Mini-Project 2

Josh Popp

April 2020

1 Methods

Discovery and characterization of variance QTLs in human induced pluripotent stem cells

We assume the count data are generated by a zero-inflated negative binomial (ZINB) distribution. Let:

- r_{ijk} be the number of molecules for individual i, cell j, gene k
- R_{ij} be a size factor for each cell
- μ_{ik} be proportional to relative abundance
- ϕ_{ik} be the variance of expression noise
- π_{ik} be the proportion of excess zeros
- x_{ij} be a q-vector of confounders per cell
- β_k be a q-vector of confounding effects on gene k

(Copied the above from the paper as necessary context)

We model the observed reads of each gene, in each cell, for each individual r_{ijk} as following a Poisson distribution parametrized by (is this what the semicolon means?) the product of the cell's size factor, the influence of the cell's confounders on the gene, and the modeled gene expression before amplification λ_{ijk}:

$$r_{ijk} \sim \text{Poisson}(\cdot, R_{ij} \exp(x'_{ij}\beta_k)\lambda_{ijk})$$

Then we model the expression before amplification with a point-Gamma distribution to account for dropout, where there is a chance of the expression coming from the true distrubtion of reads (a Gamma distribution parametrized by relative abundance and variance due to true expression noise), as well as a gene-specific chance (π_{ik}) of dropout:

$$\lambda_{ijk} \sim \pi_{ik}\delta_0(\cdot) + (1 - \pi_{ik})\text{Gamma}(\cdot; \mu_{ik}, \phi_{ik})$$

We do not account for the increased likelihood that lowly-expressed genes will be dropped out aside from making π gene-specific, allowing the model to learn this relationship.

Under this model, the mean and variance of observed gene expression (expression before amplification) are as follows:

$$E[\lambda_{ijk}] = (1 - \pi ik)\mu_{ik}$$
$$V[\lambda_{ijk}] = (1 - \pi ik)\mu_{ik}^2\phi_{ik} + \pi_{ik}(1 - \pi_{ik})\mu_{ik}^2$$

I understand that it wouldn't be appropriate to write out explanations for the math in every methods section like this, but once I hit a string of equations like this I have a hard time following. In this paper, it

took me several run throughs to understand what was actually being modeled, and it wasn't until class that I realized the meaning of the Gamma-point process. So I was trying to break down the meaning of each step through here, and make it very clear what is being modeled and how. I also wanted to emphasize that while they call λ the expression, it's not the true expression since it's modeled with dropout - unless they're treating dropout as a biological process.

2 Replication

GPFates

Setup/ Installation

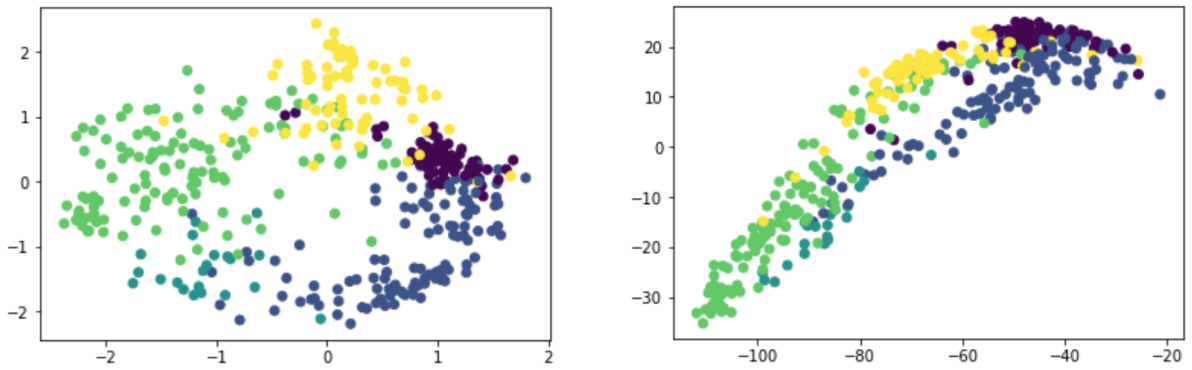
Setting up the software necessary to run GPFates took about an hour due to challenges in getting dependencies in line, finding/ downloading github repos, getting directory structure right, and using a Jupyter notebook in MARCC for the first time.

What they provide

On the GPFates github, they provide an example folder containing an expression matrix and metadata. This metadata contains not only day assignments, read counts, etc, but also columns for some of their results like dimensionality reduction and inferred pseudotime. With this, they provide a tutorial, which shows some of how to run through their pipeline, but most of it is commented out. They skip to the last stage of analysis, access the columns they have saved in their metadata, and use that to produce the visualizations which are exactly as shown in the paper. I tried doing this by going through the steps from the start, and faced some challenges in reproducing but after reworking some parts ended up with similar results.

Dimensionality Reduction

I performed dimensionality reduction using their method (left), and also did PCA on the data to compare (right), and was interested to see that PCA picked up on a very similar pattern. Shown are the dimensionality reduced representations, colored by day (day 0 purple -> day 7 yellow)

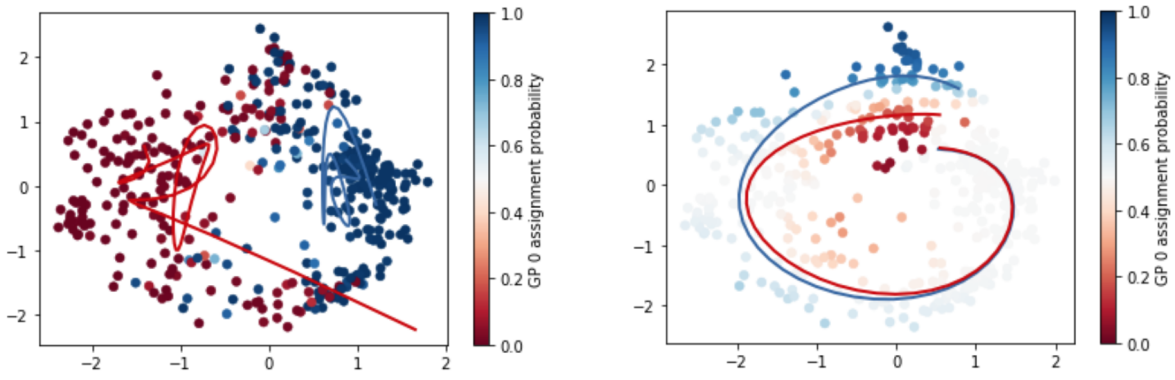


(a) GPLVM **(b)** PCA

Figure 1: Dimensionality Reductions

Pseudotime and (Multi-) Fate Inference

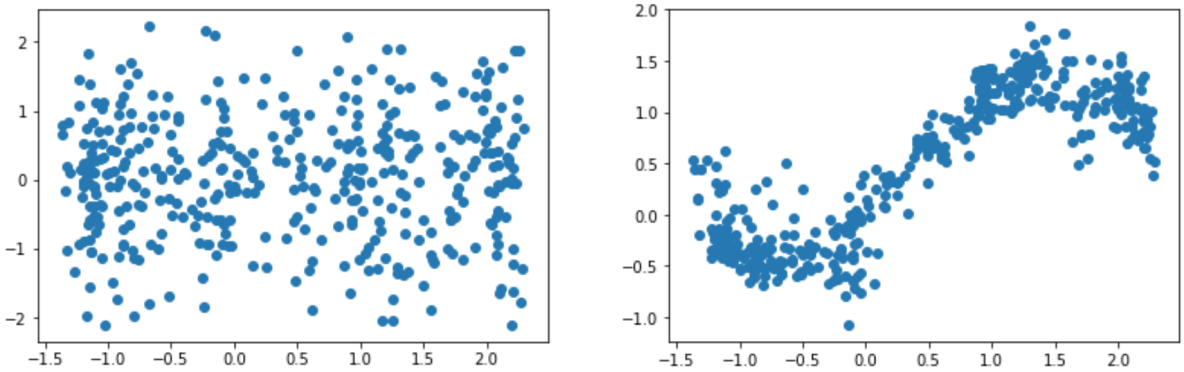
I ran through their pipeline to infer pseudotime from the dimensionality-reduced data, and fed pseudotime and these data into their overlapping mixture of Gaussian processes method to detect fates. I got the gem on the left, so for a sanity check I compared this to what you get when you skip the actual analysis and just use the columns they gave you in the metadata, which is shown on the right.



(a) My results **(b)** Their results

Figure 2: Fates according to GPFates

To try to understand the subtle differences between these two, I took a look at the metadata and found that rather than using `gplvm_0` and `gplvm_1` as I was using (and as their example teaches you to create) they were using `gplvm_2d_0` and `gplvm_2d_1`. I didn't see where a second dimension could be coming from, and this wasn't anywhere in their tutorial, so I checked how my variables lined up with theirs and it appeared to be almost perfectly one to one. Then I looked at how my pseudotime lined up with theirs, and it was not almost one to one (see left). When I ran OMGP, it converged what seemed like instantly, so I tried re-running that a few times changing a couple parameters (in particular, tried changing days, which are used as priors, from 0-indexing as they have in their tutorial/ metadata to 1-indexing as they have in the paper). I ran into a few errors, switched this back, it still converged instantly, ran it a couple times, and eventually I got a new pseudotime which, crazy enough, is a sine transformation of theirs (see right).

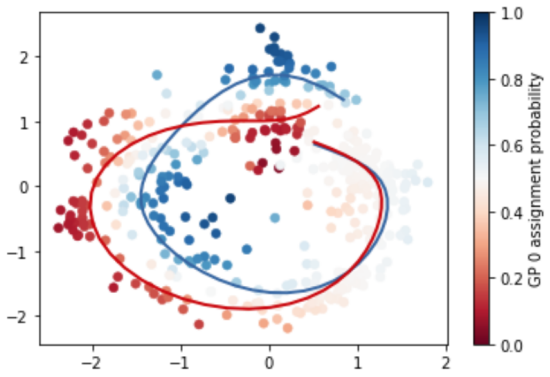


(a) First try **(b)** nth try

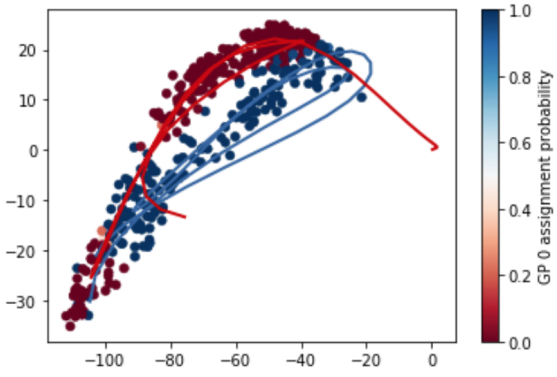
Figure 3: Their pseudotime (x) vs my pseudotime (y)

This is pretty wild, and it's not immediately apparent to me why this would happen. My thought is that as you move from 1 to 1 along a unit circle, y follows this same pattern. Our data, in their inferred low-dimensional space, approximately lies along a circle (or an ellipse or something) and pseudotime increases as you traverse the circle from 1 to 1. So maybe there's something there - I want to think that their `gplvm_2d` variables that they use might partially be behind this (the circle idea roughly corresponds to considering two dimensions simultaneously as compared to one) but I showed that they line up almost perfectly so I'm not sure. Either way, this new pseudotime gives us results that look roughly like theirs (left), BUT the bifurcation point appears to be occurring prematurely. Just for comparison, since PCA captured a somewhat similar trend to their GPLVM-inferred dimensions, I went back through the pipeline using my own `pca_0` and `pca_1`, to see how well this all would work on PCA, and it doesn't work great (left). However, considering they compute pseudotime and fates (final trajectories) separately, a different pseudotime method could still

be compatible with this one I think.



(a) My results with patience

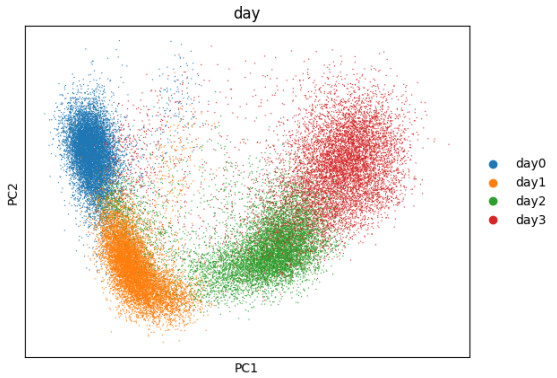


(b) My results with PCA

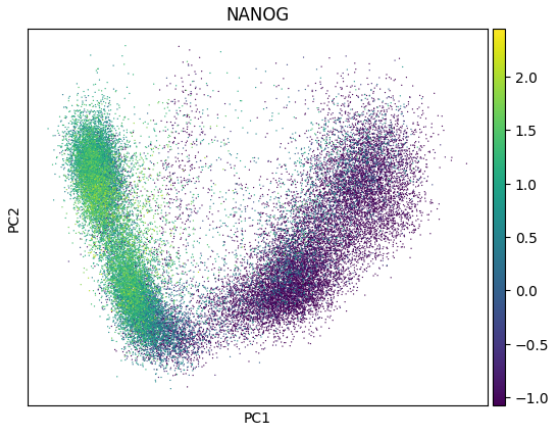
3 Final Project Proposal and Preliminary Results

For this project, we will be re-implementing and building on *Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression* by Cuomo et al. [1]. We will be reproducing their pseudotime inference and differential expression analysis (fig 1) and eQTL mapping (fig 2). For improvements, we will look at alternative pseudotime methods besides using principal component 1 (and diffusion pseudotime, which they also explored but did not present). Namely, we'll compare results of using Slingshot and Monocle3 for pseudotime inference. We will also consider two approaches to eQTL calling using pseudotime: one using pseudotime as a fixed effect in a linear mixed model (as implemented in this paper and Strober et al.), and another that resembles a pseudo-bulk approach, calling eQTLs on clusters obtained from Louvain clustering on a UMAP embedding of the data, looking at how both of these compare to a true pseudo-bulk approach, our best approximation of non-single cell RNAseq with the given data. We will investigate static, dynamic (linear, and non-linear) eQTLs and look at functional enrichments within all subsets of these, with a particular focus on lead switching events.

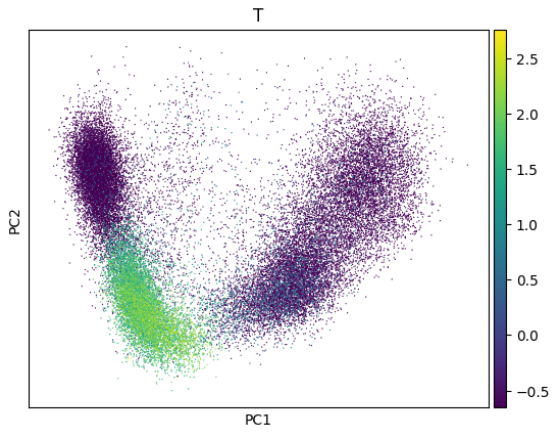
Here's a sneak peek at what's to come:



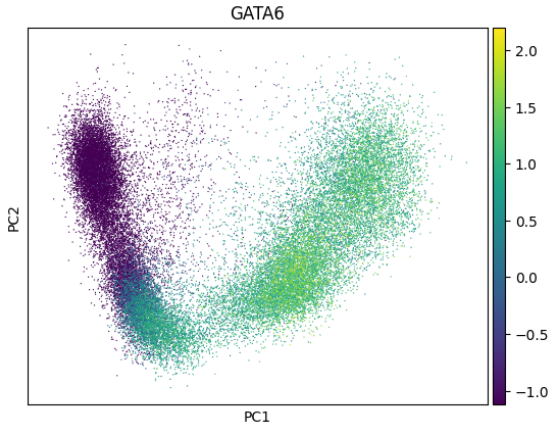
(a) PCA, colored by day



(b) PCA, colored by NANOG expression, a stem cell marker gene



(c) PCA, colored by T expression, a mesoderm marker gene



(d) PCA, colored by GATA6 expression, a definitive endoderm marker gene

4 Specific Aims

Single-cell RNA sequencing has become an incredibly powerful tool for deciphering the heterogeneity that is prevalent across biological contexts. This technology has especially catalyzed the investigation of differentiation processes by offering a higher resolution than is possible with bulk RNA sequencing. Trajectory inference methods enable the assignment of cells to a position along an inferred developmental path based on their individually measured (or inferred) expression profile, making it possible to sift through densely heterogeneous populations of partially differentiated cells and even distinct cell types to study the genetic programs underlying natural development, directed differentiation, and cellular reprogramming. However, as our understanding of the genetic programs underlying developmental processes has leapt forward, our understanding of the regulatory architecture behind these programs has been slower to progress. The dynamics of genetic regulation on gene expression have only recently begun to be explored. This proposal aims to utilize genomic data and single-cell RNA sequencing to improve our understanding of the dynamics of genetic regulation during human development, and utilize this information to improve our understanding of the biological phenomena that are at play.

The aims of the proposal are as follows.

Aim 1: To leverage biologically important latent factors to improve the identification of quantitative trait loci. The first attempts to use pseudotime to identify expression quantitative trait loci (eQTLs)

have recently been established, though they are limited in their scope [1]. The trajectory inference methods previously applied to this objective are limited, and thus will be challenging to extend to more complex biological phenomena, such as bifurcating or multifurcating developmental pathways (as occur in humans, plants, and nearly all multicellular organisms). We intend to utilize more sophisticated trajectory inference methods which are not constrained to linear trajectories to obtain pseudotime data for RNA-sequencing profiles obtained from spontaneously differentiating embryoid bodies from human cell lines. This will power the simultaneous identification of dynamic eQTLs active across a range of biological contexts (ie tissue types) in a single study. Furthermore, this methodology can be extended beyond pseudotime to other biologically important latent factors. For example, pseudospace has previously been identified as an important factor in the epithelial-to-mesenchymal transition, making this another candidate factor that could help identify context-specific eQTLs [2].

Aim 2: To improve trajectory inference through incorporation of genetic variation. One objective of trajectory inference is to cut through the heterogeneity of cellular populations to identify the distinct developmental paths pursued by subpopulations of cells. One approach to identifying such bifurcation (or multifurcation) events utilizes an overlapping mixture of Gaussian processes model, in which each distinct developmental pathway is modeled as its own Gaussian process (and these processes are identical leading up to a bifurcation event) [3]. We will investigate the impact of incorporation of genetic variant information on trajectory inference in a probabilistic framework. When genetic variation drives such bifurcation events (whether directly or indirectly, ie via genetically regulated epigenetic modification), incorporating genomic data in such a model could improve performance by adding the constraint that genetically similar individuals are likely to follow more similar (or the same) trajectories. This constraint in similarity between trajectories can be explicitly modeled in an overlapping mixture of Gaussian processes model. Furthermore, as has previously been shown [3], once such a model has been trained it is possible to identify driver genes involved in bifurcation events. These can then be used to guide the search for biologically important quantitative trait loci in a developmental pathway.

Aim 3: Identification of causal loci through genetic engineering The final objective of the search for quantitative trait loci is the identification of causal loci, though this is made difficult by the extensive linkage disequilibrium present in the human genome. However, modern gene editing technologies such as the CRISPR-Cas9 system have made highly accurate gene editing possible, including at the single-nucleotide resolution [4]. Recent methods have even made high-throughput gene editing possible at the single-cell level to assay a wide range of edits simultaneously [5]. Utilizing this pooled CRISPR system, for essentially any phenotype that can be measured in high-throughput (such as gene expression), we can interrogate a set of putative causal SNPs and investigate which are truly causal, eliminating the ambiguities introduced by linkage disequilibrium in a population. The use of single-cell sequencing can improve the quality of such assays by providing simultaneous measurement of phenotype and transfection efficiency, a source of significant heterogeneity in bulk expression datasets. We will utilize this method to experimentally validate and search for causality in the eQTLs identified by aim 1, and can further use them to validate findings of computational approaches to fine mapping such as eCAVIAR [6].

Understanding the static and dynamic genetic regulation of gene expression through differentiation is a major outstanding goal for the understanding of human development, and the treatment and cure of many developmental disorders. With the rapid advancement of single-cell RNA sequencing and gene editing, and through the use of computational methods that are able to integrate expression and genotype information, we are well positioned to identify the regulatory regions which underlie the genetic programs involved in biological development.

References

[1] Cuomo, A. S. *et al.* Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. *Nature communications* **11**, 1–14 (2020).

[2] McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nature genetics* **51**, 1389–1398 (2019).

[3] Lönnberg, T. *et al.* Single-cell rna-seq and computational analysis using temporal mixture modelling resolves th1/tfh fate bifurcation in malaria. *Science immunology* **2** (2017).

[4] Lee, H. K. *et al.* Targeting fidelity of adenine and cytosine base editors in mouse embryos. *Nature communications* **9**, 1–6 (2018).

[5] Hill, A. J. *et al.* On the design of crispr-based single-cell molecular screens. *Nature methods* **15**, 271 (2018).

[6] Hormozdiari, F. *et al.* Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics* **99**, 1245–1260 (2016).