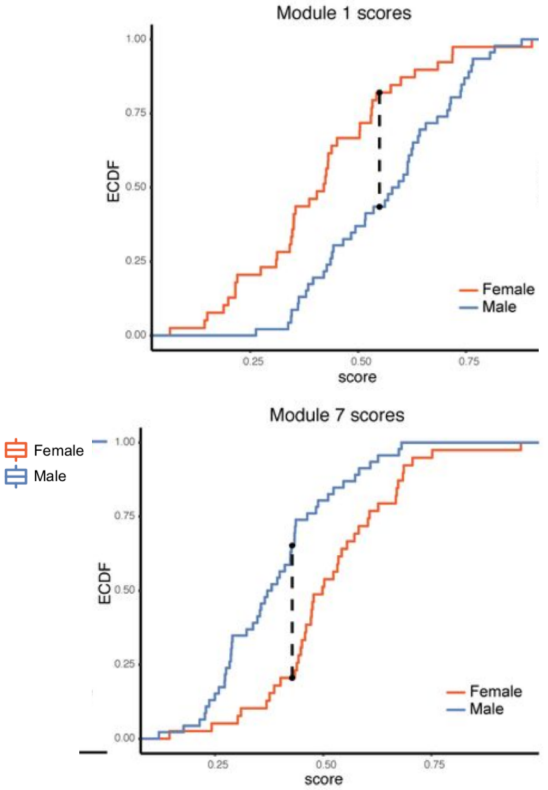
Fragiadakis et. al, *Biorxiv*, 2019

UCSF

Men and women have distinct immune response profiles







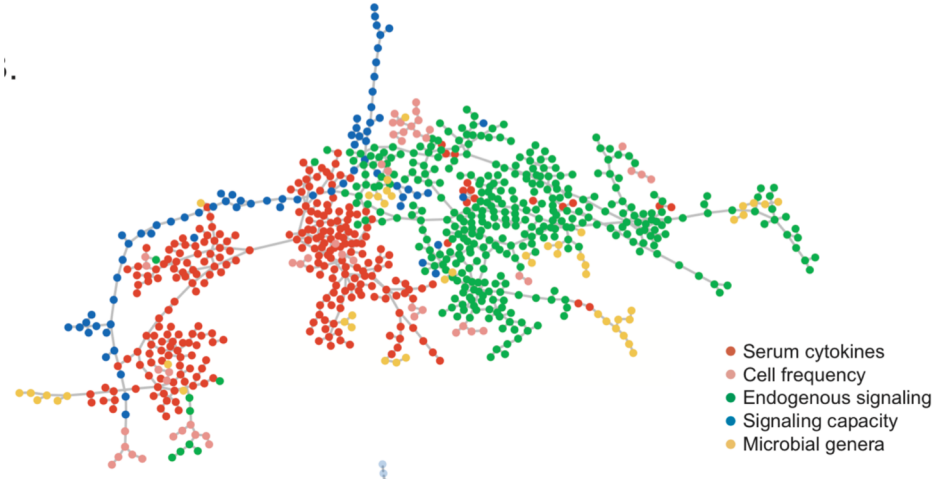
Fragiadakis et. al, *Biorxiv*, 2019

UCSF

Research

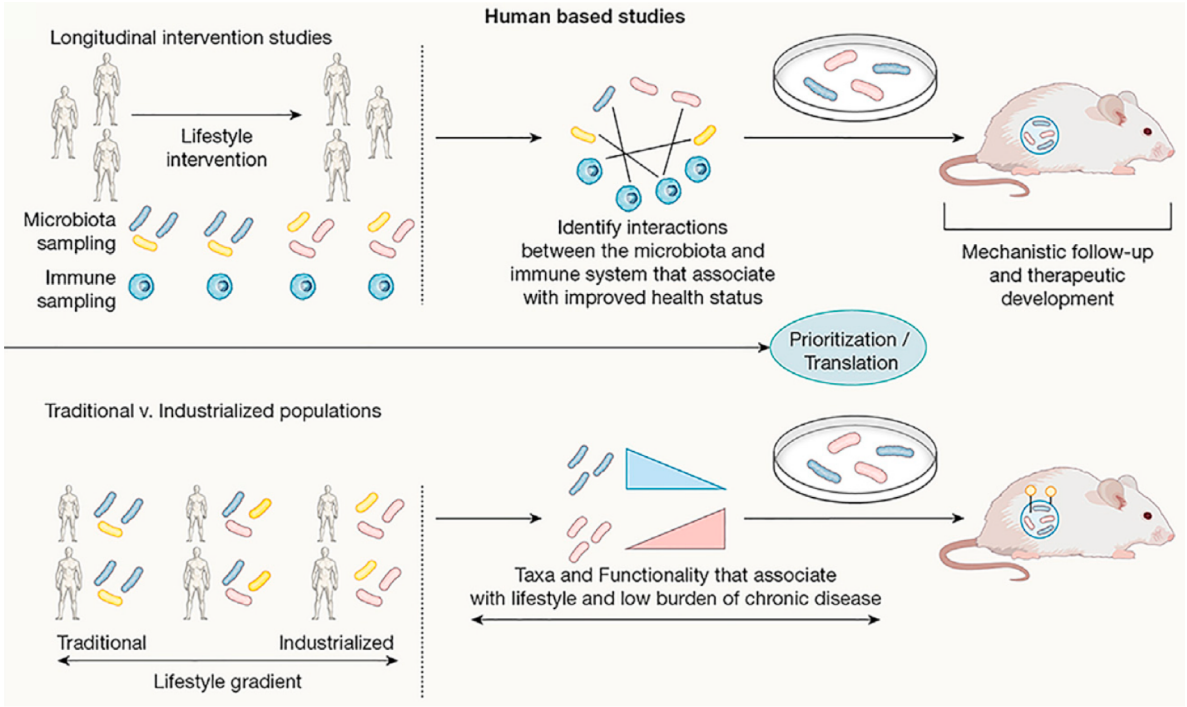
Integrating immune and microbiome states

- Combining microbiome and immune profiling data toward an understanding of organization, interactions, and fluctuations (with Spitzer lab)



UCSF

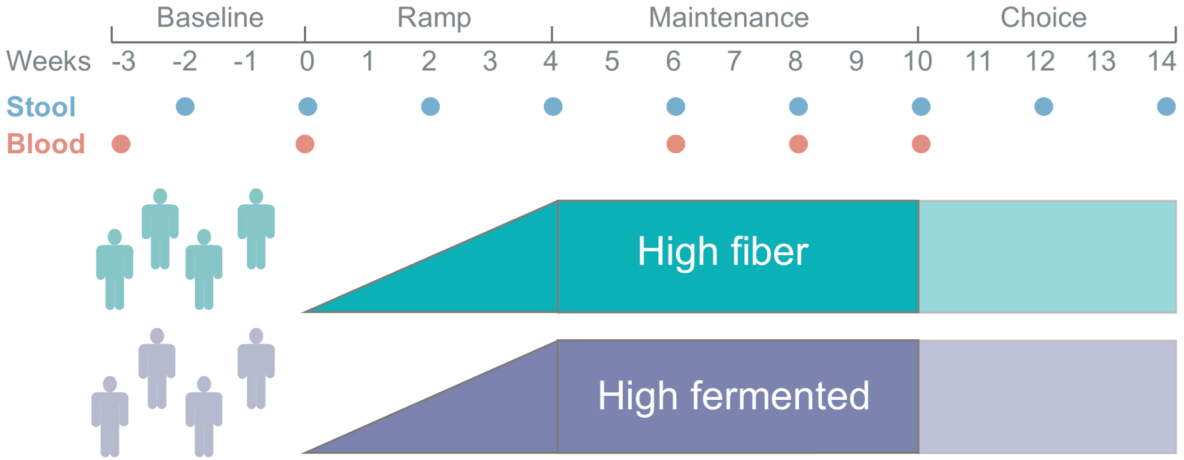
Approaches for immune and microbiome studies in humans



Spencer, Fragiadakis and Sonnenburg, *Immunity*, 2019

UCSF

Tracking the microbiome and the immune system in human dietary intervention studies



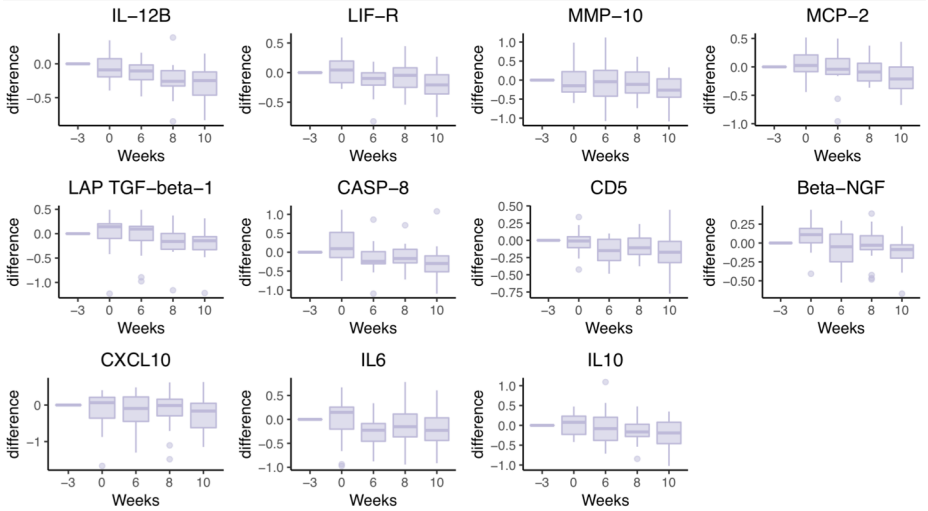
Microbiome profiling: 16S, Metagenomics, SCFA

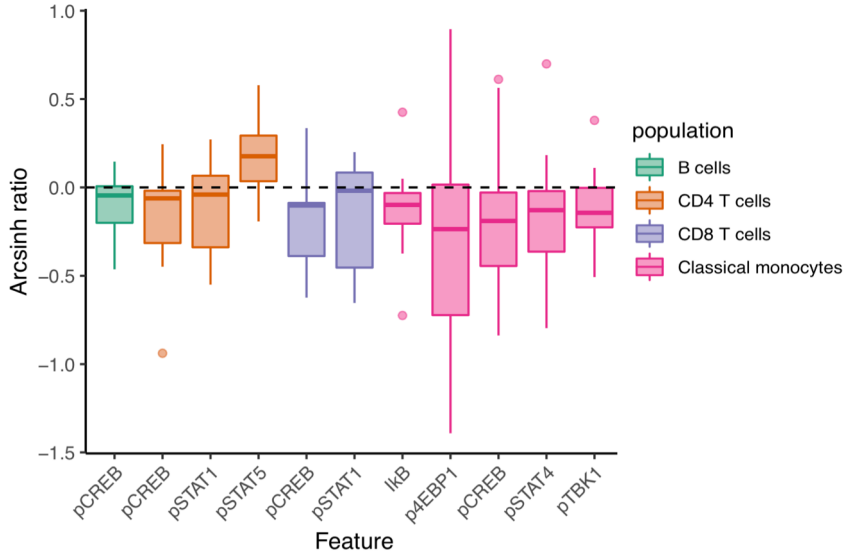
Immune profiling: CyTOF, Phosphoflow, O-link

With Hannah Wastyk, Sonnenburg Lab

UCSF

Decrease in inflammatory cytokines and signaling in participants consuming fermented foods





Citations and resources

- Cross-species immune reference map of CyTOF data with standardized panels and workflows
  - Data and interactive resource available at immuneatlas.org
  - Synchronized panel development at Bjornson et. al, Biorxiv, 2019 (**doi:** https://doi.org/10.1101/577759)
  - Primary research papers: Fragiadakis et. al, Biorxiv, 2019 (**doi:** https://doi.org/10.1101/567784) and Bjornson et. al, 2019 (**doi:** https://doi.org/10.1101/574160)
- Framework for immune and microbiome studies in humans
  - Spencer, Fragiadakis, and Sonnenburg. Pursuing human-relevant gut microbiota-immune interactions. Immunity. 2019. (**doi**: https://doi.org/10.1016/j.immuni.2019.08.002)

Collaborative projects through the CoLab structure

Data Library

The Data Library: helping biologists better engage with their data

Project UCSF Immunoprofiler

Overview | Experiments | Sequencing | Project Documents | Faq

Description

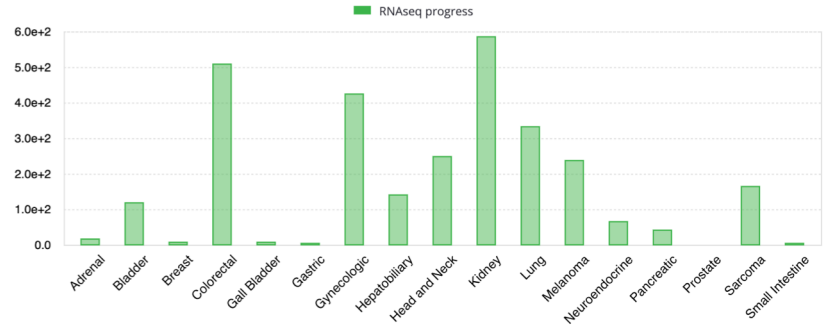
A translational platform to understand the immunological basis of cancer

What's New?

2019-05-09

We updated the Treatment table so it attaches directly to the Patient model. The old nested Regimen => Tr
adjuvant, neo-adjuvant, etc.) from the Regimen model have been moved into the Treatment table, and res
table from each Patient/Clinical tab.

RNA Progress



Quality Control

Saurabh Asthana

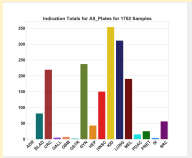
Bushra Samad

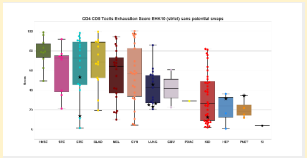
UCSF

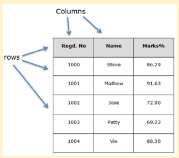
Data Library

The Data Library: helping biologists better engage with their data

Data Retrieval and QC
Sample inventory
Raw and processed data access
Subsetted data download

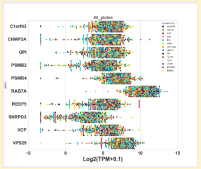

Sample totals

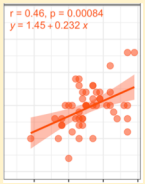

QC metrics



Exported data

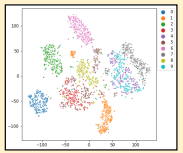
Data Analysis
Gene signatures
Multi-modal data correlation
Cross-project analysis

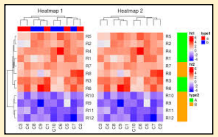

Gene signatures

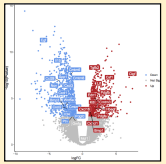


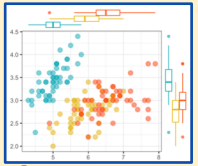
Cross-platform correlates

Data Visualization
Dimensionality reduction
Heatmaps, scatter, and boxplots
Publication-ready plots


tSNE


Heatmaps


Volcano plots


Summary plot

Training

Education, training, and outreach

- Embedded researchers model
- Workshops and trainings
  - Git, R, python, reproducible research, data type-specific analyses
- Outreach events
  - Women in Data Science, URM outreach, invited speakers/ industry partnerships

UCSF

Before we dive in….

Lowering the barrier to entry in biological data science

- Coding ~~is not harder~~ is easier than biology
- Good research is good research and similar principles apply:
  - question/hypothesis-based interrogation
  - appropriately selected and well-documented methods
  - an interest in reproducibility
- Ask ask ask (me, your lab mate, google/stack overflow), try try try (tutorials, your data)

Principles of Reproducible Research

Adapted from Karl Broman's course on Tools for Reproducible Research

- What is reproducible research in the land of biological data science?

A set of practices involving the process and structuring data and code that minimize error, maximize efficiency, and enable reproducibility by the community.

- Why bother?

Ultimately saves time ("Hey, I tweaked the alignment slightly…")

Facilitates collaboration

Minimizes error

"Your worst collaborator is you six months ago, and you don't respond to emails".

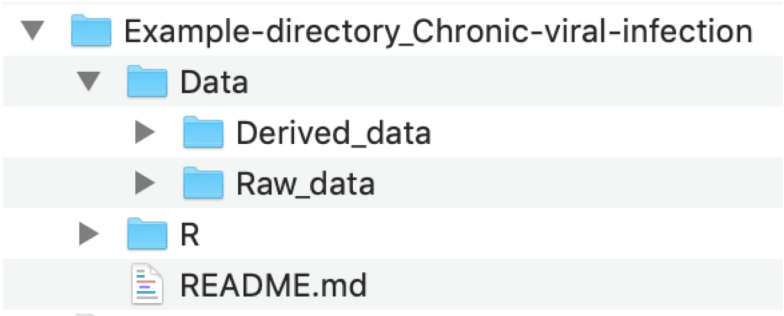
Principles of Reproducible Research

1. Organize your data and code
2. Everything with a script
3. Use version control
4. Turn repeated code into functions (and other good coding practices)
5. Turn scripts into reproducible reports
6. Package functions for future use

Organize your data and code

- Have a single directory per project
- Separate raw from derived data
- Separate the data from the code
- Use relative paths in your code
- Choose file names carefully (and don't use "final")
- Write README files

Sample structure:



Everything with a script

- Get the data in most raw form possible
- Don’t hand-edit data files (changing column names, deleting headers, etc)
- All aspects of data cleaning should be in scripts
- Each aspect of the analysis should be in a script
- Save your seeds for random number generation

Use version control

Your lab notebook for code that will save you from yourself

- (will do together in hands-on)

Turn repeated code into functions

And other good coding practices

- Turn repeated code into functions, and write both smaller, modular functions and larger more generalized functions (that likely use those functions)
- Use variables as much as possible rather than hardcoding in values
- Use a “style guide” for your code (spacing, variable syntax, etc, e.g. my_variable_syntax, or myVariableSyntax);
- Naming:
  - use clear, meaningful names
  - variables are often nouns (single_cell_matrix), functions often verbs (plot_cells())

Formatting matters

```
# version 1

ranks_of_interest <- c("Phylum", "Class", "Order", "Family"
summary_list <- list()
for (rank in ranks_of_interest){
            summary_mat <- summarize_taxa(diet_ps, Rank = r
  summary_df <- data.frame(SampleID = diet_ps@sam_data$Samp
  summary_list[[rank]] <- summary_df}
taxa_summaries <- Reduce(left_join, summary_list)
formatted_metadata <- select(diet_ps@sam_data, SampleID, Gr
dplyr::rename(Timepoint = CollectionTime) %>%
  dplyr::rename(Participant = old_record_id)
taxa_summaries <- left_join(formatted_metadata, taxa_summar
taxa_summaries$Timepoint <- as.numeric(taxa_summaries$Timep
```

```
## version 2

ranks_of_interest <- c("Phylum", "Class", "Order", "Family", "Genus")

summary_list <- list()

for (rank in ranks_of_interest){

  summary_mat <- summarize_taxa(diet_ps, Rank = rank, normalize = TRUE)

  summary_df <- data.frame(SampleID = diet_ps@sam_data$SampleID, summary_mat) %>%
                           plyr::rename(c("Unassigned" = paste(rank, "_Unassigned")))

  summary_list[[rank]] <- summary_df

}

taxa_summaries <- Reduce(left_join, summary_list)

formatted_metadata <- select(diet_ps@sam_data, SampleID, Group, old_record_id, CollectionTime) %>%
                      dplyr::rename(Timepoint = CollectionTime) %>%
                      dplyr::rename(Participant = old_record_id)

taxa_summaries <- left_join(formatted_metadata, taxa_summaries)

taxa_summaries$Timepoint <- as.numeric(taxa_summaries$Timepoint)
```


UCSF

Turn scripts into reproducible reports

Provides context for why you did what you did

```
---
title: "CyTOF pre-processing"
author: "GK Fragiadakis"
date: "10/7/2019"
output: html_document
---

Example runthrough of CyTOF data pre-processing. Reference: https://github.com/ParkerICI/premessa

This demo includes:

- Channel renaming
- File concatenation
- Normalization
- Debarcoding

Then upload for gating, additional analyses.

Start:

```{r}

library(premessa)
demo_path <- "~/Documents/CyTOF/demo fcs files/"

```
```

CyTOF pre-processing

GK Fragiadakis

10/7/2019

Example runthrough of CyTOF data pre-processing. Reference: https://github.com/ParkerICI/premessa

This demo includes:

- Channel renaming
- File concatenation
- Normalization
- Debarcoding

Then upload for gating, additional analyses.

Start:

```
library(premessa)
demo_path <- "~/Documents/CyTOF/demo fcs files/"
```

Rmarkdown + knitr

UCSF

Package functions for future use

- Resources for writing your own R package:

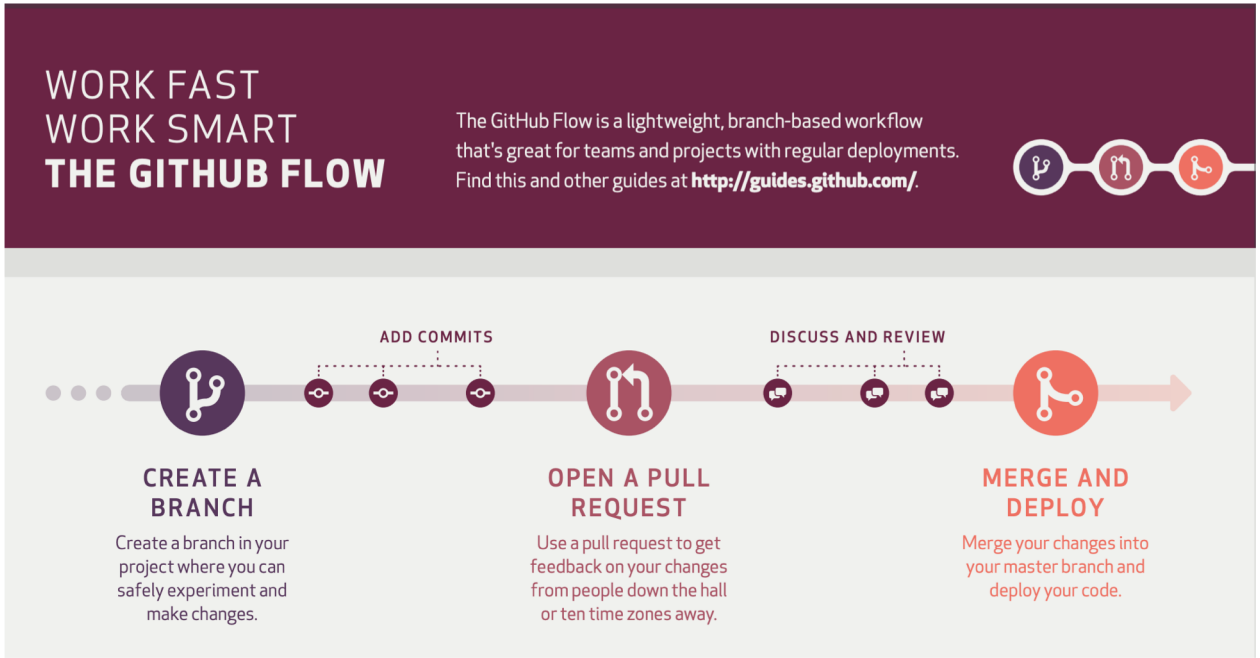
- Simplest (maybe slightly oversimplified): https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/

- Fairly simple (good place to start): https://kbroman.org/pkg_primer/

- A whole book on building R packages (very useful reference): http://r-pkgs.had.co.nz/

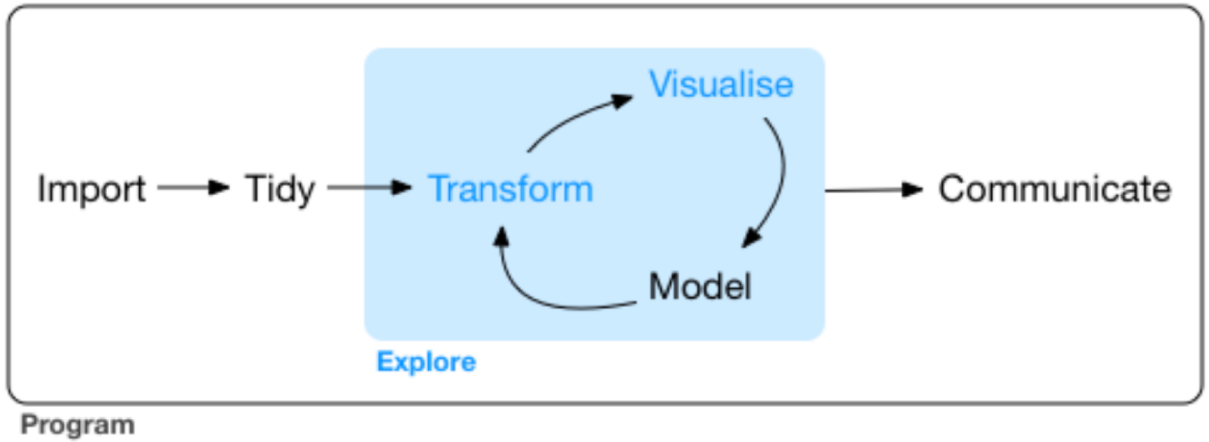
Version control (hands-on demo)

Repo at https://github.com/UCSF-DSCOLAB/BMS-225A

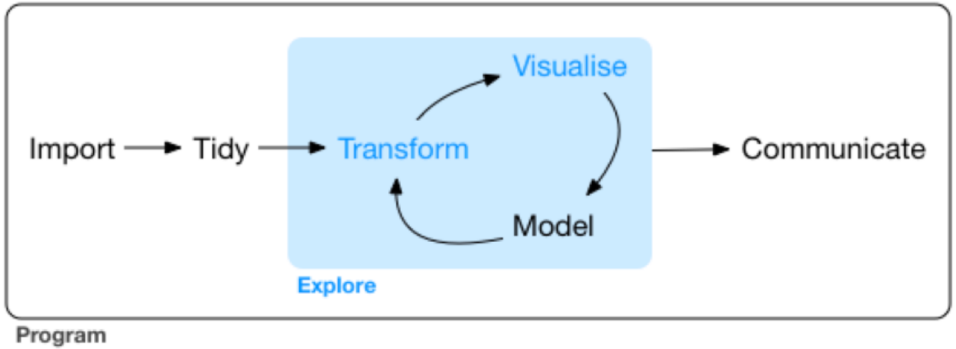


Exploratory Data Analysis

Iterative, question-based

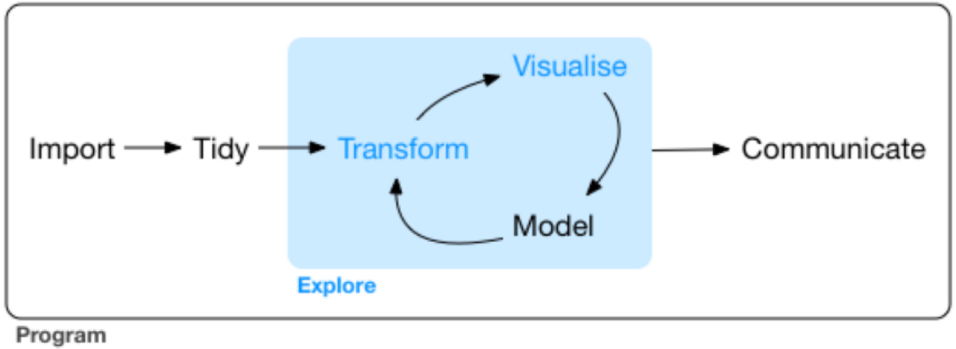


EDA pipeline: CyTOF data



- Tidying the data:
  - Pre-processing the data (normalize, debarcode, file cleanup): see the premessa package
  - Gate the data and export population values (FlowJo, Cytobank, CellEngine)
  - Combine this data matrix with sample metadata
- Exploring the data
  - What questions would you ask?

EDA pipeline: scRNAseq data



- Tidying the data:
  - Alignment
  - Filtering out cells with low gene content, dying cells, etc

- Exploring the data:
  - clustering, dimensionality reduction, differential expression
  - (see example)

Exploratory and confirmatory split

Raising the bar for confirmatory claims

One way to do this:

- Use 60% of your data for exploratory data analysis and training (can do as much modeling and exploring as you want)
- Use 20% for a query set (comparing models etc)
- Use 20% as your test set (you can only use this data ONCE)

Try it at home, come get help

- DSCoLab office hours: Tuesdays 2-4pm in S-447 at Parnassus
- Check out our website: dscolab.ucsf.edu
- Email our group at dscolab@ucsf.edu
- Email me at Gabriela.Fragiadakis@ucsf.edu

DSCoLab

Dscolab.ucsf.edu

BAKAR IMMUNO

COLABS

DSCoLab:
Saurabh Asthana
Lenny Lupin-Jimenez
Arjun Rao
Bushra Samad

UCSF:
Max Krummel
Vincent Chan
Matt Spitzer
Jody Baron
Jimmie Ye
Peter Turnbaugh
Jeroen Roose
Eric Huang

Stanford:
Sonnenburg Lab
Hannah Wastyk
Gardner Group
Nolan Lab
Zach Bjornson

Get in touch!
dscolab@ucsf.edu
Gabriela.Fragiadakis@ucsf.edu