Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data (Campbell & Yau 2018)

Pseudotime algorithms: employed to extract latent temporal information from cross-sectional data sets allowing dynamic biological processes to be studied in situations where the collection of time series data is challenging or prohibitive

PhenoPath: a novel statistical framework (**hybrid regression-latent variable model**) that **learns how pseudotime trajectories can be modulated through covariates that encode such factors**.

- **longitudinal studies** are often challenging to conduct and cohort sizes limited by logistical and resource availability
- **cross-sectional surveys** of a population are relatively easier to conduct in large numbers and more prevalent for molecular 'omics based studies.
    - do not directly capture the changes in disease characteristics in patients but it may be possible to recapitulate aspects of temporal variation by applying **"pseudotime" computational analysis**

The objective of **pseudotime analysis** is to take a collection of high-dimensional molecular data from a cross-sectional cohort of individuals and to map these on to a series of one-dimensional quantities, called **pseudotimes**

These **pseudotimes measure the relative progression of each of the individuals along the biological process of interest**, e.g., disease progression, cellular development, etc., allowing us to **understand the (pseudo)temporal behaviour of measured features without explicit time series data**

- possible when individuals in the cross-sectional cohort behave asynchronously and each is at a different stage of progression
- by creating a relative ordering of the individuals, we can define a series of molecular states that constitute a trajectory for the process of interest

Pseudotime methods generally rely on the **assumption that any two individuals with similar observations should carry correspondingly similar pseudotimes** and algorithms will attempt to **find some ordering of the individuals that satisfies some overall global measure** that best adheres to this assumption

- differ in the way "similarity" is defined
- when applied to molecular data, typically capture some dominant mode of variation that corresponds to the continuous (de)activation of a set of biological pathways

gained particular popularity in the domain of **single-cell** gene expression analysis (where **each "individual" is now a single cell**) e.g. to model the differentiation (cf. https://github.com/agitter/single-cell-pseudotime)

- use advanced machine learning techniques (e.g. can characterise cell cycle, model branching behaviours)

these single-cell applications were **predated by more general applications** in modeling disease progression

- provided early inspiration for single-cell pseudotime methods

To date, little cross-over between these distinct application domains (different contexts of application)

- interesting possibilities by translating recent advances in single-cell pseudotime modelling to disease progression modelling

recent single-cell pseudotime approaches for branching pseudotime trajectories, these **can only be retrospectively examined for their association with prior factors of interest**

→ develop **a statistical model in which these factors could be explicitly incorporated** into pseudotime analysis

- would provide **a mechanism to account for known genetic, phenotypic or environmental factors allowing gene expression variability to be decomposed into different contributory factors**
- would allow us to answer questions related to the **interaction between heterogeneity in these external factors and biological progression**

a novel **Bayesian statistical framework for pseudotime trajectory modelling that allows explicit inclusion of prior factors of interest**

- **allows to incorporate information in the form of covariates** that can modulate the pseudo- temporal progression allowing sub-groups within the cross- sectional population to each develop their own trajectory
- **combines linear regression and latent variable modelling and allows for interactions between the covariates and temporally driven components** of the model
- first approach to allow for **modelling pseudotime trajectories on heterogeneous backgrounds** allowing its **utility in both single and non-single cell** applications

PhenoPath provides a **probabilistic ordering of high-dimensional gene expression measurements across objects** (e.g., cells, tumours, patients, etc)

- achieved by **compressing** the information contained within the data on to a **unidimensional axis**

    - construct an axis such that **relative positions along the axis correspond to some meaningful biological or disease progression**
- novelty: introduce the notion that **objects may have different labels (covariates)** attached to them corresponding to different innate properties or exposure to external stimuli
    - these factors might cause the objects to evolve over (pseudo)time differently
- **simultaneously learns a pseudotemporal axis** that is common to the different object labels, **while decomposing gene expression variability into static and dynamic components**

$\Rightarrow$ a **Bayesian** statistical framework that integrates **linear regression and latent variable modelling**

- the observed data (y_n) for the n-th individual is a linear function of both measured covariates (x_n) and an unobserved latent variable (z_n) corresponding to latent progression that we will term pseudotime

the model involves **three components**:

1. **gene expression**: a **static** component based on your covariate status (Ax_n^T)
2. a **dynamic** component related to **how far** along the biological process you are (λz_n)
3. (main novelty) an **interaction component** which allows your **covariate status to change the direction of the dynamic component** of the gene expression ($Bx_n^T z_n$)

- if only 1. used = linear regression based differential expression analysis
- if only 2. used = factor analysis

the **covariates** in this study are binary quantities, any arbitrary design matrix that can be used for standard regression may be used for x (

sparse Bayesian prior probability distributions are used to constrain the parameters (A, B, λ) so that covariates only drive the emergence of distinct trajectories if there is sufficient information within the data to do so

fast and highly scalable variational Bayesian inference framework that can handle thousands of features and samples in minutes using a standard personal computer

variational inference of such hierarchical Bayesian models can be sensitive to **hyperparameters values and parameter initialisation** we found PhenoPath to be **robust** to such choices

Distinct response trajectories of the dendritic cells under either LPS or PAM stimulation are evident, with a common cell state at the beginning of pseudotime diverging under LPS and PAM stimulation. Despite capture times not being used as an input, the PhenoPath pseudotime trajectory recapitulates the physical time progression of the cells with an R2 = 0.68 (Fig. 2b) with 7500 highly variable genes as input. We compared the ability of PhenoPath to recapitulate the physical progression of the cells through pseudotime inference to three state-of-the-art pseudotime algo- rithms (Monocle 28, DPT5 and TSCAN3) across a wide range of gene set sizes.

We found that for every input gene set size PhenoPath reported a higher correlation with capture time (Fig. 2c) than other methods tested

We found that while some genes that exhibit stimulant-pseudotime interactions can be identified as differentially expressed genes, the majority require
the explicit PhenoPath model to resolve the relative contributions of the static and dynamic expression components.

Pseudotemporal modelling in colorectal cancer.

- in a non-single-cell setting by examining RNA sequencing gene expression data from the TCGA colorectal adenocarcinoma (COAD) cohort
- microsatellite instability (MSI) status as a phenotypic covariate
- Pseudotime inference using PhenoPath was applied to 4801 highly variable genes across 284 COAD samples
- common pseudotemporal scale but distinct development trajectories for MSI-high and MSI-low tumours

- expression of T- regulatory cell (Tregs) immune markers (Fig. 4b) increased along the trajectory and found
- **GO analysis**: enrichment of immune-related pathways

  $\rightarrow$ PhenoPath has ordered the tumours according to levels of tumour immunogenicity and Tregs infiltration of the tumours

To corroborate this proposition, we used an **bulk RNA sequencing deconvolution tool**, quanTIseq2 which uses transcriptomic profiles of immune cells to estimate immune cell content of each tumour

- tumours identified by quanTIseq as having high regulatory T cell or immune cell content scores were most correlated with PhenoPath
  pseudotime $\rightarrow$ PhenoPath had unbiasedly identified an immunogenic contribution to colorectal cancer progression through unsupervised analysis

92 **putative covariate-pseudotime interactions including known tumour suppressor gene**

- PhenoPath identified the MLH1 gene whose interaction effect size was far larger than any othergene. This associationprovides animportant positive control since MLH1 is a well-known DNA mismatch repair gene

a standard **differential expression analysis**

- many of these 92 genes are differentially expressed between MSI groups,
- PhenoPath is able to resolve the dynamic contribution to these expression differences
  - expression of these genes in MSI-low tumours is relatively constant
  - in MSI-high tumours, there is a spectrum of expression levels that linearly changes over pseudotime following the increasing immune cell infiltration in the MSI-high tumours.

[NB: pseudotime ~ as a proxy for immune response; interactions with covariate = interactions immune response ~ microsatellite status]
We next sought to uncover whether the **other genes exhibiting interactions between the immune response and microsatellite instability** displayed a concerted action in any cancer-related interactions between the immune response and microsatellite instability displayed **a concerted action in any cancer-related pathways**. We took the **top 20 genes by interaction effect size** and performed an unsupervised pathway enrichment analysis using Reactome30

Pseudotemporal modelling in breast cancer

- pseudotemporal analysis of the TCGA breast cancer cohort using estrogen receptor (ER) status as a phenotypic cov- ariate.
- applied PhenoPath to 1135 breast cancers over 4579 highly variable genes
- identified distinct ER status specific pseudo- temporal trajectories
- markers of vascular growth pathways or angiogenesis showed common pseudotemporal progression independent of ER status. This
- a **GO enrichment analysis** indicated that the genes driving the inferred pseudotemporal trajectory were indeed enriched for vascular growth pathways

  $\rightarrow$ through unsupervised analysis, PhenoPath had ordered the breast tumours and measured breast tumour progression in terms of angiogenic development

Survival analysis using stratified (by ER status) Cox proportional hazards modelling with covariates

- the pseudotime covariate was significant ($p = 0.0032$) $\rightarrow$ gave evidence that increasing pseudotemporal progression in these breast tumours conferred reduced overall survival rates

to understand how angiogenic development differs by ER status, we examined the landscape of **genes exhibiting covariate-pseudotime interactions**

- 42% of the genes affected by an interaction between the pseudotem- poral trajectory and ER receptor status. The large percentage was expected given the heterogeneity of breast cancers and the strong stratification power of ER status in breast cancer subtyping
- positive control: ESR1 one of these

a **pathway enrichment analysis using Reactome30** to discover **whether any of the top 20 interacting genes (by β value) converge on a cancer- related pathway**

- (at a FDR <5%) enrichment for Unfolded protein response and ATF6α activating chaperone genes

  $\rightarrow$ PhenoPath analysis suggests a relationship between the ER status of the tumour to the (vascular) growth via pathway- specific action mediated by ATF6α

Many of these genes [key genes with significant interactions] exhibit a **convergence**—they have markedly different expression at the beginning of the trajectory based on ER status yet converge towards the end. **We derived a mathematical formula to infer such convergence points** and calculated these for all genes showing significant interactions (see Supplementary Results for details). Remarkably, **the vast majority converge towards the end of the trajectory** (Fig. 7c), implying **a common end-point in vascular development for both ER+ and ER− cancer subtypes**. This effect can be seen in the trajectory plots in Fig. 6a, where the ER+ and ER− tumours converge at the end of their trajectories. This suggests that while there exist **low levels of angiogenesis pathway activation, ER status dominates** gene expression while as **angiogenesis pathway activation increases it comes to dominate expression patterns over ER status**.

https://github.com/agitter/single-cell-pseudotime

Methods

$N \times G$ data matrix Y for N samples and G features (y_{ng} an entry of this matrix)

- corresponds to the measurement of **a dynamic molecular process that we might reasonably expect to show continuous evolution** such as gene expression corresponding to a particular pathway.

Learn a one- dimensional linear embedding that would be our “best guess” of such progression via a factor analysis model:

$$y_{ng} = \lambda_g z_n + \epsilon_{ng}, \epsilon_{ng} \sim N(0, \tau_g^{-1})$$

- z_n = the latent measure of progression for sample n
- λ_g = the factor loading for feature g: it essentially describes the evolution of g along the trajectory.

ith the entry in the nth row and pth column given by xnp

However, it is conceivable that the evolution of feature g along the trajectory is not identical for all samples but is instead affected by a set of **external covariates**. Note that we **expect such features to be “static” and should not correlate with the trajectory itself**.

Introducing the $N \times P$ covariate matrix X (entry given by x_{np}), we allow such measurements to perturb the factor loading matrix

$$\lambda_g \to \lambda_{ng} = \lambda_g + \sum_{p=1}^{p} \beta_{pg} x_{np}$$

- β_{pg} quantifies the effect of covariate p on the evolution of feature g.

Despite Y being column-centred we need to reintroduce **gene and covariate-specific intercepts** to satisfy the model assumptions, giving a generative model of the form :

$$y_{ng} = \eta_g + \sum_{p=1}^{p} \alpha_{pg} x_{np} + (\lambda_g + \sum_{p=1}^{p} \beta_{pg} x_{np}) z_n + \epsilon_{ng}, \epsilon_{ng} \sim N(0, \tau_g^{-1})$$

aim: **inference of z_n that encodes progression along with** β_{pg} which is informative of novel interactions between continuous trajectories and external covariates

$\rightarrow$ we place a **sparse Bayesian prior on** β_{pg} of the form:

$$\beta_{pg} \sim N(0, \chi_{pg}^{-1})$$

- the posterior of χ_{pg} is informative of the model's belief that β_{pg} is non-zero

perform **co-ordinate ascent mean field variational inference with an approximating distribution** q

Ranking covariate-pathway interactions

For each gene g and covariate p we have β_{pg} that **encodes the effect of p on the evolution of g along the trajectory z.**

Aim: identify interesting interactions for further analysis and follow-up.

The variational approximation for β_{pg} is given by

$$q_{\beta_{pg}} \sim N(m_{\beta_{pg}}, s_{\beta_{pg}})$$

which after (approximately) maximising the ELBO will give estimates $\hat{m}_{\beta_{pg}}$ and $\hat{s}_{\beta_{pg}}$ for every gene and covariate

We classify or label an interaction as of interest if

$$\frac{\hat{m}_{\beta_{pg}}}{\hat{s}_{\beta_{pg}}} > k$$

where k is a positive constant.

In other words, the interaction is **not of interest if $\beta_{pg} = 0$ falls within k posterior standard deviations of the posterior estimate of the mean of the interaction.**

$\rightarrow$ This is equivalent to a **decision theoretic loss criteria** governing whether the true value for β lies in the tails of the posterior marginal or not.

Supp. mat.

TCGA data processing

TPM matrices were retrieved from a recent transcript-level quantification of the entire TCGA study

clinical metadata, including the phenotypic covariates used in Phenopath were retrieved using the RTCGA R package

transcript level expression estimates were combined to gene level expression estimates using Scater

A PCA visualisation of the COAD dataset showed two distinct clusters based on the plate of sequencing. Rather than try to correct such a large batch effect, we retained samples with a PC1 score of less than 0 and a PC3 score greater than -10, and removed any “normal” tumour types. For input to PhenoPath we used the 4,801 genes whose median absolute deviation in log(TPM+1) expression was greater than sqrt(0.5)

A PCA visualisation of the BRCA dataset (Supplementary Fig. 4b) showed a loosely dispersed outlier population that separated on the first and third principal components. We performed Gaussian mixture model clustering using the R package mclust[8], and removed samples designated as cluster 2 in Supplementary ulation Fig. 4b, giving 1,135 samples for analysis.

For input to PhenoPath we used the 4,579 genes whose variance in log(TPM+1) expression was greater than 1 and whose median absolute deviation was greater than 0.

Simulation setup

We sought to quantify the extent to which

- having **a joint model of pseudotimes and interactions aids identification of the interactions**
- such **interactions confound trajectory inference** using traditional methods.

Mathematically, PhenoPath infers the posterior distribution $p(z, \beta|Y)$ of the pseudotimes z and in-teraction parameters β given the data Y.

The two step procedure we compare against is analogous to first **inferring an estimate** of the pseudotimes $\hat{z}|Y$ before **inferring an estimate of the interaction parameters** given the fixed pseudotimes $\hat{\beta}|\hat{z}, Y$.

To do this we simulated RNA-seq data where **a certain proportion of the genes exhibited different behaviour over pseudotime depending on the external covariate status**. We then re-inferred the pseudotimes using a variety of algorithms (including PhenoPath), and performed post-hoc differential expression (DE) analysis testing for interaction effects between the trajectory and the covariate using common DE algorithms.

Differential expression was performed with PhenoPath, Limma voom, ([13]) DE- Seq2, and MAST [14]. Raw counts were used for Limma voom and DESeq2 (as recommended), while log-normalised values were provided to MAST which is designed to work with log(TPM+ 1). In all cases (with els of expression ~x:pseudotime (interaction) was compared to the nested model expression ~x + the exception of PhenoPath where differential expression testing is implicit in model inference), the mod- pseudotime (no interaction).

One common theme in many pseudotime inference algorithms is the emphasis on learning **a non-linear mani- fold embedded in the high dimensional space** (see e.g. [16, 17]) which could in theory decrease the accuracy of trajectories inferred with **PhenoPath that assumes linear changes** in expression over (pseudo-)time. Since the pseudotimes are always unobserved it is impossible to precisely quantify the true non-linearity of the “pseu- dotemporal manifold”.

However, we can fit trajectories and compare the results with those inferred using the “nonlinear” algorithms.

The trajectory inferred by pseudotime algorithms is obviously dependent on the genes used. A common approach is to select a subset of highly variable genes (HVGs) ([23]),

Across each dataset we PhenoPath was generally robust to the number of HVGs selected,

One of PhenoPath’s strengths is its ability to work with arbitrary design matrices incorporating binary, cat- egorical, or continuous covariates.

An alternative approach to that of the PhenoPath model is to **split the samples based on covariate-status and perform pseudotime inference separately on each** before combining the results post-hoc. Several downsides !

- examining each set of cells separately leads to smaller numbers of samples per test and therefore a reduction in power to detect interactions
- if the covariate is continuous one would have to resort to arbitrary binning of samples to perform such an analyses
- this becomes even more burdensome if multiple covariates are present (have to consider many different groups of samples, e.g. AC, AD, ... for 4 levels ABCD) and fit the pseudotimes separately for each
- the number of groups grows exponentially in the number of factors, meaning many groups will actually have few or no samples in it
- subtle issues arise with respect to biological interpretation. Upon splitting the samples, how do you know the inferred pseudotimes correspond to the same biological process?

BRCA Survival Analysis

We fitted a stratified (ER status) Cox proportional hazards model to the overall survival data for 720 TCGA BRCA patients with survival and expression data using patient age at onset and PhenoPath pseudotime as co- variates.

This survival analysis indicated that the model coefficient associated with the pseudotime contribu- tion was significant (p = 0.0032).

Analysis of deviance between nested models, with and without pseudotime as a covariate, indicated that the performance of the more complex model that includes pseudotime produces a better fit to the survival data p = 0.004124.

Proportional hazards tests and diagnostics based on weighted residuals was performed to confirm that the proportion hazards assumption was not violated.

Identifying crossover points in BRCA

Unless the gradient of change along the trajectory is exactly equal for both phenotypes (i.e. $\beta = 0$ exactly), the gene expression will cross at a given point in the trajectory.

Inference of this point would allow us **to identify sections of the trajectory not affected by the covariate and consequently sections of the trajectory that are**.

- if the crossover point occurs towards the **beginning** of the trajectory, it would mean **gene expression is similar at the beginning but diverges** as we move along the trajectory.
- if the crossover points occur towards the **end** of the trajectory, it would imply the expression profiles for the two phenotypes are **different at the beginning of the trajectory, but converge as the trajectory progresses**. An interpretation of this would be that **the effect on expression from the trajectory slowly dominates** over the effect of phenotypes on the trajectory.

It is important to note that **the latent trajectory values loosely follow a $N(0, 1)$ distribution**.

- the '**middle**' of the trajectory is any value around 0,
- values of -1 or less could be thought of as the '**beginning**'
- values greater than 1 may be thought of as the '**end**'.

Crucially, we can derive an analytical expression from the PhenoPath parameters for the *crossover point z^**.

The condition for the crossover point is that **the predicted expression for each phenotype is identical**.

Therefore (in the context of BRCA cancer)
$y_g^{ER_+}(z_g^*) = y_g^{ER_-}(z_g^*)$

which leads to the condition

$$\alpha_g x_{ER_+} + (c_g + \beta_g x_{ER_+}) z_g^* = \alpha_g x_{ER_-} + (c_g + \beta_g x_{ER_-}) z_g^*$$

which is in turn solved by
$z_g^* = \frac{\alpha_g}{\beta_g}$

We fitted the crossover points z^* for all significant genes in the BRCA dataset. We find that **the vast majority of the crossover times z^* occur towards the end** of the trajectory, with a median value of around 0.4.

$\rightarrow$ at the beginning of the trajectory most genes are differentially expressed based on ER status, while as the trajectory progresses it comes to dominate at the gene expression converges