Using Dirichlet Gaussian Processes to Analyze Gene Expression of Cancer Metastasis Progression

Perla Molina

BEEHIVE Lab

Autumn 2023

Introduction

It’s me!

Perla Molina

- First Year PhD in DBDS
- Bachelor’s in Data Science at USF
  - DaVita Internship
  - AWM President
- Why Stanford?
  - Easy move
  - Meaningful research
  - Data science realm
- Research interests
  - Cancer and disease
  - gynecology/women’s health
- Obsessed with kpop and horror movies

UNIVERSITY OF SAN FRANCISCO

Background Info + Material

The Biological Problem

Cancer metastasis is the cause of death for **50-90%** → no current therapies to specifically target metastasis [2] → the need to look at metastatic data

Understanding genetic dynamics of cancer metastasis remains incomplete [2] → lots of undiscovered territories, especially with progression over time

Previous computational analysis reveals “an ordered series of immunological changes that correspond to metastatic progression” [2] → importance of looking at differential changes at molecular and genetic levels → potential for target-based therapies

What is DP_GP?

- Bayesian nonparametric model for time series trajectories [1]
  - P is the number of genes
  - T the number of time points per sample, assuming observations at the same time points across samples, but allowing for missing observations (missing data)
- Bayesian part → probabilistic framework that can analyze uncertainty
- “**DP clusters the trajectories** of gene expression levels across time, where the **trajectories are modeled using a Gaussian process**.” [1]

$$Y \in \mathbb{R}^{P \times T}$$

$$G \sim DP(\alpha, G_0);$$

$$\theta_h \sim G;$$

$$y_j \sim p(\cdot|\theta_h).$$

Why DP_GP?

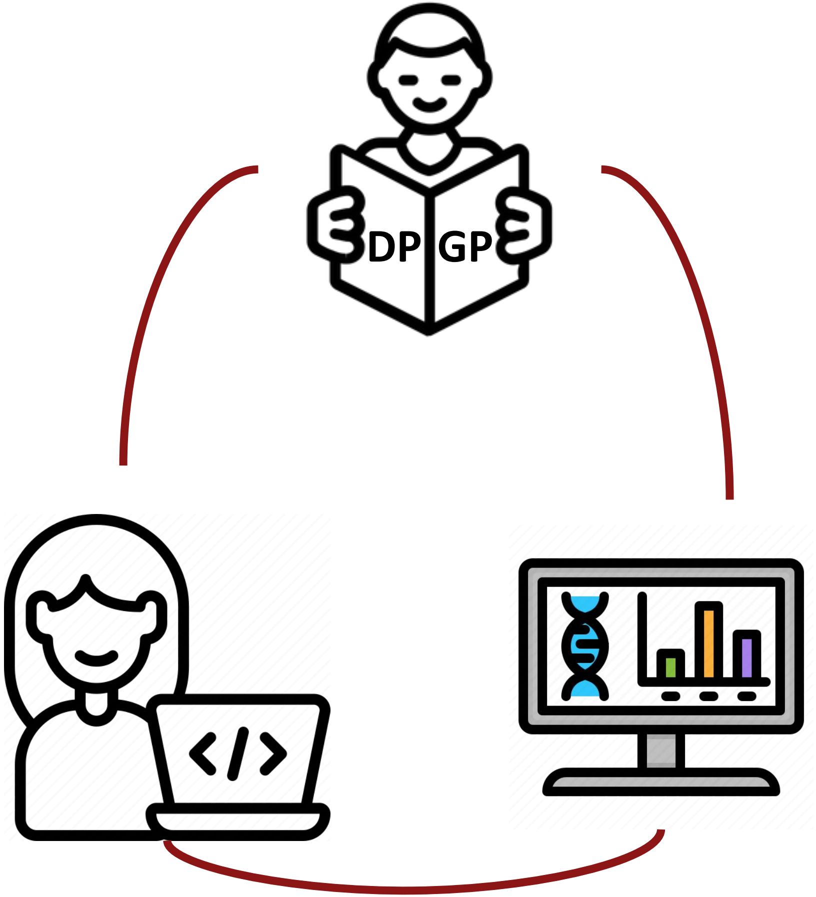
Benefit: Do not have to assume the given number of clusters at beginning, a priori [1] (other methods mostly do) → huge benefit for analyzing differential growth over time

Benefit: Does not assume independence of clusters (like k-Means, hierarchical clustering, etc) → important in clustering gene expression over period of time

Objectives

My Task

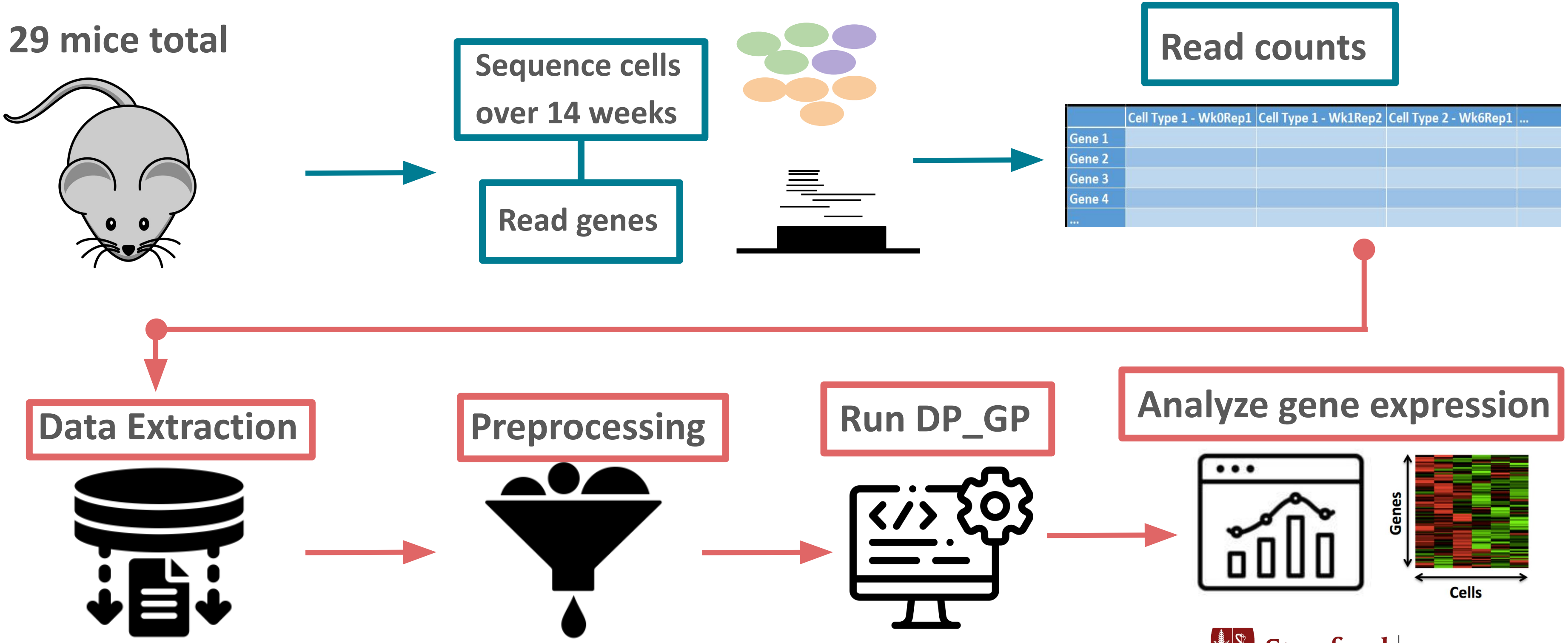
- Learn how DP_GP works
- Update and install DP_GP software
- Extract, preprocess, and format data
  - Look at top 3 and lowest 2 frequent cell types
- Analyze gene expression of metastatic lung cancer in mice over time using DP_GP



Methodology

Study Design

Wet lab
Me



29 mice total

Sequence cells over 14 weeks

Read genes

Read counts

	Cell Type 1 - Wk0Rep1	Cell Type 1 - Wk1Rep2	Cell Type 2 - Wk6Rep1	...
Gene 1				
Gene 2				
Gene 3				
Gene 4				
...				

Data Extraction

Preprocessing

Run DP_GP

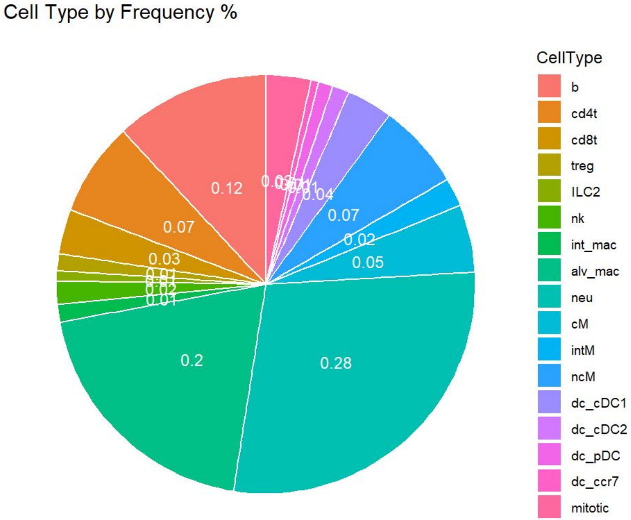
Analyze gene expression

What I Used

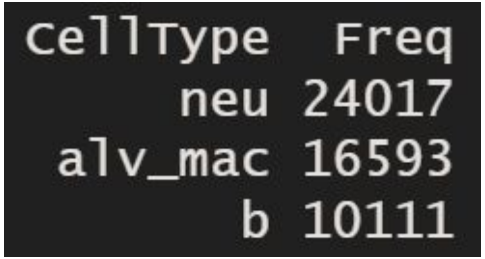
- R
    - Extract data for each cell type
    - Format by individual timepoints
        - Sum counts of each replicate & timepoint
    - Preprocess & select significant genes
        - CPM value threshold >= 10 [3]
        - Log2 fold change threshold >= 4 and adjusted Wilcox p-value < 0.01 [4][5]
            - Bonferroni correction
    - Normalize significant genes
        - Average the sums of replicates for each time point
        - Z-score normalization
- Python
    - Fix & update outdated code
    - Install updated DP_GP
    - Run DP_GP on final output data from R (45 iterations per cell type dataset)

Results

Top & Lowest Frequent Cell Types

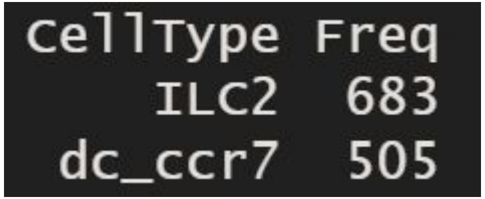


Top 3



```
CellType  Freq
     neu 24017
 alv_mac 16593
       b 10111
```

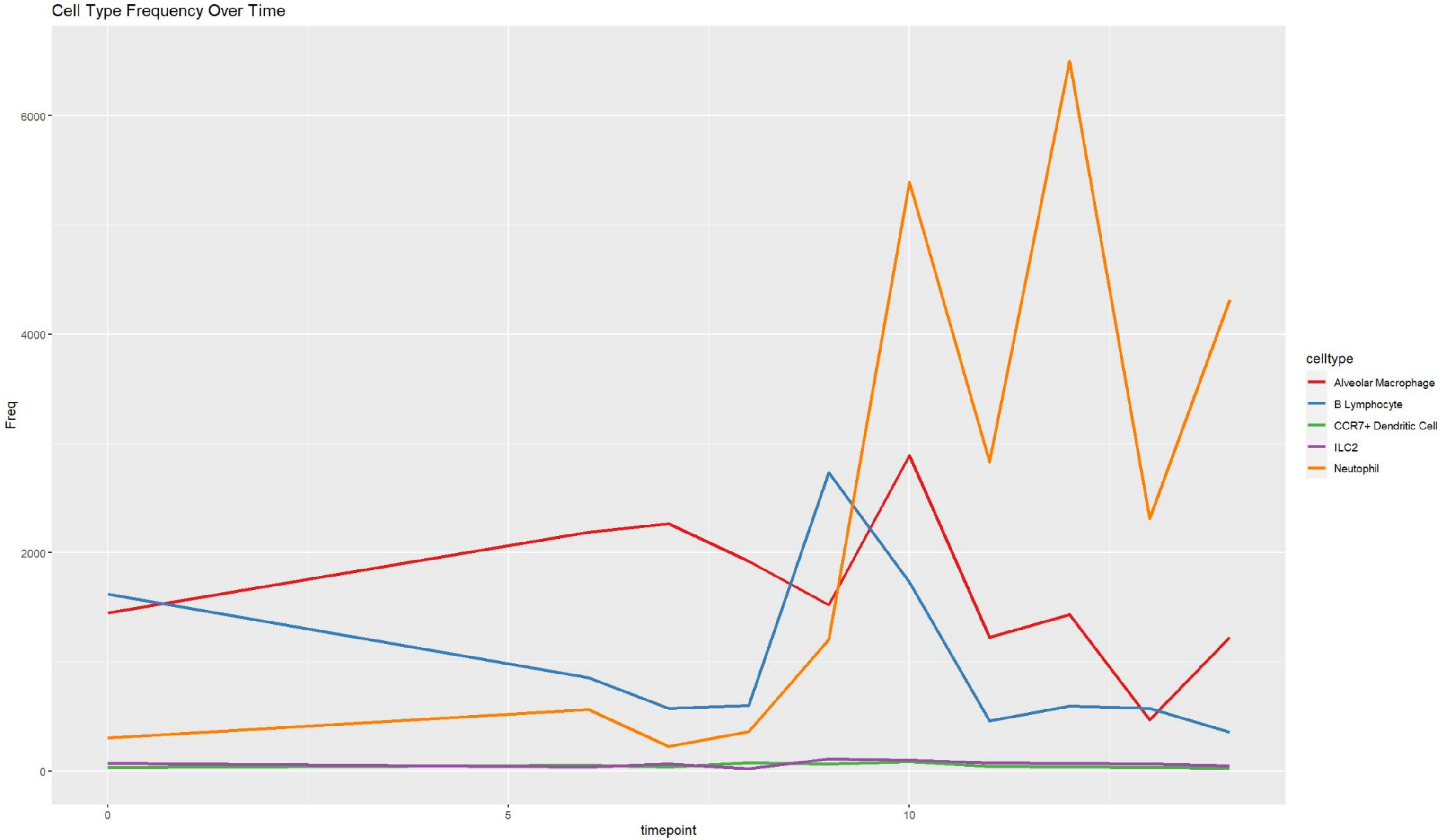
Lowest 2



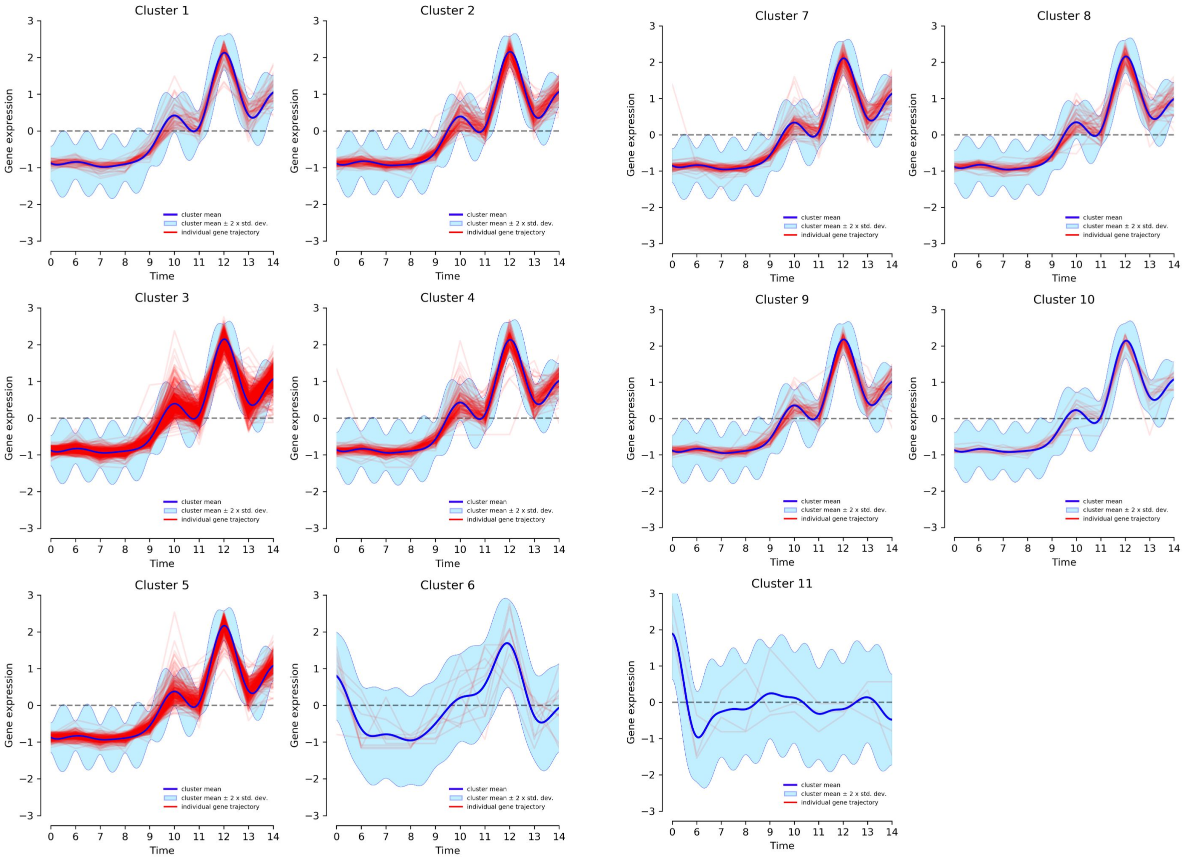
```
CellType Freq
    ILC2  683
  dc_ccr7  505
```

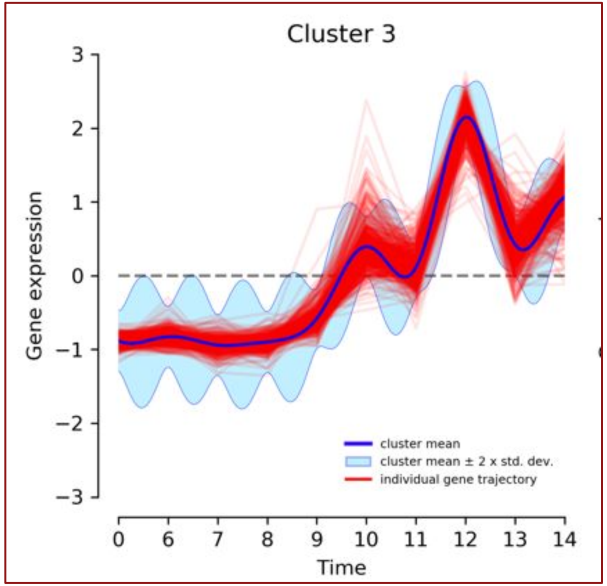
- neu = Neutrophil cells
- alv_mac = Alveolar macrophages
- b = B lymphocyte cells
- ILC2 = Type 2 Innate Lymphoid Cells
- dc_ccr7 = CCR7+ Dendritic cells

Top & Lowest Frequent Cell Types Over Time

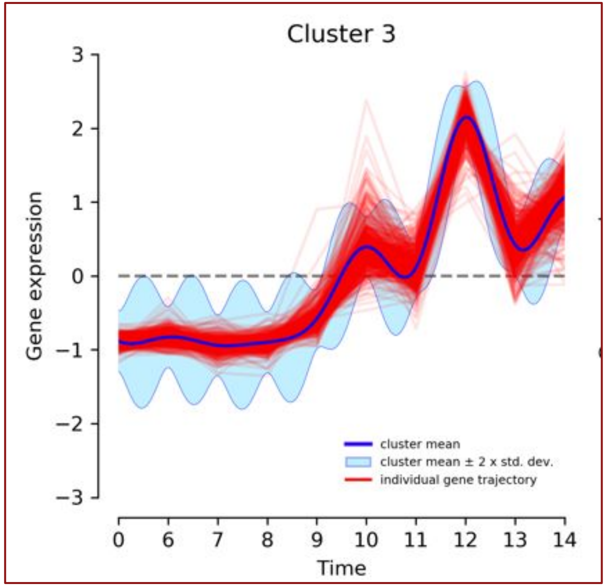


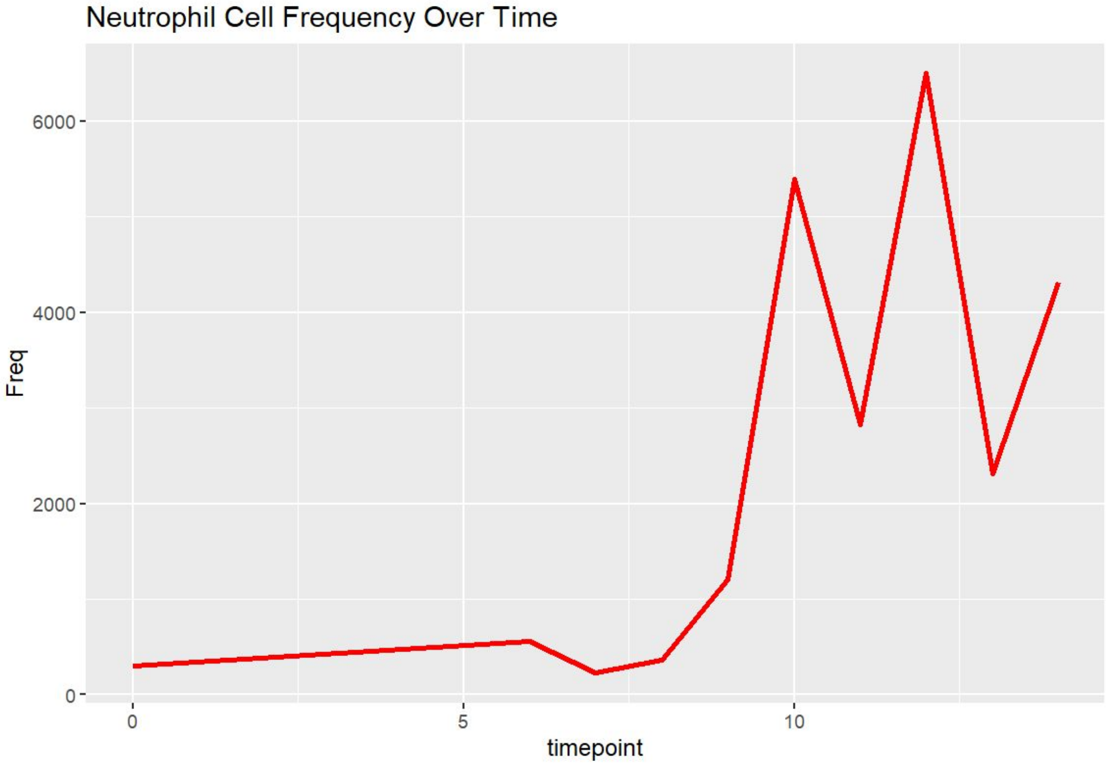
DP_GP Gene Expression for Neutrophil Cells



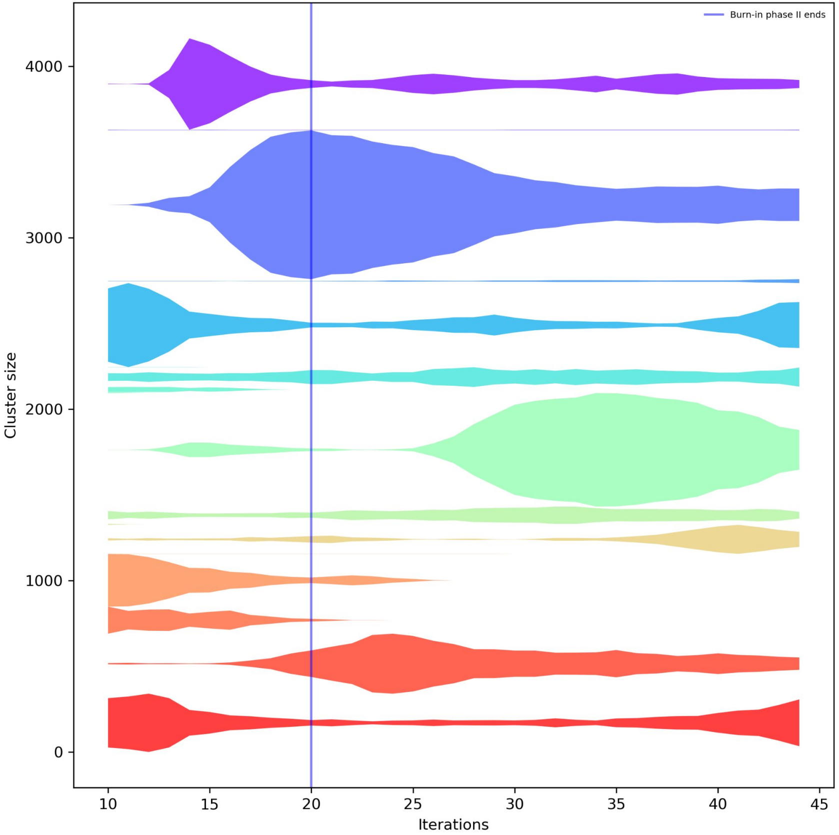


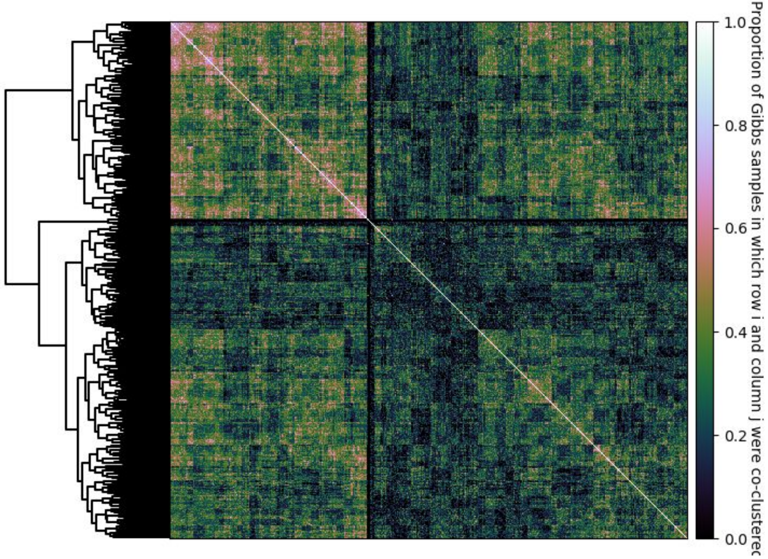
Closer Look at Neutrophil Cells Over Time



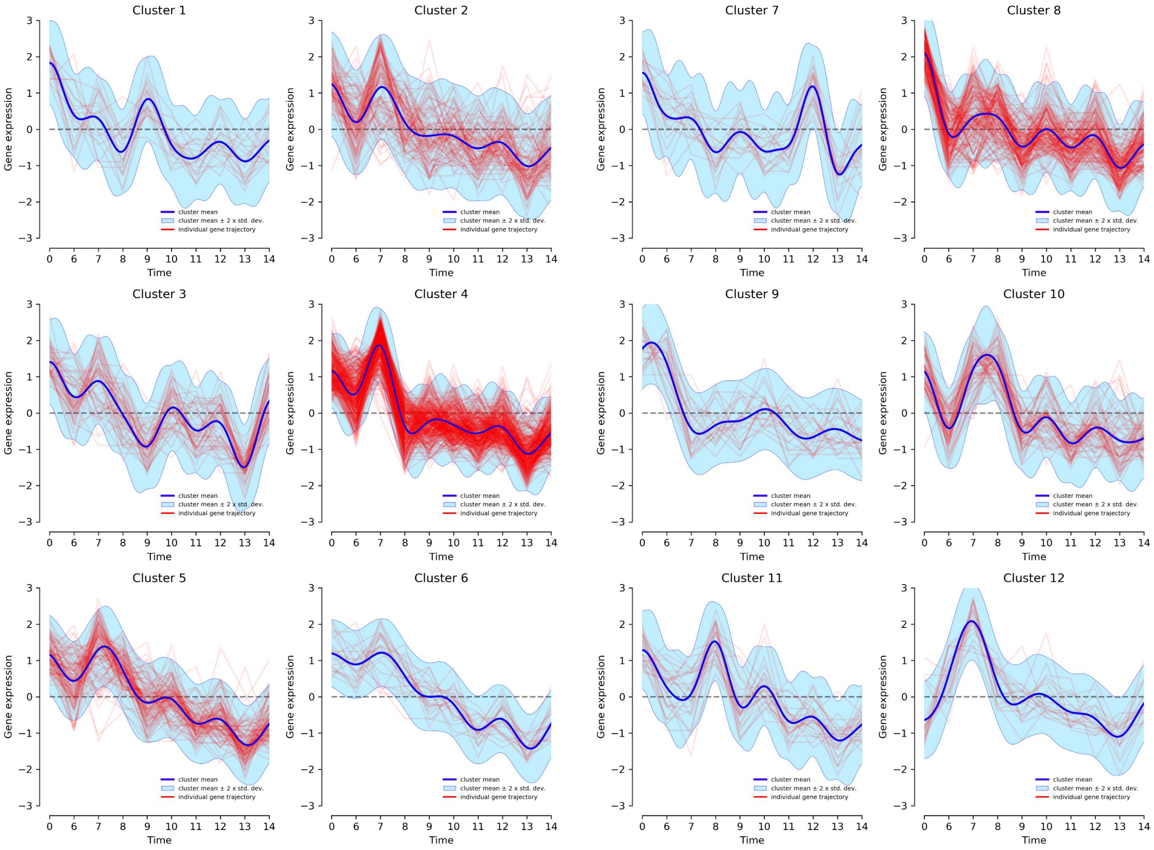


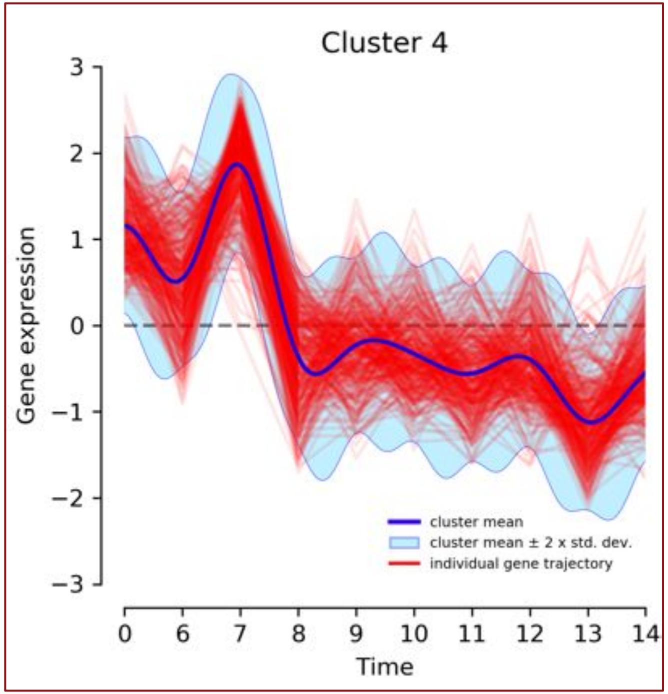
DP_GP Iteration Results for Neutrophil Cells



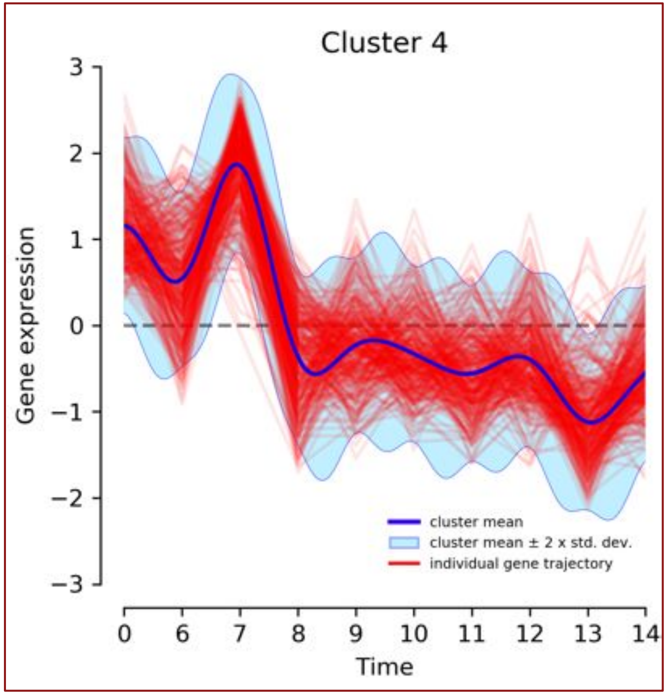


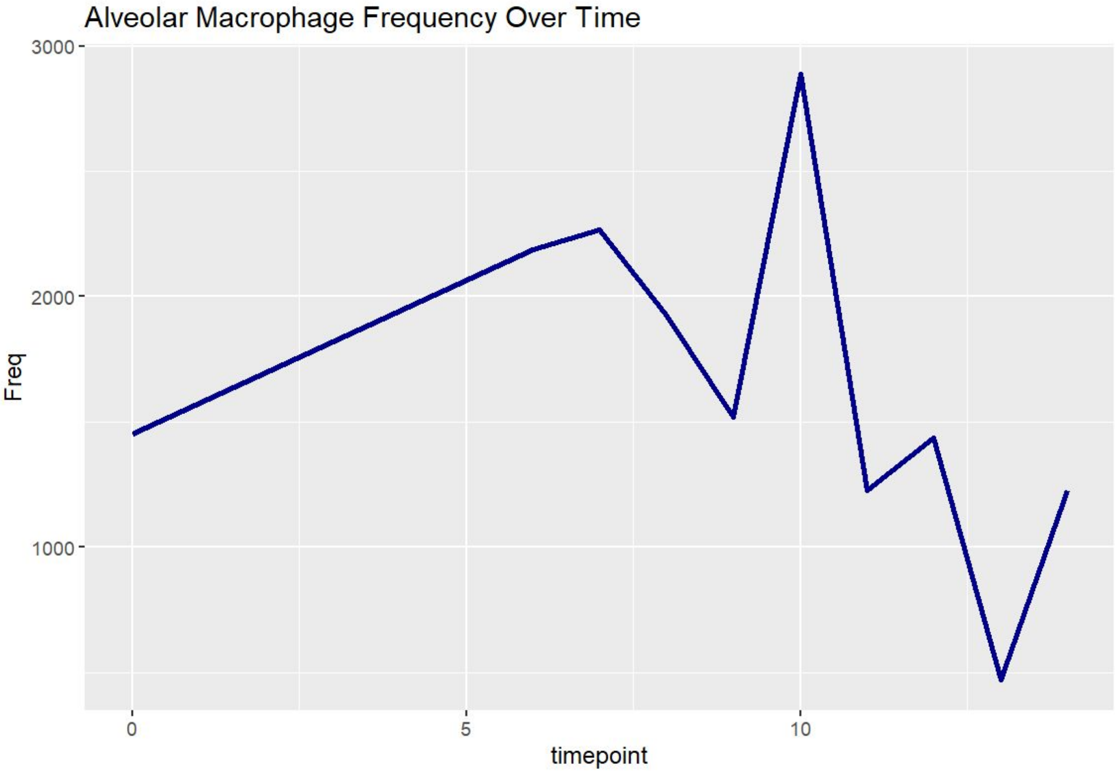
DP_GP Gene Expression for Alveolar Macrophages



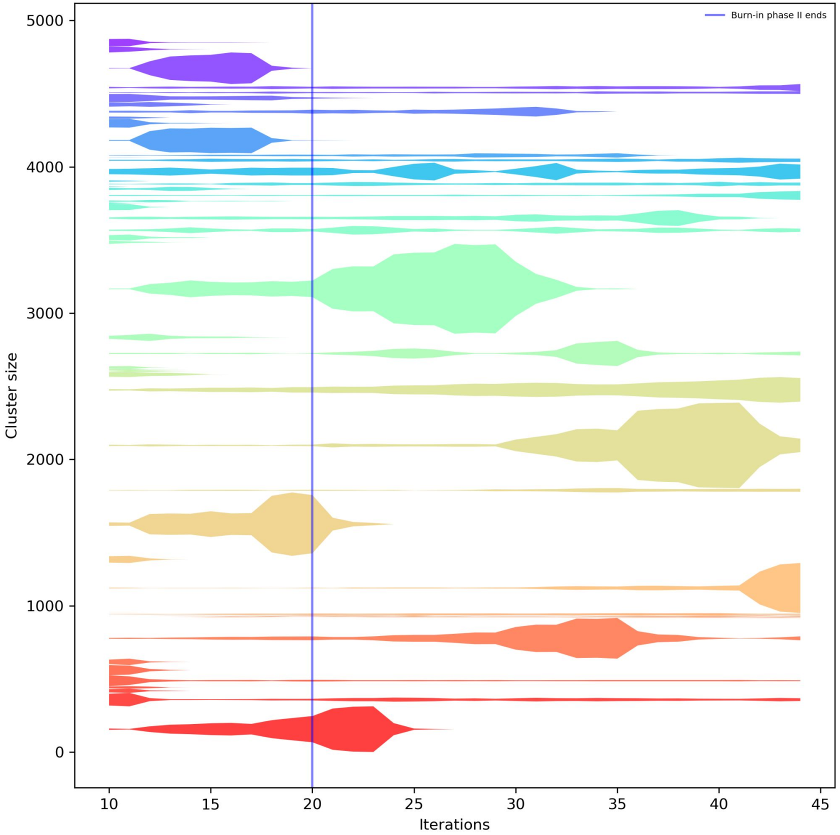


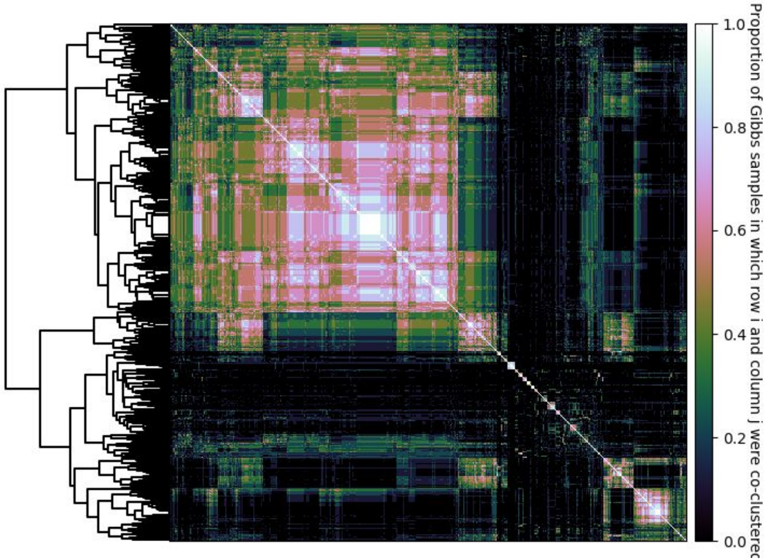
Closer Look at Alveolar Macrophages Over Time



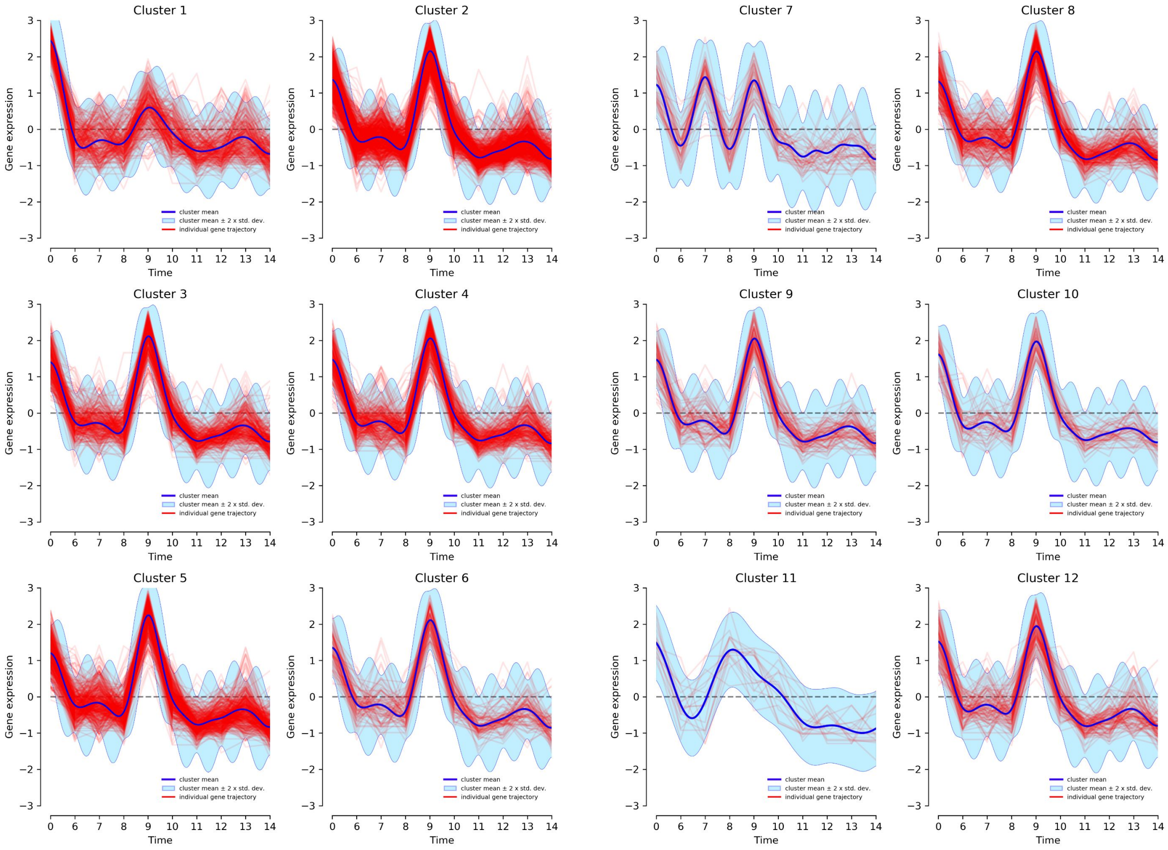


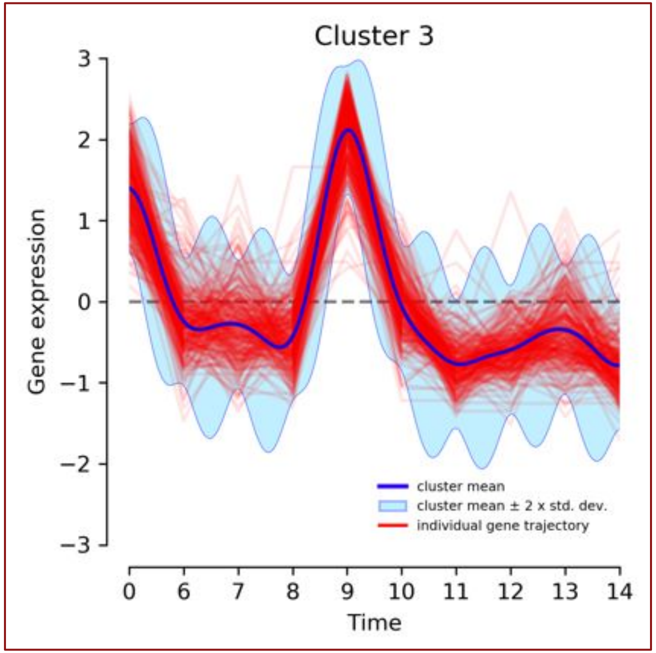
DP_GP Iteration Results for Alveolar Macrophages





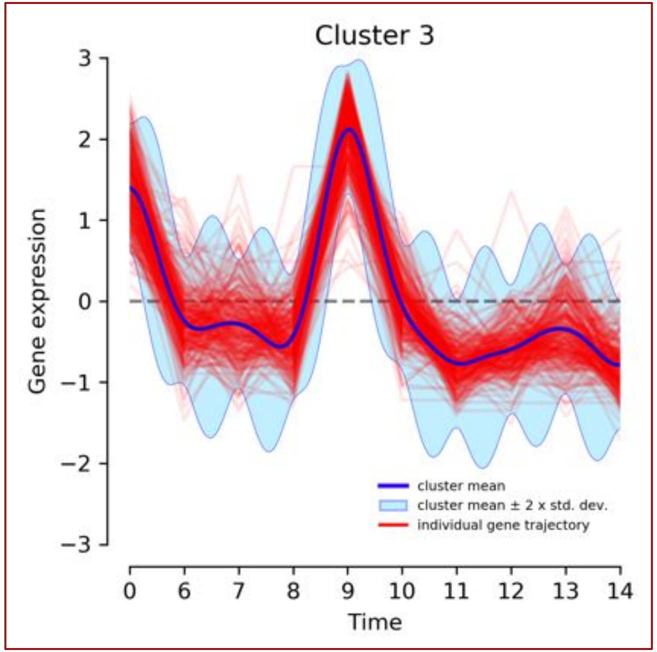
DP_GP Gene Expression for B Lymphocytes

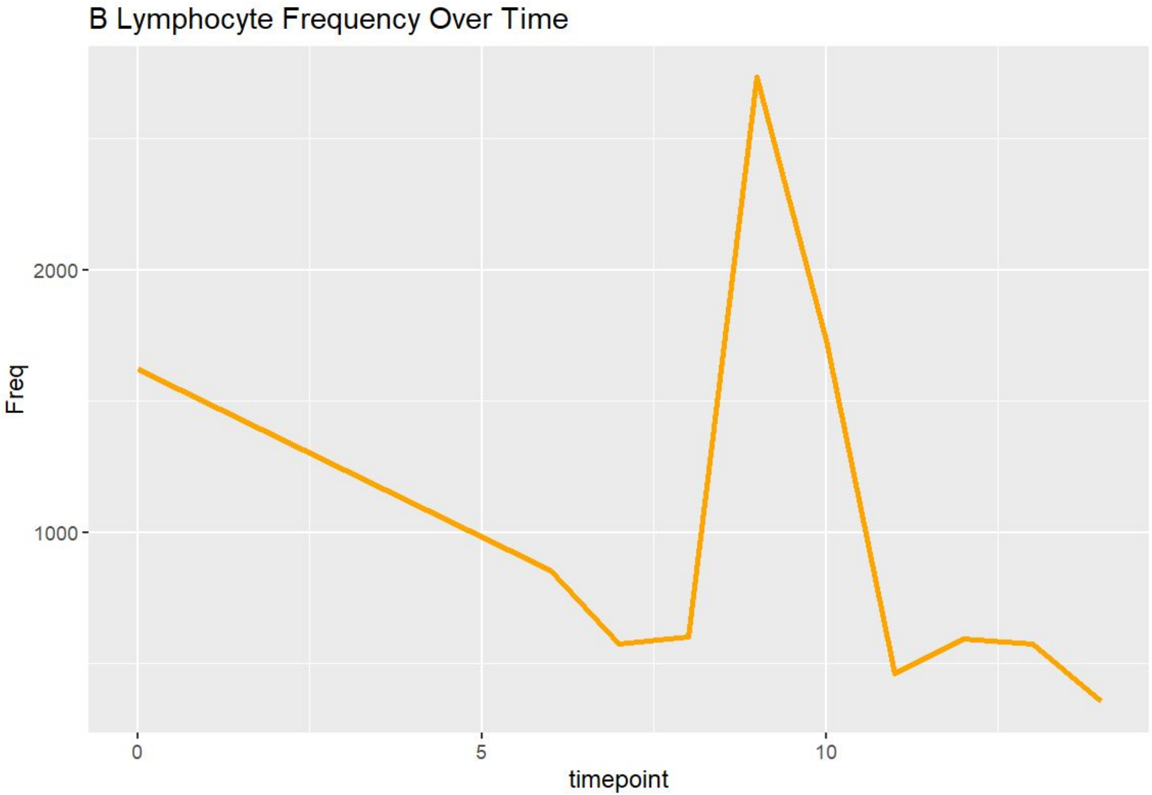




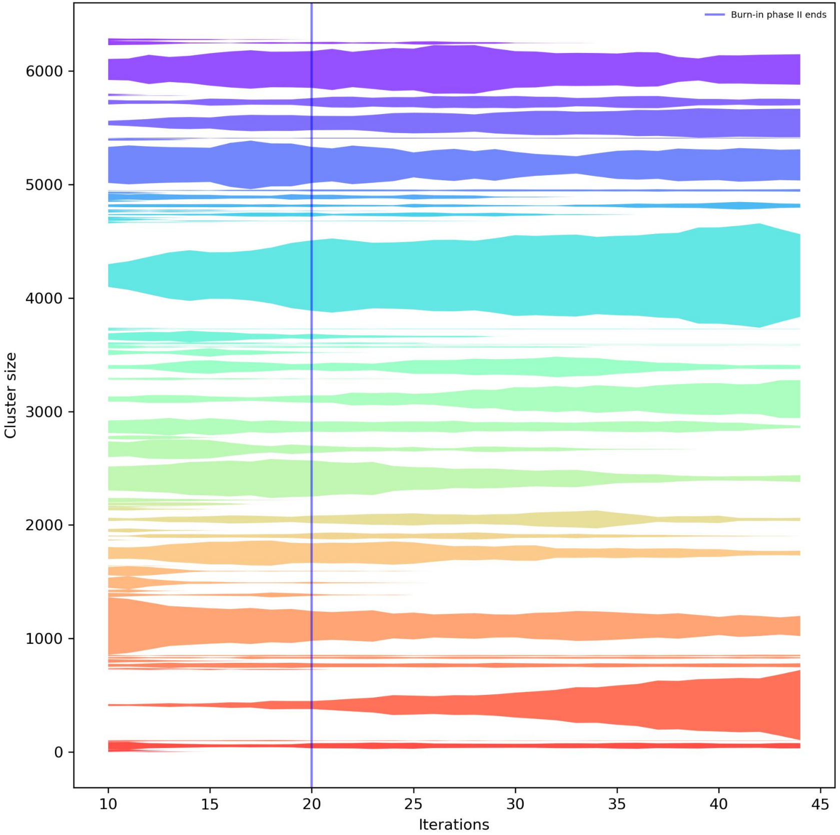
Stanford | MEDICINE

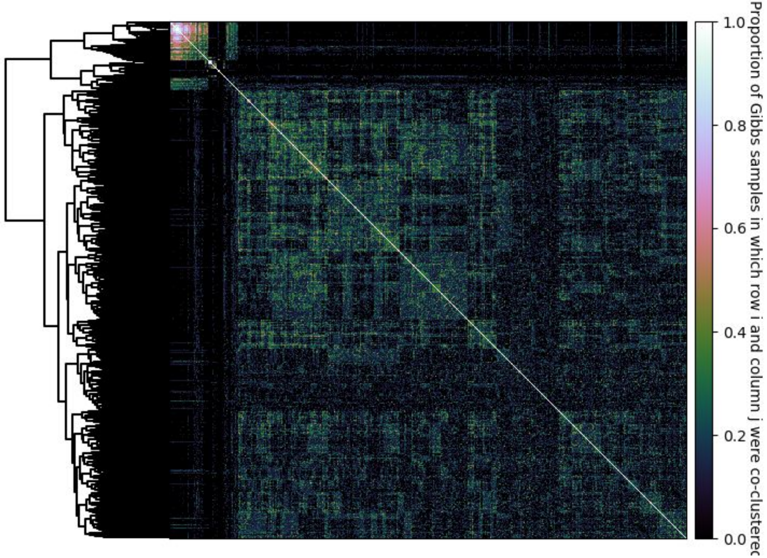
Closer Look at B Lymphocytes



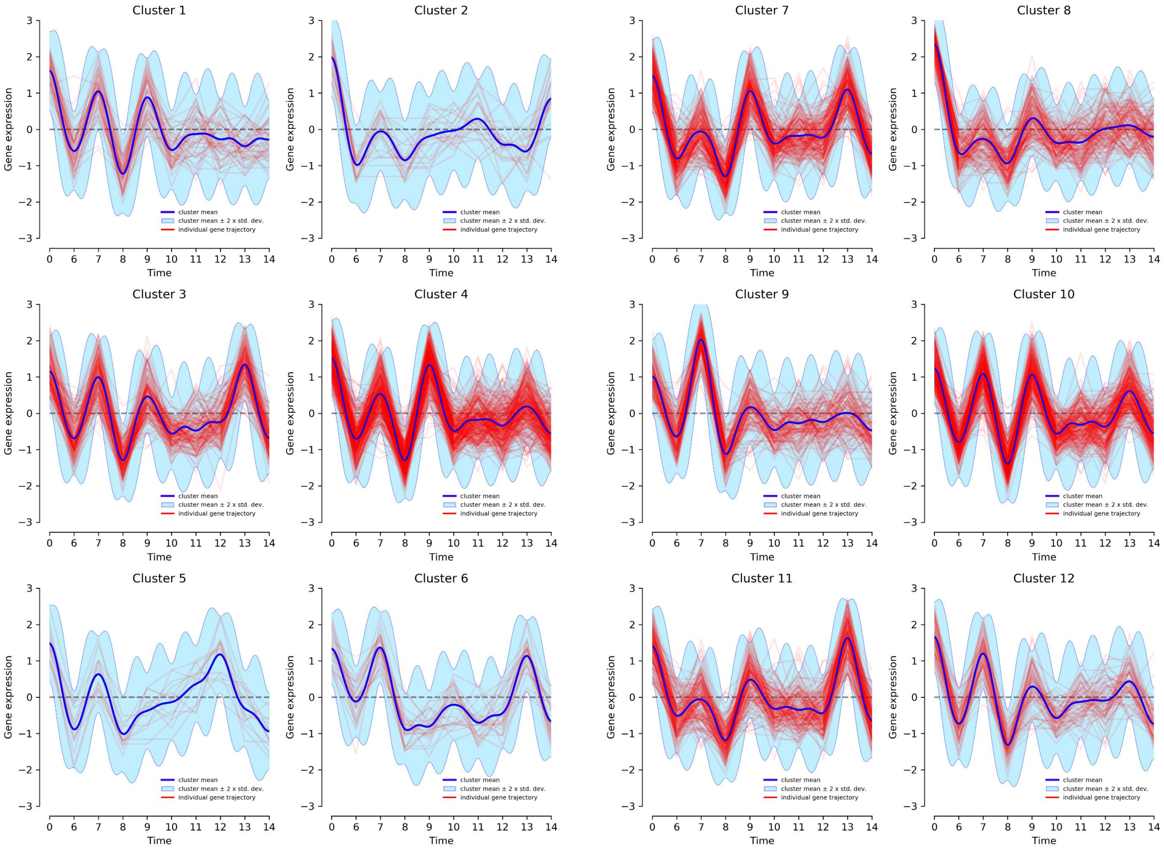


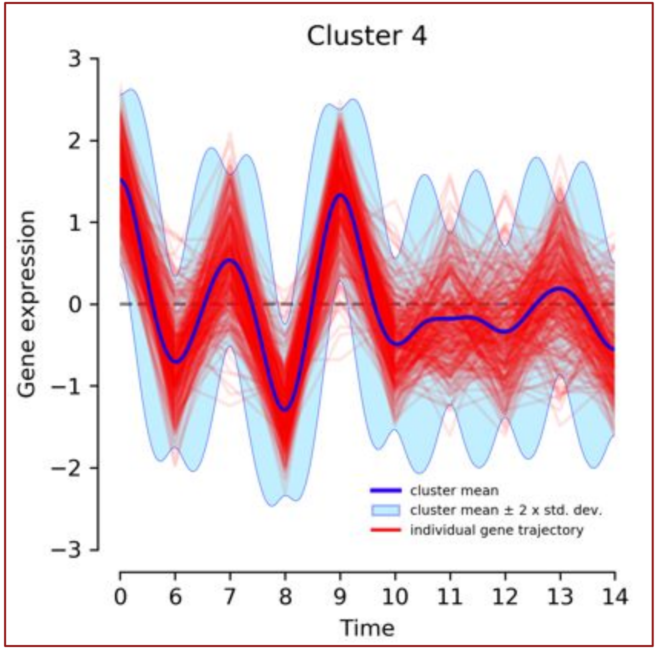
DP_GP Iteration Results for B Lymphocytes



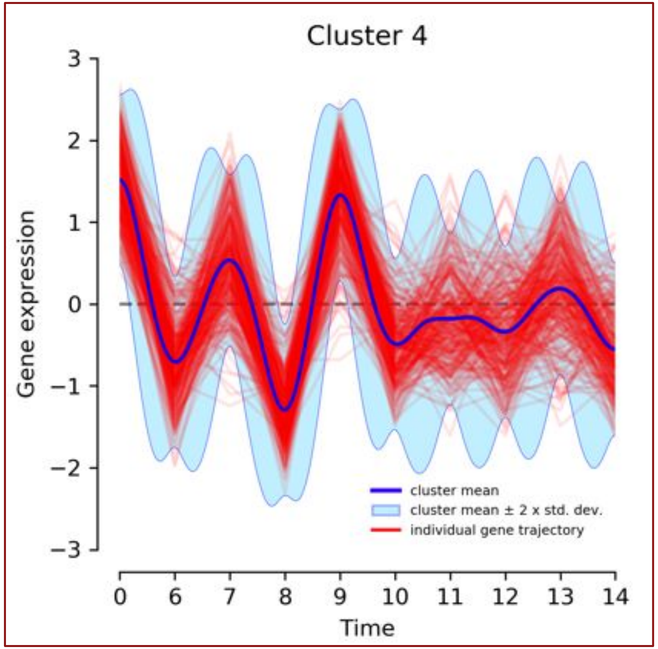


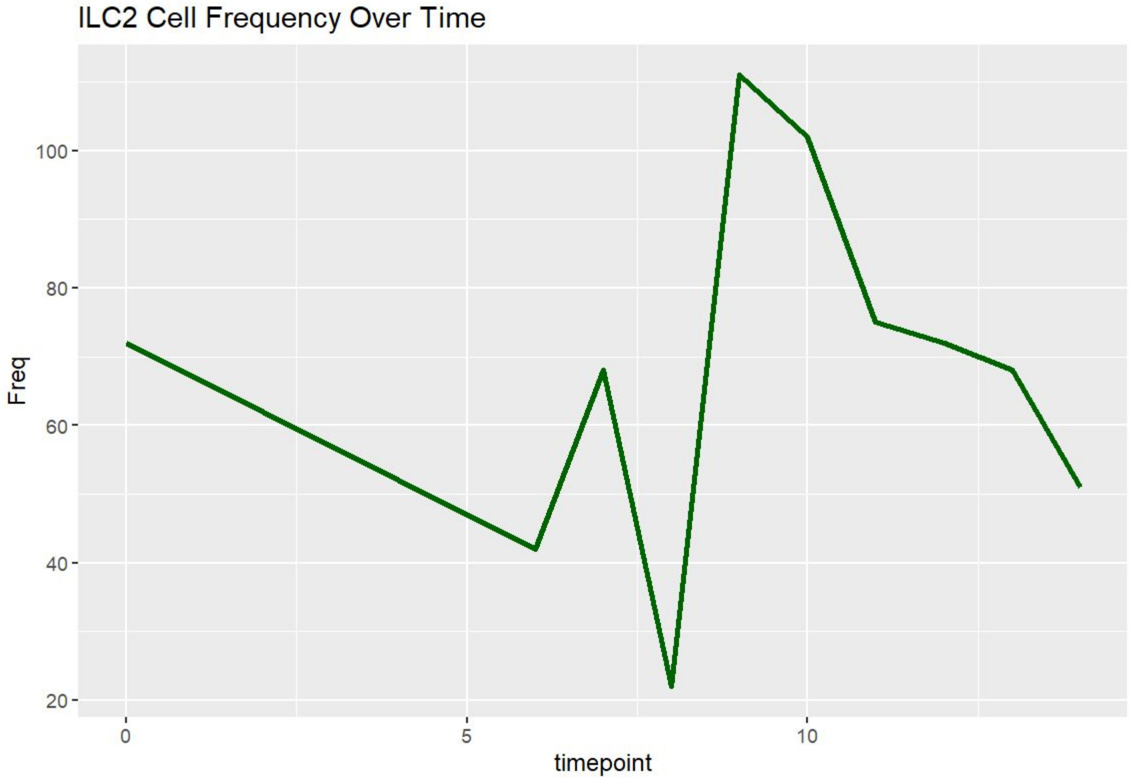
DP_GP Gene Expression for Type 2 Innate Lymphoid Cells



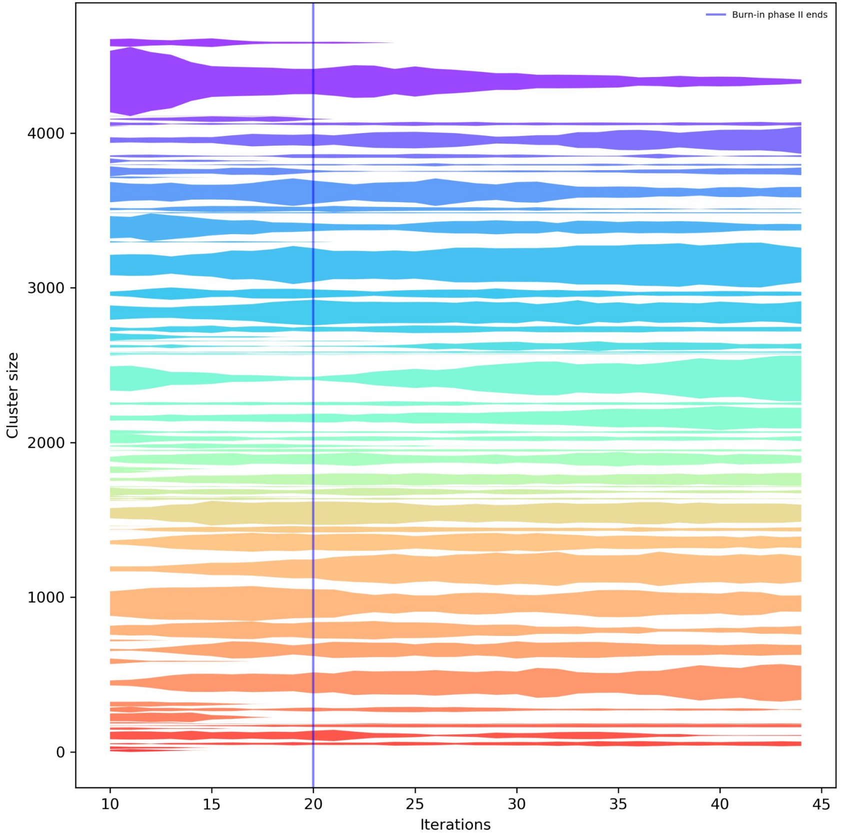


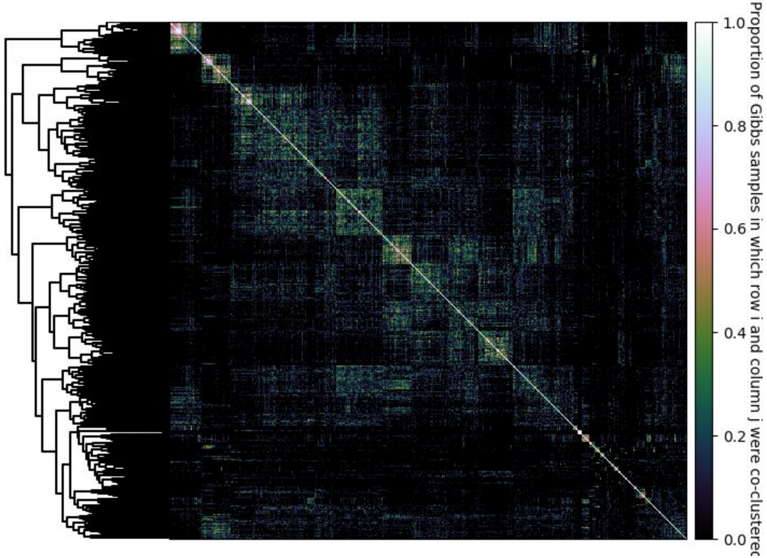
Closer Look at Type 2 Innate Lymphoid Cells



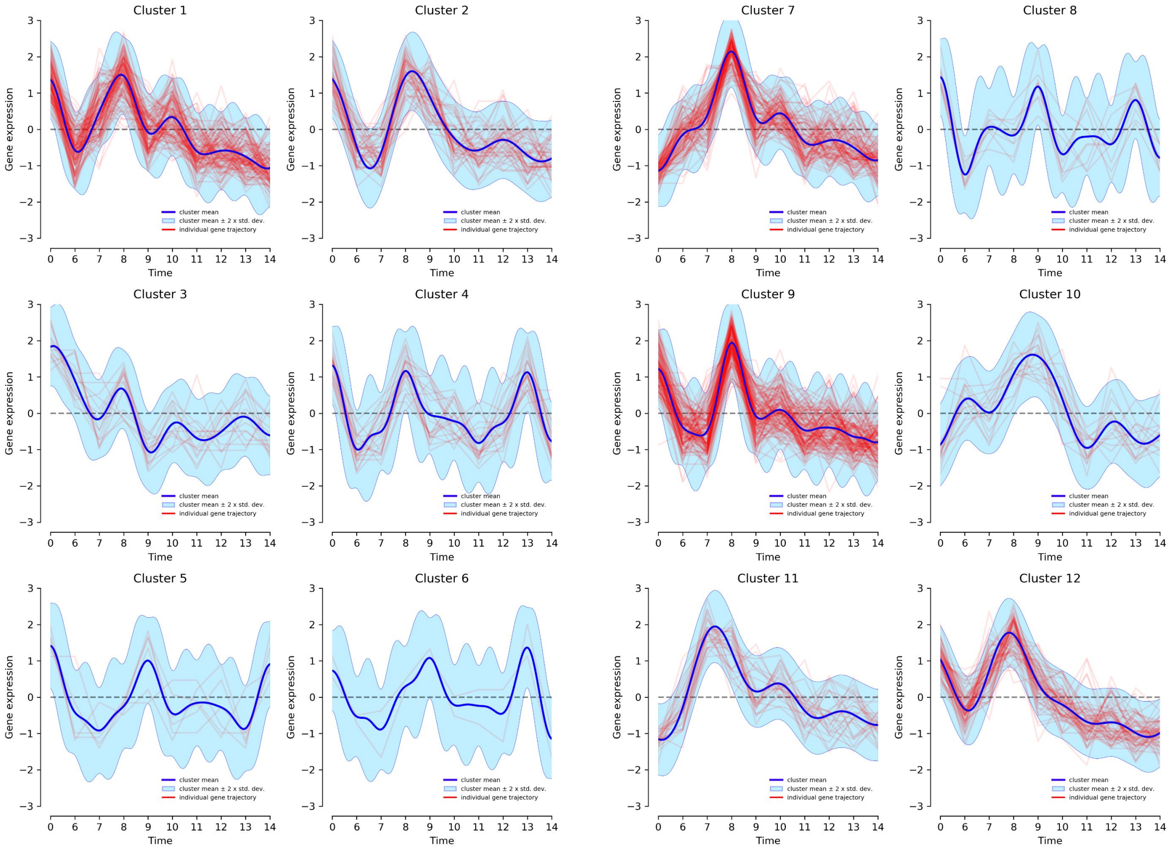


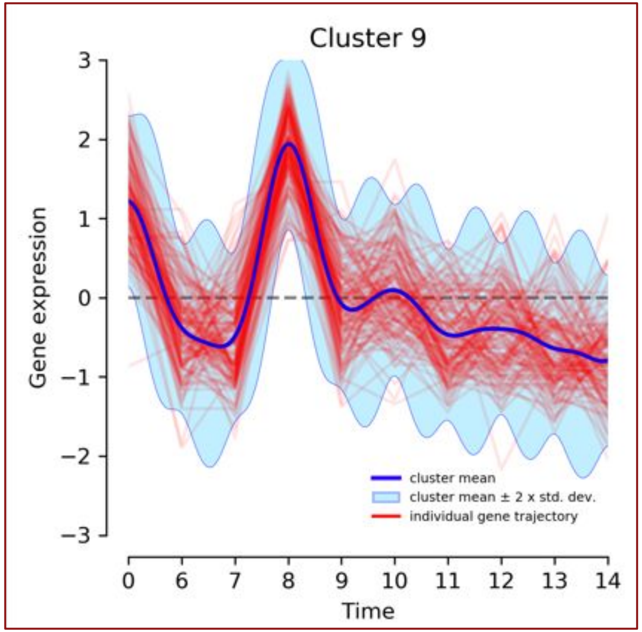
DP_GP Iteration Results for Type 2 Innate Lymphoid Cells



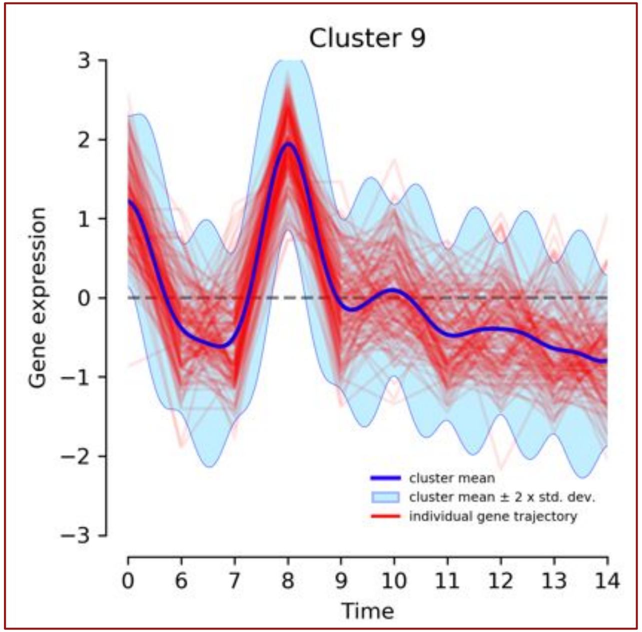


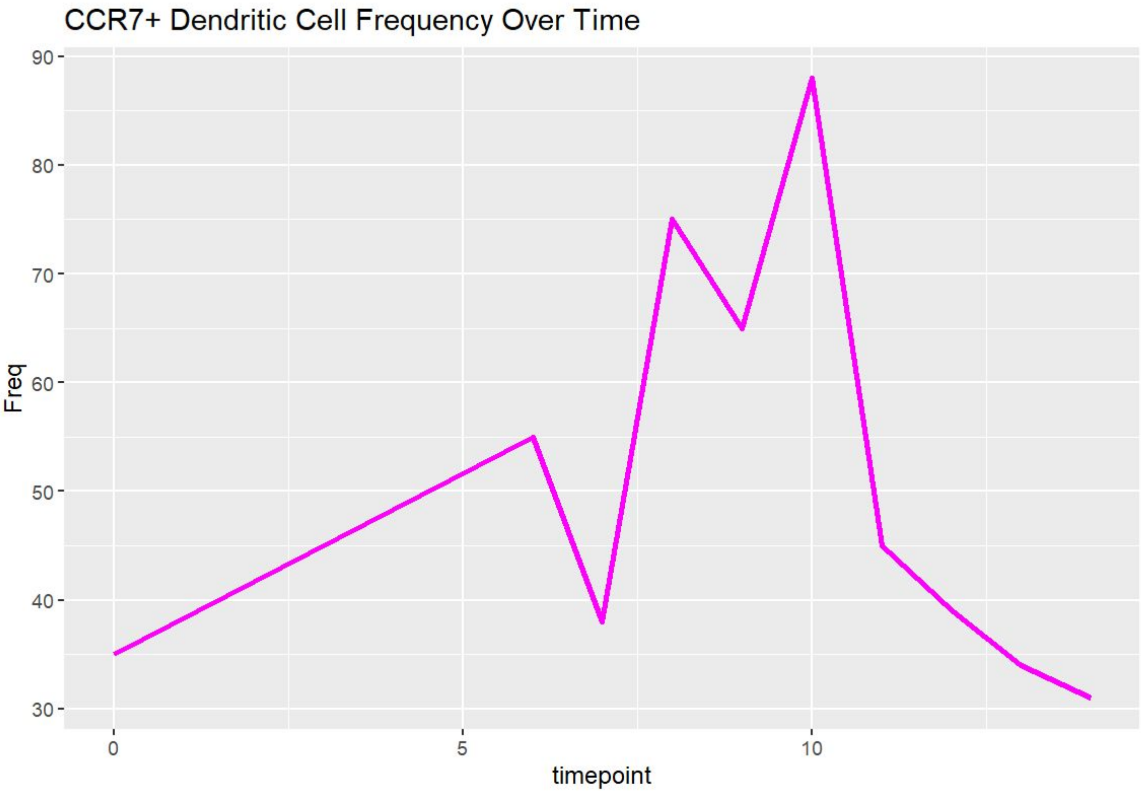
DP_GP Gene Expression for CCR7+ Dendritic Cells



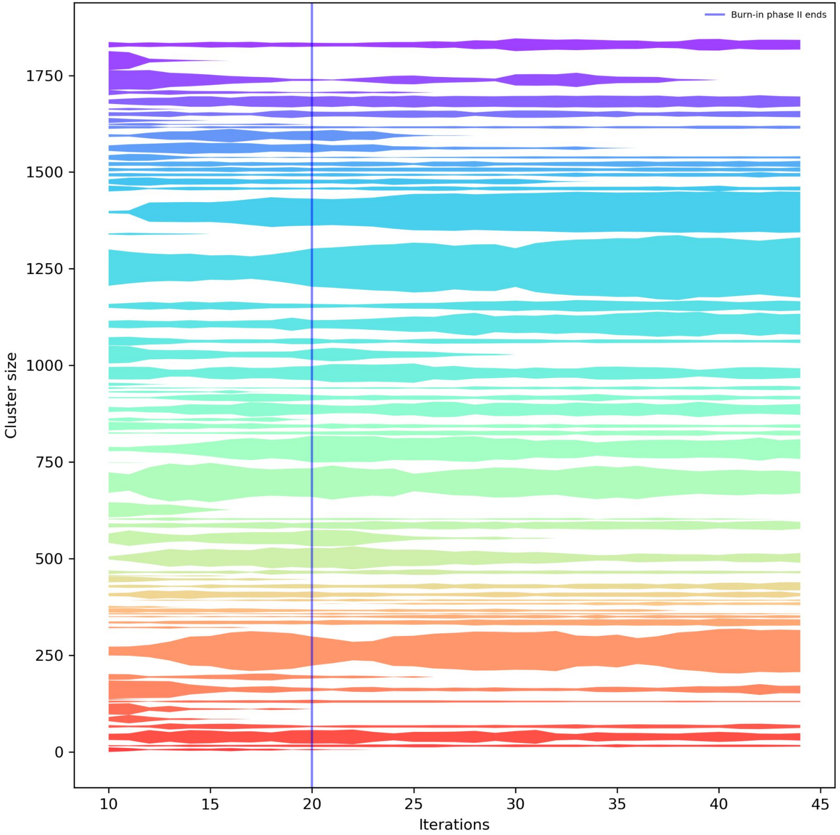


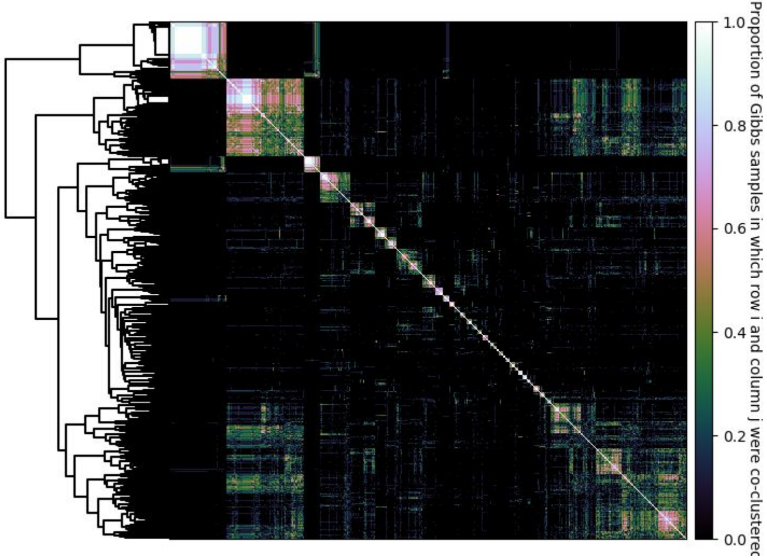
Closer Look at CCR7+ Dendritic Cells





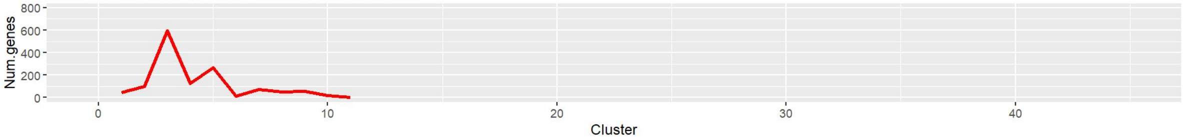
DP_GP Iteration Results for CCR7+ Dendritic Cells



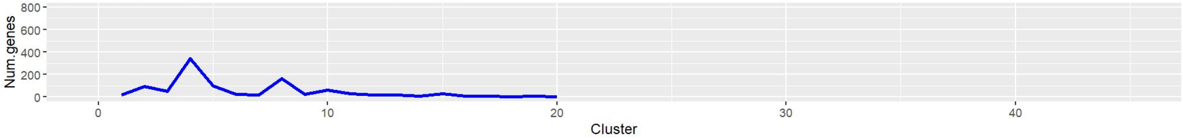


Benefit of DP_GP Clustering

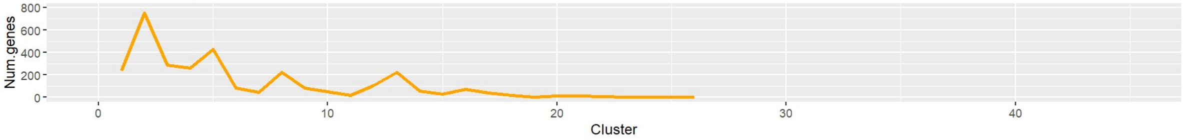
Number of Genes per Cluster (Neutrophil Cells)



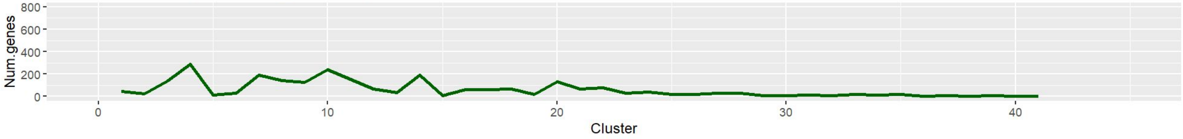
Number of Genes per Cluster (Alveolar Macrophages)



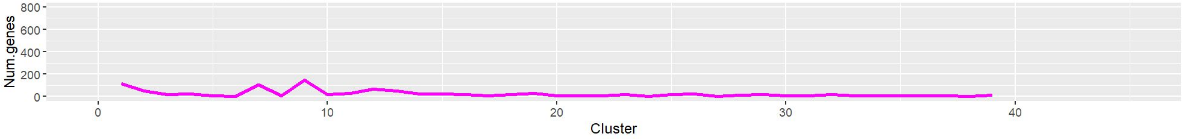
Number of Genes per Cluster (B Cells)



Number of Genes per Cluster (ILC2 Cells)



Number of Genes per Cluster (CCR7+ Dendritic Cells)



Conclusions + Implications

Key Conclusions + (Possible) Implications

Conclusion	Implication
Correlations between gene expression and number of cells over time	More cells increases gene expression → connections between up-down regulation and frequency of cell type → potential for target-based therapies and indications for dosages
Most frequent cells were related to immunity/immune-response functions	The body does try its best to combat cancer, but still didn't fully perform its job → potential for target-based therapies & research for cell-cell interactions in cancer metastasis
Key spikes of gene expression at certain time points	Relativity to the progression of illness (i.e. extreme gene expression (up or down-reg) correspond to specific cycles or pathways of cancer) → potential for outcomes research (pinpoint time of cancer progression that is the worst/least worst) OR unknown issue during data collection in lab

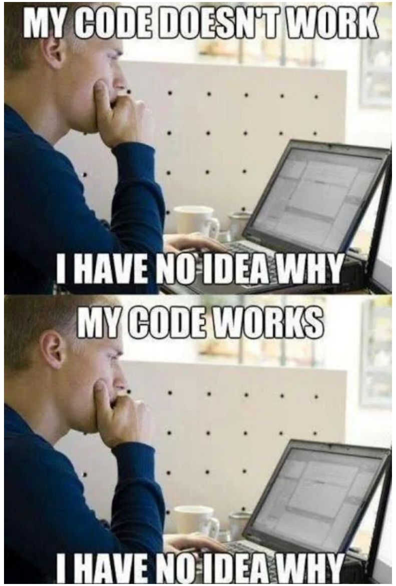
Key Conclusions + (Possible) Implications

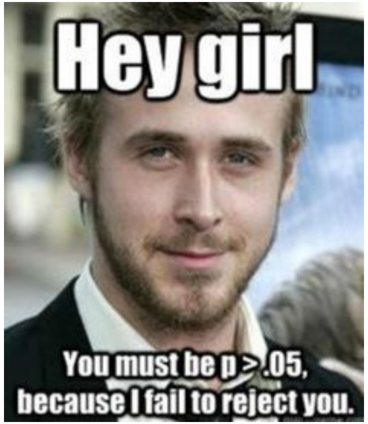
Conclusion	Implication
Number of clusters increased with the less frequent cell types	Less frequent clusters have extreme variability and uncertainty → rare cell types or possible “by-stander” cells in cancer progression → less certain about their purpose and importance → optimization problems and possible indication of cancer progression having less impact on gene expression for these specific cell types
Overall optimal number of clusters ~<10	DP_GP will need to be iterated more to condense clusters

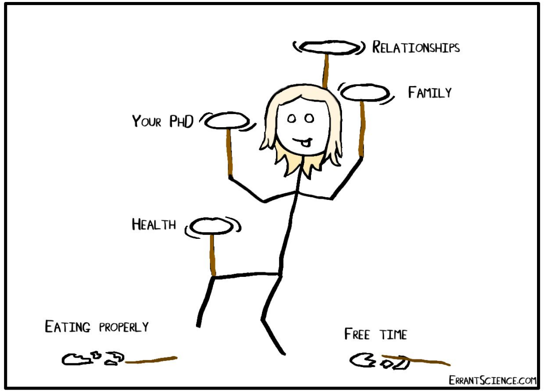
Challenges & Opportunities

Challenges

- Updating DP_GP code
- Extracting and formatting data
- Figuring out the best statistical thresholds
- Adjusting to grad school life right after undergrad (balancing classes and research)





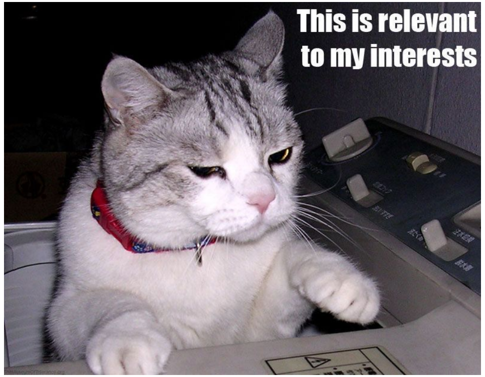


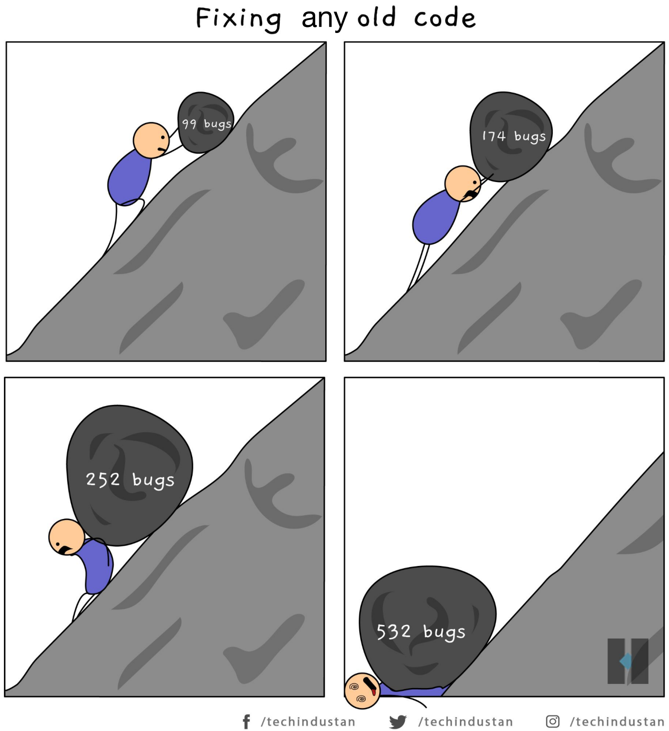
TRYING TO DO EVERYTHING DURING A PHD CAN BE A BIT TRICKY

When you start grad school right after undergrad
Ah shit, here we go again.

Opportunities (What I Learned)

- The importance of making user-friendly and up-to-date code/software/methods
- The endless possibilities of research in gene expression alone and the various different ways to analyze it
- Developed an interest in RNA-seq analysis/research





If I Had More Time

- Analyze all cell types from the data (17 total)
- Combat the memory issues to run more iterations ⟶ better clusters
- Run a DP_GP analysis on immune-response cells vs non-immune-response
- Find a way to collectively run DP_GP on the entire dataset (not just individual cell types)
- Conduct cross-validation techniques or methods for gene expression results
- Improve DP_GP software to efficiently work with single cell data (initially used to analyze bacterial growth)

References

References

[1] McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, et al. (2018) Clustering gene expression time series data using an infinite Gaussian process mixture model. PLOS Computational Biology 14(1): e1005896. https://doi.org/10.1371/journal.pcbi.1005896.

[2] McGinnis, Christopher S., et al. “The Temporal Progression of Immune Remodeling during Metastasis.” bioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2023, www.biorxiv.org/content/10.1101/2023.05.04.539153v1.

[3] Law, Charity. RNA-Seqbasics: From Reads to Differential Expression - Github Pages, combine-australia.github.io/RNAseq-R/slides/RNASeq_basics.pdf.

[4] McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009 Mar 15;25(6):765-71. doi: 10.1093/bioinformatics/btp053. Epub 2009 Jan 28. PMID: 19176553; PMCID: PMC2654802.

[5] “Lab #5 Differential Expression.” Data Analysis, www.bioconductor.org/help/course-materials/2015/Uruguay2015/day5-data_analysis.html.

Q&A + Acknowledgements

Acknowledgements

Vidal Arroyo
(Biophysics - Engelhardt)

Department of
Biomedical
Data Science
AT STANFORD MEDICINE