GENE EXPRESSION NORMALIZATION FOR SPARSE OR HETEROGENOUS mRNA SEQUENCING DATA
and
TIME-COURSE STUDY OF EMBRIOLOGICAL DEVELOPMENTAL TIMING IN CHIMERIC TISSUES

by

Jared T. Brown

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Statistics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 23 August 2021

The dissertation is approved by the following members of the Final Oral Committee:

Christina Kendziorski, Professor, Biostatistics and Medical Informatics
Michael Newton, Chair, Biostatistics and Medical Informatics and Professor, Statistics
Anru Zhang, Assistant Professor, Statistics
Daifeng Wang, Assistant Professor, Biostatistics and Medical Informatics
Beth Drolet, Chair, Dermatology

Dedicated to Dr. Robb M. Thomson

our conversations inspired this path

and your poetry captured its beauty

thank you, Granddad

ACKNOWLEDGEMENTS

Many thanks need to go to my advisor, Christina Kendziorski, for all her help through this project. It was her careful guidance through collaborations that allowed me to grow into the scientist that I am today. In terms of my primary research, I need to particularly thank her for finding the perfect balancing point between encouraging me to explore new and exciting paths and helping me focus on the research problems of greatest relevance. I always found great wisdom in our conversations and (almost) always left refreshed and excited for the work to come.

Likewise, I owe a debt of gratitude to all the collaborators I have worked with during my time in Madison. In particular I need to thank Chris Barr. It was while working with Chris that I first cut my teeth on proper academic research. His patience and endless excitement for the work we were doing helped develop the passion for collaborative research I am bringing to the rest of my career.

Many thanks also the current and past members of the Kendziorski lab who I had the pleasure of working with. Their thoughtful critiques of my work helped significantly improve the end results and our lab meetings often inspired me to incorporate new approaches in my work. Not least, their constant support helped ground me with a feeling of belonging to something meaningful as we all pursued this challenging goal.

I would like to similarly thank the remaining members of my thesis committee, Michael Newton, Anru Zhang, Daifeng Wang, and Beth Drolet for their support, suggestions, and thoughtful critiques of this work. They have all helped improve my research and I am proud to follow in their footsteps.

Finally, and perhaps most importantly, many thanks to my family for their support these last six years. I always had in the back of my mind that all of you saw this project as a worthwhile and meaningful one, and that helped significantly in seeing it through to the end. In particular, I have to thank my wife, Tina, for your endless encouragement. Your support, even as I commandeered our living room as a Covid era office made this what it is.

Thank you all.

CONTENTS

LIST OF FIGURES

GENE EXPRESSION NORMALIZATION FOR SPARSE OR HETEROGENOUS mRNA SEQUENCING DATA

and

TIME-COURSE STUDY OF EMBRIOLOGICAL DEVELOPMENTAL TIMING IN CHIMERIC TISSUES

Jared T. Brown

Under the supervision of Professor Christina Kendziorski
at the University of Wisconsin-Madison

Abstract

Recent years have seen rapid development in the technologies used to sequence gene expression across the genome; automated systems have increased the speed with which multiple biological samples can be processed and droplet/microfluidics systems now allow the preparation/sequencing of thousands of single cells. The newly generated datasets thereby pair novel experimental designs with altered statistical properties of the underlying data. To address these changes, we develop two new normalization techniques and a novel analysis pipeline for time series data.

A critical component of gene expression analysis, normalization aims to remove nuisance variation from expression data. New, droplet/microfluidics-based systems can, however, challenge current approaches due to the corresponding sparsity in the data. Specifically, proportions of zeros

are dependent upon technical sources of variation. Existing methods typically assume an approximately continuous distributions of expression and so struggle to remove the dependency relationship between zeros and nuisance variation. To address this, we introduce Distributional Normalization (Dino), a normalization procedure which transforms the expression data by careful sampling from an estimated distribution conditioned on the observed expression levels.

Modern sequencing protocols – which typically process samples containing heterogenous cell-types – also have the potential to confound biological and nuisance variation. In particular, total expression levels of individual cells, the typical measure of nuisance variation, may differ across cell-types. As such, existing normalization techniques risk the removal of true biological variation. To address this, we introduce Robust Heterogeneity Integration Normalization (Rhino), a normalization method which aims to identify sets of genes with constant expression levels across cell-types and thereby normalize in a manner which is sensitive to biologically derived differences in total expression levels.

Finally, taking advantage of the improved automation of sample maintenance and preparation, we focus on a time-series experiment studying differential developmental timing between human, mouse, and chimeric mixed human/mouse stem cells. Our aim is to determine whether the rates of development can be altered through cellular signaling. We therefore develop an analysis pipeline to identify both differentially regulated genes over time and genes which appear accelerated/decelerated compared to control, thereby allowing us to characterize changes in the rates of development between experimental conditions.

1 Introduction

The experimental platforms facilitating RNA-sequencing (RNA-seq) experiments have advanced dramatically over the past decade. In a traditional, or *bulk*, RNA-seq experiment, the abundance, or *expression*, of mRNA molecules is assayed in a tissue sample for all genes in the genome, resulting in a functionally averaged measure of mRNA abundance across the individual cells in the sample. Hundreds of studies driving biological discoveries have been based on such bulk experiments and have thus increased the interest in RNA-sequencing protocols. Such discoveries in turn have fueled a virtuous cycle in which new technological developments have allowed hundreds more studies to occur, implementing previously impossible or excessively expensive measurements, and thereby assessing biological questions with greater sensitivity and precision (A. R. Wu *et al.*, 2017; Hwang *et al.*, 2018; Bacher and Kendziorski, 2016; Haque *et al.*, 2017; Kolodziejczyk *et al.*, 2015).

Of particular note is the maturation of microfluidics systems for single-cell RNA-seq (scRNA-seq) in which genome-wide gene expression is assayed individually for each cell in a biological sample. Technologies implementing this technique include Drop-seq (Macosko *et al.*, 2015), inDrop (Klein *et al.*, 2015), and perhaps most commonly observed in published studies, the commercially available 10X Chromium platform (Zheng *et al.*, 2017). Previous methods for scRNA-seq tended to treat individual cells as individual biological samples, requiring sorting, isolation, and processing of individual cells in a way that practically limited the number of single cells that could be simultaneously sequenced to the 10s or 100s. In contrast, microfluidics systems like the above perform single-cell isolation by forming an emulsion of water droplets in oil on the microfluidics chip and capturing single cells together with barcoded primers in the water droplets

as they flow through the system; a process which occurs many times per second. One of the particular contributions of these techniques is the novel addition of barcodes to the primers captured with individual cells which are unique to that water droplet and thus unique to the co-captured cell. These cell-specific barcodes on the primers allow downstream sequencing to treat the entire emulsion of droplets and cells as a single biological sample with sequenced reads being sorted after-the-fact into their source cells by the unique cellular barcode. In this way, and for the first time, it has become feasible to simultaneously sequence genome-wide gene expression in 1,000s to 100,000s of individual cells.

While these novel platforms are unquestionably both impressive in and of themselves as well as impactful for the scientific community, the ability to derive statistically meaningful results naturally remains predicated on the degree to which variations in gene expression are due to biological factors of interest. In practice, however, multiple sources of nuisance variation – including some technical sources particular to the sequencing platform – confound true, biologically derived differences between cells and experimental conditions.

Generally, the class of statistical techniques denoted as *normalization methods* aim to first identify such technical and/or nuisance variation and then transform the sequenced gene expression in order to remove nuisance variation prior to downstream analysis. A main contribution of this thesis concerns novel statistical methods and their corresponding software implementations to remove nuisance variation. The identification of particular types of nuisance variation and methods to address them are discussed in Chapters 2 & 3. We should also note that while characterizing normalization as the *identification and removal by mathematical transformation of nuisance variation* is the more common definition of *normalization methods* (and the one used in this document), some techniques alternately describe normalization as the

accounting for nuisance variation while simultaneously performing downstream analysis, typically by modeling such variation through a cell-specific covariate.

To date, many normalization techniques have been developed and published (several are described in Chapter 2) to correct for nuisance variation in RNA-seq datasets, some specifically for scRNA-seq data. Unfortunately, existing techniques often struggle when applied to modern scRNA-seq data generated from microfluidics platforms. Being statistical techniques, the success of these methods is naturally contingent on the model assumptions being made matching the statistical properties of the data they are applied to. Historically, bulk (as opposed to single-cell) sequencing has resulted in high expression levels for many, even a majority of genes in a dataset to the point that continuous distributions could feasibly be used for modeling purposes. Likewise, older single-cell techniques which sequenced on the order of 10s to 100s of cells at once demonstrated sufficiently high expression (if still much lower than bulk) that continuous models were successful. By contrast, the above introduced microfluidics systems often sequence similar total numbers of cDNA molecules, but must divide that sequencing budget across thousands, or tens of thousands, of cells. This in turn results in cell-specific estimates of expression which are not only lower than previous techniques, but also characteristically sparse.

This dramatic reduction in per-cell reads fundamentally alters the statistical properties of the data. In particular, continuous assumptions no longer hold, even approximately. The distributional failure can be seen most strongly when it is observed that genes normalized by existing methods are not fully independent of nuisance variation.

While there are multiple sources of variation affecting counts, the one that is of primary interest in most normalization efforts is the cell-specific value denoted as library size (LS or

occasionally LS_j to denote the specific LS value for cell *j*). A detailed definition of LS is given in Chapter 2. However, one can think of library size as the total sequenced gene expression within a cell. Ideally, this total could be controlled across cells so that expression levels between cells could be directly compared. This is not the case in practice, and differences in LS values between cells have been shown to dramatically confound sequenced expression levels and therefore require proper normalization to correct for this nuisance variation.

Existing methods for normalization typically ensure that some properties of normalized gene expression, such as the mean, are independent of LS. However, as we show in Chapter 2, the full expression distribution is not independent of LS; a fact which can lead to systematic errors, particularly large numbers of false positives, in downstream testing and analysis. Beyond this, many existing techniques struggle algorithmically as well as statistically. The high proportion of zeros across genes, often greater than 90% for a given cell (Townes *et al.*, 2019), results in algorithmic instability or outright failure in existing methods.

To solve these challenges, we developed a method titled Distributional Normalization, or Dino (Brown, Ni, *et al.*, 2021). Dino directly accommodates the discrete nature of scRNA-seq count data by modeling expression as a mixture of negative binomial (NB) distributions. As such, the high proportion of zeros typically present in modern scRNA-seq data can be directly accommodated by correspondingly low mixture component means. Further, the use of a mixture model accommodates deviations from a pure NB distribution, particularly in the case of multi-modal expression as might be the case for genes that define differences between cell types. To generate normalized expression, Dino deviates from existing methods that perform deterministic transformations of expression data. Instead, Dino normalizes by sampling from the posterior distribution of cell- and gene-specific means, thereby generating normalized expression for which

the full distribution is largely independent of nuisance variation. The full details behind the inspiration for, development of, and testing of the Dino method are the subject of Chapter 2.

Dino is specifically designed to be robust to variations in the input data. As mentioned above, the choice of a mixture distribution accommodates both heterogenous cell populations as well as a certain (sometimes large) degree of uncertainty in the true underlying distribution of the data. To expand on this robustness, Dino further allows the user to provide alternate cell-specific size-factor estimates as the covariate measure of nuisance variation, overriding the default calculation of LS as the measure of nuisance variation. However, Dino itself is unable to compute alternate estimates of nuisance variation.

In fact, most existing normalization methods do not estimate a measure of nuisance variation, relying instead on LS to be a valid – and fast to compute – estimate. While the many successful studies employing such methods show that this can be true, at least approximately, such is not always the case. Generally, any time that LS is correlated with – or otherwise dependent upon (i.e., if the dependency relationship is non-linear) – the biological variability of interest, normalization to remove the effects of LS will compromise downstream analysis. At best, this will take the form of the removal of both technical and some biological variability, reducing the power to detect true differences between groups of cells. At worst, however, such events can artificially generate the appearance of differences where none exist in the actual tissues under study.

The normalization methods which estimate size-factor measures of nuisance variation other than LS acknowledge this fact. Most commonly they make implicit biological assumptions such as: the “median” gene is equivalently expressed across all tissues, and so expected (average) expression for this gene should be constant between all subsets of cells in the normalized data

(quotes because "median" takes a more complicated definition in the actual algorithms). The understanding of the importance of clarifying biological assumptions dates back at least to 2010 when Robinson and Oshlack, while developing their normalization method Trimmed Mean of M-Values (TMM), observed that simultaneously normalizing RNA-seq data from kidney and liver cells using LS as the measure of nuisance variation systematically biased testing between the two tissues. In particular, the liver cells, which presumably upregulated more genes compared to the kidneys, appeared systematically downregulated due to the division by too large of a size-factor (when LS was used as that size-factor), resulting in a panel of housekeeping genes appearing differentially expressed (Robinson and Oshlack, 2010). Despite this, the majority of modern studies still use normalization methods based upon LS.

To address this, we are developing a novel normalization method called Robust Heterogeneity Integration Normalization, or Rhino. Full details of the inspiration for, methods of, and preliminary testing of the Rhino method are the subject of Chapter 3. While Rhino aims to improve several aspects of the normalization pipeline, its primary purpose is the robust estimation of improved size-factor estimates of nuisance variation. To this end, Rhino is based around the underlying biological assumption that, between any two cells (or samples), there is some non-empty set of genes which are equivalently expressed, and as such, post normalization, the expected (average) expression of these genes should be constant between those two cells. The primary difference between Rhino and previous methods which made similar assumptions is two-fold. First, the set of equivalently expressed genes may change between different pairs of cells, and so optimal estimates of nuisance variation should use all such genes, both for improved accuracy *and* improved precision of estimation. Second, the identification of such genes is model based,

avoiding heuristics such as *the median gene is equivalently expressed* or *the 5% of genes with the greatest absolute log-fold change are differentially expressed*.

While both of great significance to the field as well as forming the foundation upon which much of this dissertation is based, the technologies of microfluidics are far from the only impactful technological developments of recent years. The increasing speed and automation by which cells and tissues can be processed and sequenced has also impacted the types of studies that can be conducted. Of particular interest is the increase in time-series genomics studies being conducted, many of which remain better suited to bulk RNA-seq. It is in this context that we transition into the final chapter of this dissertation.

One of the great potentials of modern treatments is the development of regenerative medicine by which grown cells, tissues, and even full organs could be used for therapeutic effect. However, many of the more dramatic promises of this field of study remain unrealized, partially due to the difficulty and complexity inherent in the attempt to grow in-vitro tissues and organs. While the use of pluripotent stem cells offers to mitigate many of the outstanding challenges, it is curious to note that stem-cells grown in-vitro still differentiate and develop at about the same speed as the species from which they were derived, despite the absence of maternal chemical signaling (Barry *et al.*, 2017; Kanton *et al.*, 2019; Espuny-Camacho *et al.*, 2013; Maroof *et al.*, 2013; Gaspard *et al.*, 2008; Pollen *et al.*, 2019; Nicholas *et al.*, 2013). This suggests the existence of an "intrinsic developmental clock." Given the observation that pluripotent mouse stem cells develop in accordance with the 20-day mouse gestation time while pluripotent human stem cells develop in accordance with the 9-month human gestation time, Dr. Christopher Barry and I set out to determine whether, and if so, to what degree and in which way factors that are active during the

more rapid murine development could accelerate the development of human cells (Brown, Barry, *et al.*, 2021).

In particular, the lack of maternal influence on the developmental clock suggests that differences in developmental timing are the result of differences in the timing of cell-to-cell signaling events. We therefore directly test this hypothesis by co-culturing mouse and human stem cells such that the human cells are exposed to both ambient and direct contact signaling from the more rapidly developing mouse cells. Beyond answering the question of whether developmental rates can be moderated by external signaling factors, the contribution of this thesis is also the development of an analysis pipeline to handle these data.

Previous methods exist to aid in the analysis of time-series genomic data. However, such methods are typically restricted to the identification of dynamic or differentially regulated genes over time between conditions. For our purposes, it is not sufficient to identify differential regulation. Rather, in order to address questions of developmental *rate*, we were required to characterize the relative acceleration or deceleration of genes expression trajectories between conditions and then summarize those measures across the genome to identify systems of accelerated or decelerated activity, and ultimately assess global differences in developmental rate. We approached this task by leveraging multiple statistical modalities. These approaches and the findings of our study are the subject of Chapter 4.

2 Normalizing full expression distributions with Dino

2.1 Background

Normalization is a critical first step in the analysis of RNA-seq data. We define normalization here as a procedure which transforms sequenced expression levels to remove nuisance variation, possibly calculating a measure of nuisance variation as a first step. As such, proper normalization removes the dependency relationship between expression and technical or nuisance artifacts. In contrast and as a preview to the motivation for this chapter, unsuccessful normalization may result in transformed expression values which retain some degree of dependence on the measured nuisance variation.

The simplest and, perhaps, the most common measure of nuisance variation in sequencing experiments is the cell-specific library size (LS or LS_j for the specific value associated with cell j). LS is simply defined as the cell-specific sum of expression across the genome. In the usual representation of gene expression as a $G \times J$ integer matrix with G-many genes on the rows and J-many cells on the columns, LS is calculated as the J column sums. Estimated this way, LS and gene sequenced expression levels are typically linearly related. This relationship is further theoretically reinforced by the observation that, if the number of reads sequenced for a given cell increases (LS increases), then the expected number reads aligning to any given gene should increase proportionately; i.e., if total reads double, so too does the expected (sequenced) expression level for each gene.

Many existing methods leverage this simple relationship. The most computationally rapid and possibly the most commonly used are based on scale-factor transformations of the expression

data. In such approaches, a scale factor is calculated for each cell and the expression level for each gene in that cell is divided by that same scale factor. In Counts Per Ten thousand (CPT), a version of which is implemented in the Seurat analysis pipeline (Butler *et al.*, 2018), scale-factors are calculated by dividing LS_j by ten thousand, thus resulting in normalized expression which sums to ten thousand for each cell. The older Counts Per Million (CPM) (Law *et al.*, 2014) is similar, but scales cell-specific expression to one million.

Other methods leverage the expected linear relationship between LS and expression in a similar way. Like CPT and CPM, Scran (Lun *et al.*, 2016) transforms sequenced expression through the use of scale factors. In this case, however, scale-factors are estimated based on a modified version of the Median-Ratio (MR) algorithm (Anders and Huber, 2010) applied across pools of cells. Being based on MR, Scran aims to be more robust to potential sources of bias than is the case with LS which is only valid to the extent that differences in LS between cells are due to technical, rather than biological factors. Further, by pooling expression across cells when estimating size factors, Scran aims to be more robust to the zeros often present in the pre-microfluidics scRNA-seq that were becoming common when the method was developed, an algorithmic constraint of MR.

Around the same time Scran was developed, Bacher *et al*. (Bacher *et al.*, 2017) demonstrated that single-cell normalization could be improved by the adoption of gene-specific scale-factors, indicating that the common use of *global* scale factors could compromise performance. In their method, scNorm, scale-factors are computed for groups of genes with similar relationships (slopes) between expression and LS through the use of quantile regression with LS as a model covariate. Their method, however, was developed for pre-microfluidics platforms, such as the protocols implemented in the Fluidigm products, and so does not directly apply to the

expression values being considered here. In particular, in the FAQ section of the GitHub repository and Bioconductor vignette for scNorm, it is noted that, for datasets with more than about 80% zeros, the model parameters may fail to converge. By contrast, typical microfluidics-based datasets often have greater than 90% zeros (Townes *et al.*, 2019).

This dramatic difference in dataset sparsity is driven by at least two factors. As mentioned in Chapter 1, microfluidics protocols treat the entire emulsion of cells captured in water droplets suspended in oil as a single sequencing sample whereas previous single-cell techniques treated each individual cell as a single sequencing sample. The method by which the emulsion is generated allows the rapid preparation of 1000s of cells simultaneously for sequencing, with unique, droplet-specific, and thus cell-specific barcodes on the polydT primers allowing for reads originating from a given cell to be identified and sorted after sequencing. However, by dividing the available sequencing budget (the total number of reads which can be sequenced) for a given sequencing sample among 1000s of cells, the resulting LS for any given cell is correspondingly reduced, even for deep-sequencing applications.

The second factor driving the sparsity of microfluidics expression estimates is the use of unique molecular identifiers, or UMIs. Closely related to the cell-specific barcodes, UMIs are a polydT primer-specific barcode. The advantage when using UMIs is that technical copies of a given mRNA molecule can be identified and removed following sequencing. Such copies typically arise when the concentration of cDNA molecules in the sequencing library is increased by polymerase chain reaction (PCR) amplification, copying the extant cDNA molecules, sometimes many fold. The use of PCR is generally necessary to boost cDNA densities to levels that can be reasonably detected, but risks the event where multiple reads in the final sequencing dataset are derived from the exact same mRNA molecule in the cell under study. By tagging each mRNA

molecule prior to the construction of the PCR amplified cDNA library, such duplicates can be identified following sequencing, and just one copy retained for analysis during "de-duplication". This practice has the benefit of dramatically reducing biases in observed gene expression due to PCR artifacts and transcript length (Tung *et al.*, 2017; Grün *et al.*, 2014; Islam *et al.*, 2014; Zheng *et al.*, 2017). Further, UMI-based data appears to be better behaved than previous types, closely following a Poisson distribution of counts with means scaled by LS, rather than the zero-inflated distributions of older data (Svensson, 2020). However, the de-duplication of UMIs further reduces the per-cell LS, and therefore increases the observed sparsity of the data. Note that in the following text, we use UMI to refer to both the molecular tag described above as well as the resulting integer counts in the corresponding expression matrix. Whenever the intended use is not clear from context, we specify which is meant.

More recently, Hafemeister and Satija developed scTransform which, to our knowledge, represents the first normalization method specifically targeted at UMI count data (Hafemeister and Satija, 2019). In their approach it is observed, as with scNorm, that genes typically required unique model parameters and, also like scNorm, these parameters are similar for genes with similar expression levels. In this case, however, the unique parameters generally do not correspond to different slopes between expression levels and LS, although such differences are allowed, but rather to gene-specific overdispersion parameters. Specifically, scTransform approaches normalization as a parametric regression problem wherein parameter estimates are smoothed via spline regression across genes to improve robustness to overfitting. Following parameter estimation, normalized expression is derived by calculating the cell- and gene-specific Pearson residuals from the model. In this way, scTransform aims to remove the effects on both expression mean and variance of technical variation in LS.

With the significant amount of work in this area, it might seem as though this were a field in which the potential gains of novel approaches would be more marginal than meaningful. However, for both the above methods and others, the normalized expression from UMI datasets remains unfortunately dependent upon LS. Specifically, even though existing approaches may remove the dependency relationship from certain characteristics of the expression distribution – all methods seem to reasonably well remove the relationship between *average* expression and LS for example – the normalized distributions themselves remain dependent upon nuisance variation as measured by LS for UMI count data.

The most dramatic way in which to observe this effect is to consider the zeros in the expression matrix; gene-cell pairs for which no UMIs were sequenced. For a given gene, the proportion of zeros is naturally correlated with LS; as LS increases, the proportion of zeros decreases, or equivalently, the average expression increases. Under a scale factor approach (which includes most of the above), however, a pre-normalization zero directly translates into a post-normalization zero, and thus the proportion of zeros remains dependent upon LS in the normalized expression data. Likewise, in order to remove the dependency of average expression on LS (as most methods do), the relatively rare non-zeros among the low-LS cells must be normalized to higher values than the more common non-zeros among the high-LS cells. Both of these effects are demonstrated in Figure 2.2.

To solve this problem, that is, to normalize in a way that removes the dependency between LS and gene expression from the full distribution of expression, we developed Dino. Dino is a model-based normalization method, like scNorm and scTransform, in that it constructs, for each gene, a parameterized model of expression. The Dino method, however, possesses two unique attributes which set it apart in both function and results. First, the Dino model is comprised of a

mixture of Negative Binomial components, thereby allowing it to accommodate the potentially high levels of cell-type heterogeneity which can be present in these data. Second, Dino normalizes by resampling from the posterior distribution of the cell- and gene-specific estimated mean expression values, thereby removing the dependency of the normalized distribution on LS. Focusing on the zeros as the most dramatic example, by normalizing by resampling, low LS zeros can be sampled to values other than zero, and high LS, low count values can be sampled to zero, thereby equalizing the proportion of zeros in LS following normalization.

2.2 Materials and methods

2.2.1 The Dino model

As mentioned above, Dino normalizes expression data by resampling from the posterior distribution of cell- and gene-specific estimated mean expression values *conditional* on observed UMI counts and LS, rather than directly transforming the UMI counts by some function of LS. As such, Dino aims to estimate cell- and gene-specific mean parameters λ_{gj}. To formalize the notation, let the sequenced expression data be stored as a count matrix Y with genes on the rows and cells on the columns. Then, for G-many genes and J-many cells, Y is of dimension $G \times J$ and the integer element y_{gj} denotes the observed UMIs for gene g in cell j.

Being count data and, in particular, being count data derived from an experimental procedure which closely mimics random sampling of unique elements (referring to the UMIs) from a large pool, the counts y_{gj} lend themselves naturally to a Generalized Linear Model (GLM). Not only is there this theoretical bases for the choice of a GLM, but previous methods have successfully modeled sequencing count data by Negative Binomial (NB) or Poisson distributions (Anders and Huber, 2010; Hafemeister and Satija, 2019; Townes *et al.*, 2019). Beyond this, and as discussed

above, the model means, λ_{gj}, are expected to scale linearly with LS, suggesting a log-linear model given an appropriate transformation of LS (Anders and Huber, 2010; Lun *et al.*, 2016; Hafemeister and Satija, 2019; Townes *et al.*, 2019). Given this, we might consider a model of counts y_{gj} as

$$y_{gj} \sim f^P\left(\lambda_{gj}\delta_j\right);\ \delta_j \coloneqq \mathrm{LS_j}/c$$

where f^P is a Poisson mass function, δ_j is the scaled LS, and c is a centering parameter (by default, the median LS across cells).

Such a model, while useful for setting up the Dino method, is limited in two ways. First, it may be that the counts y_{gj} are over-dispersed with respect to the Poisson distribution as has been the case for previous count models of sequencing data, although there is some evidence that the use of UMIs has at least reduced and possibly removed such overdispersion (Svensson, 2020). Second, and more importantly, the presence of heterogenous sub-populations of cell types may render the underlying distribution of counts neither a Poisson nor a NB distribution; rather, the underlying distribution may be multimodal. When considering high-throughput, single-cell sequencing, the existence of multiple sub-populations of cells is common if not generally expected.

To accommodate both concerns and to facilitate the forthcoming calculation of the posterior distribution on the λ_{gj}, we impose a hierarchical mixture of Gammas distribution on the λ_{gj}. Ignoring for the moment the *mixture* part of this formulation, the use of a Gamma component in this way is reassuring in that the resultant marginal distribution on the counts y_{gj} is then NB, the same model assumption from scTransform (Hafemeister and Satija, 2019), DESeq2 (Love *et al.*, 2014), ZINB-WaVE (Risso *et al.*, 2018), and SAVER (Huang *et al.*, 2018). The use of a *mixture* of Gammas thus alters the model only to the extent that the corresponding marginal distribution of the counts y_{gj} is then a mixture of NB distributions.

Noting that the Dino model is fit independently to each gene, the subscript g can then be dropped in the full Dino model:

$$y_j \sim f^P(\lambda_j \delta_j)$$

$$\lambda_j \sim \sum_K \pi_k f^G\left(\frac{\mu_k}{\theta}, \theta\right)$$

where K is the number of mixture components, π_k is the weight of component k such that $\Sigma_k \pi_k = 1$, f^G is a gamma density using the shape and scale parameterization, μ_k is the mean of the corresponding Gamma component, and θ is a gene-specific scale term.

The details of how the parameters of the mixture model are estimated (a modified expectation maximization algorithm) are left until the next section. However, following estimation, the hierarchical model of counts results in a simple posterior distribution. In particular, the Gamma distribution is a conjugate prior to the Poisson, and so the posterior distribution on the means λ_j can be expressed as

$$\mathbb{P}(\lambda_j | y_j, \delta_j) \propto f^P(\lambda_j \delta_j) \sum_K \pi_k f^G\left(\frac{\mu_k}{\theta}, \theta\right)$$

which, after a slight modification, reduces to a mixture of Gamma distributions on λ_j

$$\mathbb{P}(\lambda_j | y_j, \delta_j) = \sum_K \tau_{kj} f^G\left(\frac{\mu_k}{\theta} + \gamma y_j, \frac{1}{\left(\frac{1}{\theta} + \gamma \delta_j\right)}\right)$$

where τ_{kj} denotes the conditional likelihood that mean λ_j derives from mixture component k given the observed data and γ (whose inclusion is the above mentioned "slight modification") is a concentration parameter. Testing has shown that sampling normalized values from the unaltered posterior distribution can result in excessive variance in the normalized data, reducing the power

to test for biological variation while (default) values of $\gamma = 15$ were more successful across a variety of datasets (Supplemental Figures A1-3). The concentration parameter, γ, can alternately be seen as a type of regularizer, pulling normalized values (samples from the posterior distribution) towards their scale-factor variant. In fact, in the limit as γ increases, the normalized values converge in probability to y_j / δ_j.

2.2.2 Parameter initialization and estimation

In order to accelerate calculations, the Dino model approximates the above mixture of NBs when estimating parameters. In particular, while the mixture of NB distributions does not have a clean, closed-form update to the Expectation-Maximization (EM) algorithm, a similar mixture of Poissons does. For this reason, Dino takes a two-step approach to parameter estimation. First, Dino approximates the data as a mixture of Poissons to fit the mixture component means. Second, with mean parameters fixed and with cell-specific component-membership likelihoods determined by the EM algorithm, Dino estimates the hierarchical Gamma distribution parameters – thereby returning the model to a mixture of NB distributions – in a manner aimed at matching the Empirical Bayes prior distribution on λ_j to the observed distribution of UMIs.

It is for this reason – matching the distribution on λ_j to the distribution of the UMIs – that the parameter choice for the number of mixture components, K, is chosen as the minimum of 100 (default) and the (rounded) square-root of the number of y_j which are greater than zero (there is a lower bound of at least 2 components). This *over-parameterization* is therefore clearly not intended to estimate the true number of clusters/modes which may exist in the expression distribution for a particular gene. For example, if the true underlying distribution is a mixture of two NBs, the choice of K will still be (much) larger for all but the lowest expressing genes. Instead,

this approach should be understood as a parametric approximation to a non-parametric model. This approach is inspired by the methods in the ash model for False Discovery Rate (FDR) correction by Stephens *et al.* 2017 in which it was shown that any arbitrary unimodal distribution could be approximated by a mixture of uniform distributions, with infinitely many components as "a non-parametric limit" (Stephens, 2017). As a similar approach, Dino uses an arbitrarily large number of components to approximate the unknown, underlying expression distribution of the UMI counts. Similar results surrounding the use of large mixture models were demonstrated even earlier by Cordy and Thomas in their work on distributional deconvolution (Cordy and Thomas, 1997).

Figure 2.1 demonstrates this effect where the empirical densities of UMI counts for three genes from experimentally derived cells (subsetting cells in a small neighborhood of the median LS) follow complex, multimodal distributions (blue). A simple, unimodal NB regression on the data fails to capture this distribution (green), while the estimated distribution of the λ_j from the Dino model (red), which is functionally a mixture of K Gamma distributions, much more accurately captures the underlying distribution. Similar results can be seen for simulated genes (Supplemental Figures A4-5).

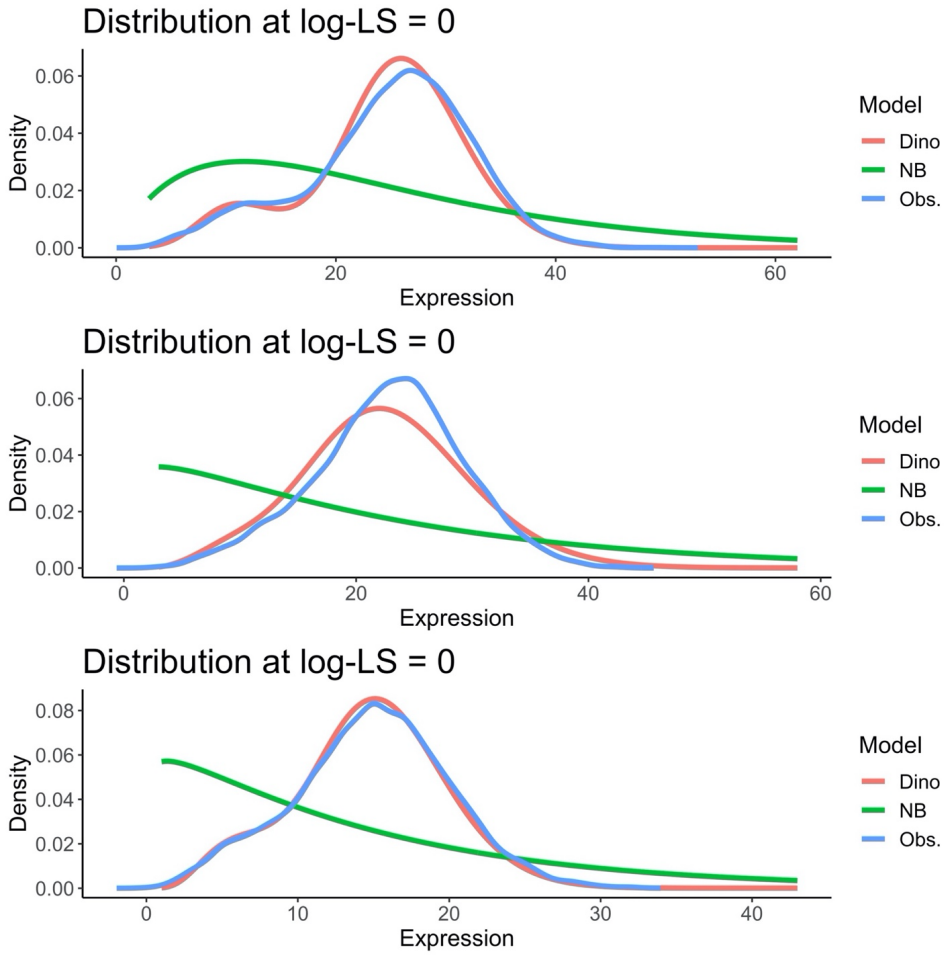


Figure 2.1: **A single Negative Binomial distribution fails to capture the heterogeneity of observed expression.**

For each of three different genes (RPL13, RPL10, RPS4X, top to bottom) from the PBMC68K_Pure dataset we plotted the empirical density of observed UMI counts (blue) from cells with scaled log-LS between -0.5 and 0.5. These curves were overlayed by the distribution (at log-LS=0) implied by the fitted parameters from a single Negative Binomial GLM (green) and the prior distribution estimated by Dino (red) used for resampling normalized expression.

With K fixed, the means, μ_k, in the mixture of Poissons approximation to the mixture of NB can be initialized. Aiming to speed model fitting with a robust parameter initialization, the μ_k are sampled from K evenly spaced points from an Empirical Cumulative Distribution Function (eCDF) of the underlying UMI counts at LS $\delta = 1$. Since this underlying distribution is complicated both by its confounding with LS and by a censoring threshold – values less than zero cannot be

observed – we developed a fast estimation procedure based on the inversion of Censored Least Absolute Deviations (cLAD) regression, also called censored quantile regression (Powell, 1984, 1986; Branham, R. L., 1982). This approach, originally described in the supplement to Brown *et al*. 2021 is presented in Appendix A.1 (Brown, Ni, *et al.*, 2021).

Following initialization, the mixture of Poissons model is estimated by an accelerated EM algorithm. As mentioned above, the mixture of Poissons model has simple, closed-form updates to the EM iterations. Model convergence is then further accelerated by the incorporation of quasi-Newton corrections to the EM updates. Letting the usual EM update be denoted by

$$\Sigma_{t+1} = g(\Sigma_t);\ \Sigma_t \coloneqq [\pi_t, \mu_t]$$

the corrected update for a step length of 1 is

$$\Sigma_{t+1} = g(\Sigma_t) - S * gl(\Sigma_t)$$

where S is an approximate inverse Hessian matrix based on the Broyden-Fletcher-Goldfarb-Shanno SR2 update, and $gl(\Sigma_t)$ is the likelihood gradient at the given parameters (Jamshidian and Jennrich, 1997). In practice, step lengths of 1 are attempted and adjusted if necessary to conform to the strong Wolfe conditions.

Having estimated the μ_k, and, as a consequence of the EM algorithm also the τ_{kj}, the parameters of the mixture of Gamma distributions can be estimated to reformulate the counts model into the mixture of NBs described above. Specifically, we set $\mu_k \coloneqq \lambda_k$ and estimate θ from an application of kernel density estimation. In particular, given a sample of (strictly positive) data, such as the μ_k, the underlying distribution of these data can be estimated as a mixture of Gamma distributions with shape parameters $\mu_k \,/\, \theta$ and scale parameter θ (Chen, 2000). In our case, θ is estimated using a kernel bandwidth estimator applied to the μ_k which is then trimmed to a maximum of 1. The restriction of the parameter

θ is designed to limit the variance of the Gamma components to not be greater than the variance of the Poisson components from which the means μ_k were derived.

2.2.3 Datasets used for validation

Dino was tested using a variety of both experimentally derived and simulated datasets. On the experimental side, six publicly available datasets were analyzed. Five of these datasets were derived from the data published by 10X genomics, including PBMC68K_Pure, PBMC5K_Prot, MaltTumor10K, MouseBrain, and PBMC68K while one, EMT, was derived from a third-party study.

All experimental datasets were tested, but PBMC68K_Pure is primarily used for displaying the following results. This dataset represents a set of peripheral blood mononuclear cells (PBMCs) which were first purified into known cell-types, after which each purified sample was individually sequenced on the 10X Genomics Chromium platform (their primary UMI-based microfluidics system) (Zheng *et al.*, 2017). This procedure allows individual cells to be annotated with specific cell-types, further allowing cell type-specific analysis and validation. To further improve the accuracy of validation based on such cell-type annotations, only the six cell types for which the tSNE plots of expression (van der Maaten and Hinton, 2008; Van Der Maaten, 2014) remained homogenous (no sub-groups) were retained: CD4+ T Helper, CD4+/CD25 T Reg, CD4+/CD45RA+/CD25- Naïve T, CD4+/CD45RO+ Memory, CD56+ NK, and CD8+/CD45RA+ Naive Cytotoxic. These cell types are identified by the original authors as demonstrating little sub-structure.

EMT is a dataset constructed of about 5,000 cells undergoing an induced epithelial to mesenchymal transition (McFaline-Figueroa *et al.*, 2019). The cells were grown in culture in such

a manner that cells on the exterior were preferentially induced to transition. As such, cells sampled and sequenced from the center of the tissue culture were expected to be predominantly epithelial while cells sampled and sequenced from the outer sections of the culture were expected to be primarily mesenchymal. This structure produces a dataset which spans the full differentiation pathway and which further can be compared to the source tissue region (inner vs. outer regions) to validate or to anchor analysis techniques such as pseudo-time ordering. The authors further describe eight gene sets derived from the Hallmark collection of gene sets (Liberzon *et al.*, 2015) which they consider to be characteristic of the cellular dynamics being induced. These eight gene sets were used here as validation.

The remaining datasets used for validation are considered in supplementary materials. Full details on the above-described datasets, PBMC68K_Pure and EMT, as well as the 4 remaining experimental datasets are described in appendix A.2.

Dino was also validated against simulated data. As with all simulated data, the aim was to capture as faithfully as possible the underlying distributional structure of experimentally derived UMI counts. In this case, the intended testing environment was a test of Differential Expression (DE) vs. Equivalent Expression (EE) where an EE gene is one for which average expression is constant between the two testing groups and a DE gene is one for which some difference in average expression exists. Note that these definitions, particularly that of DE, can be test specific; e.g., using a Wilcoxon Rank Sum (Wilcoxon) test implies a null hypothesis that EE genes are those with equivalent distributions, not just equivalent means while a likelihood ratio test of nested models implies that EE genes are those following the reduced model. In cases in the following where the definition of DE differs from a simple difference in latent average expression, the specific test and, if necessary, the fuller alternate definitions of EE and DE are specified.

Given these constraints, it was chosen to simulate UMI counts by careful down-sampling of experimentally derived expression data. Down-sampling has the advantage of not imposing an assumed expression distribution on the experimentally derived counts while careful choice of the (expected) proportion of UMIs to be retained for each gene following down-sampling allows one to control which genes are EE or DE, as well as the degree of fold-change in the DE genes, between two simulated groups. Full simulation details are provided in appendix A.3. In brief, however, simulated cells are generated as "cluster-pairs" where cells in each of the two sub-clusters in the cluster-pair differ on average in LS and where the genes which are EE/DE between each sub-cluster are known by construction. This is achieved by first taking two experimentally sequenced and transcriptionally similar cells and summing their UMI counts to generate a pooled pseudo-cell. This pseudo-cell is then down-sampled twice to generate one simulated cell in each sub-cluster of a cluster-pair. Systematic differences in the sampling proportion – p from a Binomial distribution – induce an expected difference in LS between the two simulated cells while further alteration of the sampling proportion for individual genes generates known DE genes. Following this procedure, multiple cluster-pairs are simulated to generate an aggregate dataset whose transcriptional heterogeneity mimics that of experimental data. Performance testing is only conducted between sub-clusters in a cluster-pair as the true EE/DE genes are not known between cluster-pairs.

2.3 Results

As previously discussed, both scale factor methods such as CPT and Scran as well as model-based methods such as scTransform can suffer from distributional dependencies on LS even after normalization. While such normalization successfully removes the effects of LS on average expression (within a cell type), and even the effects of LS on expression variance as in scTransform,

there are clear differences in the overall distributions of expression between cells in the low-LS (5-25% of LS) and high-LS groups (75-95% of LS) (Figure 2.2 A-B). Further, it can be shown that these results are not unique to the gene plotted. By constructing modified qq-plots where, for each gene, quantiles of expression in the low-LS group (y-axis) are plotted against quantiles of expression in the high-LS group (x-axis), distributions of expression can be aggregated across genes as a density heatmap. Were it the case that distributions were equal between these groups, at least for a significant proportion of genes, then the region of high density would follow the diagonal of the plot; as is the case for a typical qq-plot. However, significant off-diagonal regions of high density can clearly be observed for Scran and scTransform (Figure 2.2 C). In contrast, the resampling procedure implemented in Dino successfully removes the dependency relationship between LS and the full, normalized expression distribution across genes (Figure 2.2).

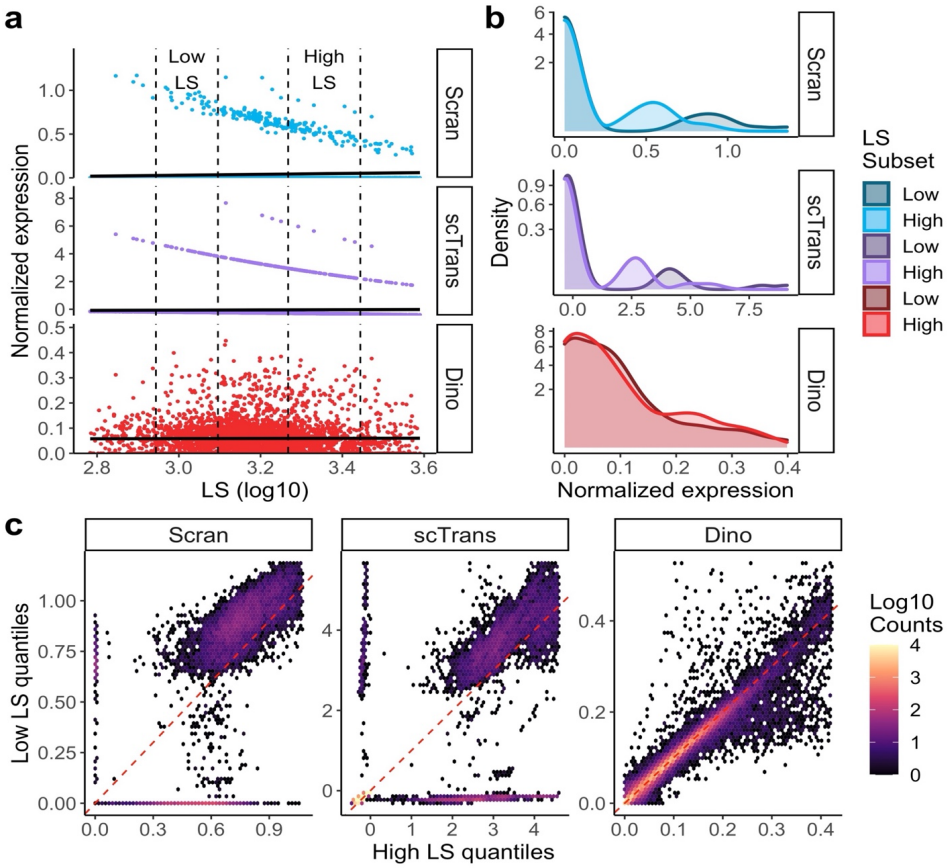


Figure 2.2: **Evaluation of gene-specific expression distributions following normalization.**

Expression data in the PBMC68K_Pure dataset were normalized by Scran, scTransform, and Dino. a)-b) Normalized expression is shown here for a homogeneous set of cells (CD4+/CD45RO+ memory cells) to minimize the effects of cell subpopulation heterogeneity. a) Normalized expression from a typical gene (NME1) under Scran, scTransform, and Dino plotted against LS. Fitted regression lines (solid black) show generally constant means across methods. Low-LS (5%-25% of LS) and high-LS (75%-95% of LS) subsets of cells are indicated by dashed lines and are used in the following panels. b) Density plots of normalized expression from low-LS and high-LS cells show that the constant mean is maintained by balancing the changing proportions of zeros, or near zeros in the case of scTransform, with expression shifts in normalized non-zeros. c) Quantile-quantile heatmaps compare normalized expression quantiles in the high-LS (x-coordinate) and low-LS cells (y-coordinate) across genes and cell-type annotations (S4.3 Section). As in panel b, there are systematic shifts in the distributions.

While these results are theoretically interesting, suggesting at least statistical limitations of existing methods, these initial results do not demonstrate practical problems for downstream analysis following the use of existing methods. To address this, we conducted a test of differential expression between cells in the low-LS and high-LS groups for each cell type in the

PBMC68K_Pure dataset. Using the default Wilcoxon rank sum test from the Seurat analysis pipeline (Butler *et al.*, 2018), we observed large numbers of genes with significant FDR corrected p-values (Figure 2.3 A). As tests were conducted within cell type annotations, the majority, if not all of these significant tests are expected to be false positives. This is not unexpected as the Wilcoxon test, as well as other single-cell specific tests such as MAST (Finak *et al.*, 2015), are sensitive to differences in the distribution of expression. Dino, by normalizing the full distribution for effects of LS, removes this dependency resulting in an order of magnitude fewer false positives (Figure 2.3 A).

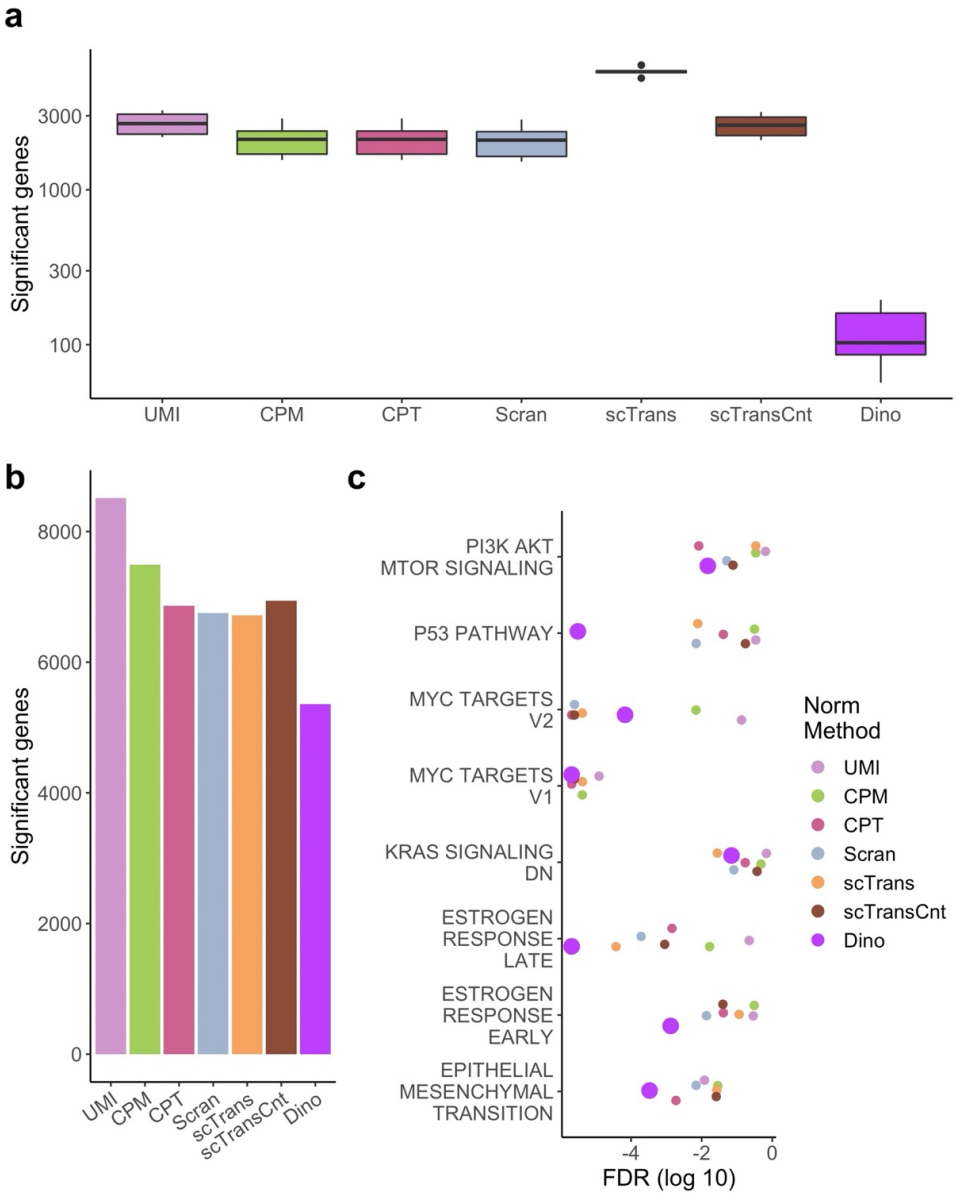


Figure 2.3: **The effects of normalization on downstream DE and enrichment analysis.**

a) Expression data from the PBMC68K_Pure dataset were normalized and genes were tested for DE using a Wilcoxon rank sum test between low-LS and high-LS cells (5%-25% and 75%-95% of LS respectively) within cell-type annotations. Box plots show numbers of significant genes. Given that cells only differ in LS, significant results are considered false positives. b) Expression data from the EMT dataset were analyzed using Monocle2 to identify genes with significantly variable expression over pseudo-time. Total numbers of significant genes are shown in a bar plot. c) Significance values of Hallmark terms enriched for DE genes from the EMT dataset, colored for each normalization method, are plotted for the subset of terms previously identified as defining expression shifts during epithelial to mesenchymal transition.

In order to further validate the performance of Dino, and importantly to demonstrate that the reduction in false discoveries is not due to a reduction in overall power, we performed a similar analysis on the EMT dataset (McFaline-Figueroa *et al.*, 2019). Following the analysis pipeline of the original authors of this data, we performed pseudo-time ordering of the EMT cells which aims to place individual cells along an estimated differentiation path (Trapnell *et al.*, 2014; Qiu *et al.*, 2017). In our case, we would expect the epithelial cells to be mapped to early pseudo-times corresponding to their position at the start of the differentiation path and mesenchymal cells to be mapped to late pseudo-times. As with the original authors of the data, we followed the pseudo-time ordering of the cells with a test for DE. In this case, however, the test aimed to identify gene with statistically significant variation as a function of pseudo-time, which, in turn, was considered a proxy measure for the position of any given cell on the transition between epithelial and mesenchymal cells. To identify these genes associated with the transition captured by the data, DE was tested through a likelihood ratio test between a smooth spline model of expression against pseudo-time and an intercept-only null model of expression against time. As before, data normalized by Dino resulted in the fewest significant genes, but, unlike before, there were still large numbers of genes detected as significant under Dino (around 5,000), consistent with a reduction in FDR, not a reduction in power (Figure 2.3 B).

To confirm this result, we performed gene-set enrichment on the gene significance values for tests based on each normalization method against the Hallmark collection of gene sets (Liberzon *et al.*, 2015). Highlighting the eight gene-sets originally identified by McFaline-Figueroa *et al.* as being characteristic of the epithelial to mesenchymal transition, it can be observed that Dino normalized data results in competitive significance in four out of the eight gene-sets and the greatest significance (corresponding to the greatest statistical power in DE

testing) in the remaining four. Of particular interest, the gene set denoted by the curators of the Hallmark collection as specifically pertaining to the *epithelial to mesenchymal transition*, is one of these last four for which Dino normalized data is close to an order of magnitude more significant (p-adj = 3.3e-4) than the next most significant alternative method (CPT, p-adj = 1.9e-3).

To expand testing beyond this individual dataset and thus show a general reduction in false discoveries and maintenance of testing power, we applied Dino to several simulated datasets which were, in turn, derived from several published datasets of UMI counts, including a simulation based on the above mentioned PBMC68K_Pure dataset. As datasets were constructed in such a way that the set of DE and EE genes is known between sub-clusters in simulated cluster-pairs, testing for DE could be conducted against a known ground truth. Therefore, we normalized the simulated data according to a panel of methods and, following normalization, conducted tests for DE using the Wilcoxon rank sum test. Receiver operator curves (ROCs) show that test results, averaged over 30 replications of the simulation-normalization-testing procedure, result for most methods in high statistical power only at the cost of high FDRs (Figure 2.4, Supplemental Figure A.6). For datasets normalized by Dino, however, high statistical power can be achieved at far lower FDRs, further supporting the results from the previous negative and positive control case studies.

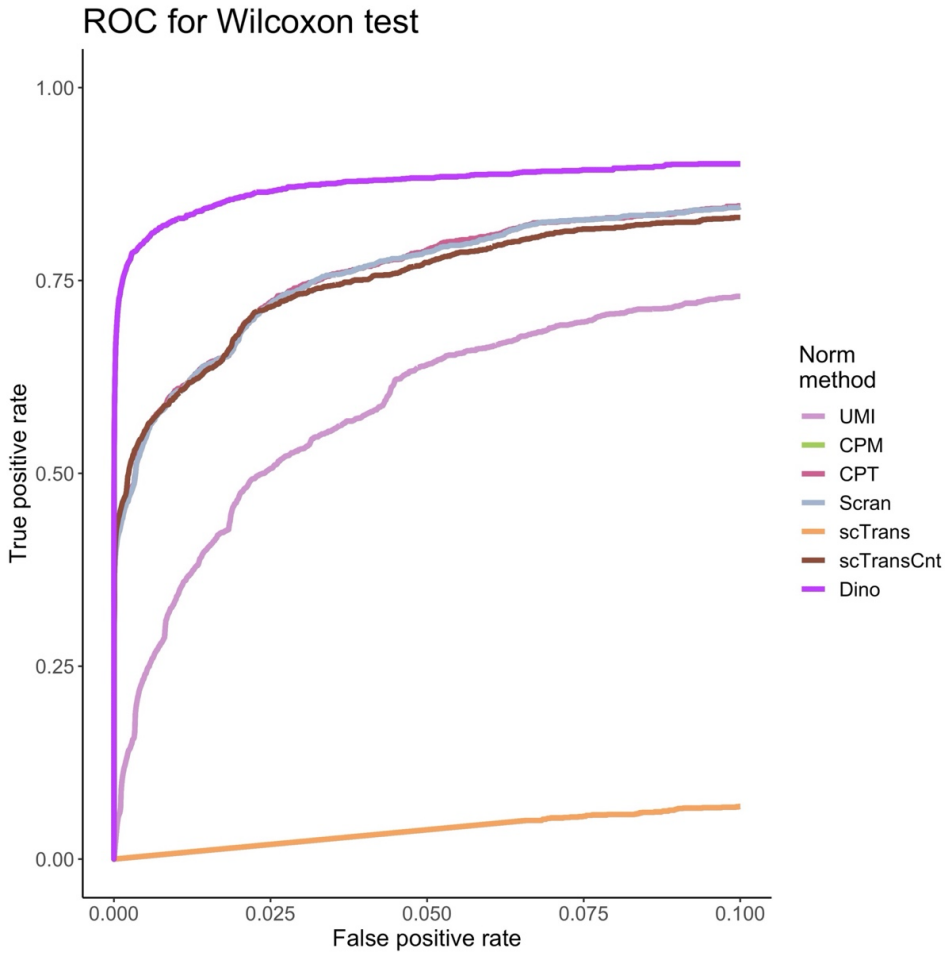


Figure 2.4: **The effects of normalization on downstream DE analysis.**

Simulated data based on the PBMC68K_Pure dataset were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR and average FPR for a Wilcoxon rank sum test, where the average is calculated across 30 simulated datasets.

To expand on these results, we further conducted the simulation-normalization-testing procedure under the MAST test designed for single-cell data and a simple t-test; both tests being applied to data simulated from the full panel of experimental data. Under the MAST test, results were similar to those observed for the Wilcoxon test, an expected result as both tests are sensitive to differences in distribution while the t-test resulted in more consistent results across methods due to that test's focus on differences in mean rather than differences in distribution (Supplemental Figures A7-8). In most cases across both datasets and testing methodologies, normalization with

Dino and thresholding significance at an adjusted p-value of 0.01 via the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995) resulted in the meaningfully higher statistical power observed in Figure 2.4. In almost all cases, the use of Dino at a minimum resulted in significantly lower FDRs at similar statistical power (Supplemental Tables A1-3).

Finally, to consider implications for another common analysis technique, we performed dimension reduction and clustering on our three datasets with either cell type annotations (PBMC68K_Pure) or pseudo-annotations (MaltTumor10K and PBMC5K_Prot). Computed clusters were then compared to the ground truth annotations by the Adjusted Rand Index (ARI). In the PBMC68K_Pure dataset, Dino performed similarly to its closest comparators, Scran and scTransform, with a slightly but not significantly higher ARI (Figure 2.5 A). To stress test the normalization methods' ability to remove nuisance variation without compromising downstream analysis, we performed a duplicate analysis where, prior to normalization, a random sample of half of the cells in the dataset were downsampled to 25% of their original LS. In all cases, this greater technical variation can be seen as an artifact in the dimension reduction where the downsampled cells (bold) of a particular cell type annotation (color) form a cluster offset from the unaltered cells. However, this effect is visually less dramatic for Dino normalized data, resulting in less confounding of cell types in the dimension reduction, and significantly higher ARI (p-value $\leq$ 1e-16 under a t-test) when the downsampling-normalization-clustering pipeline is repeated (Figure 2.5 B-C). These effects are observed again when conducting the testing pipeline on other datasets (Supplemental Figures A9-10).

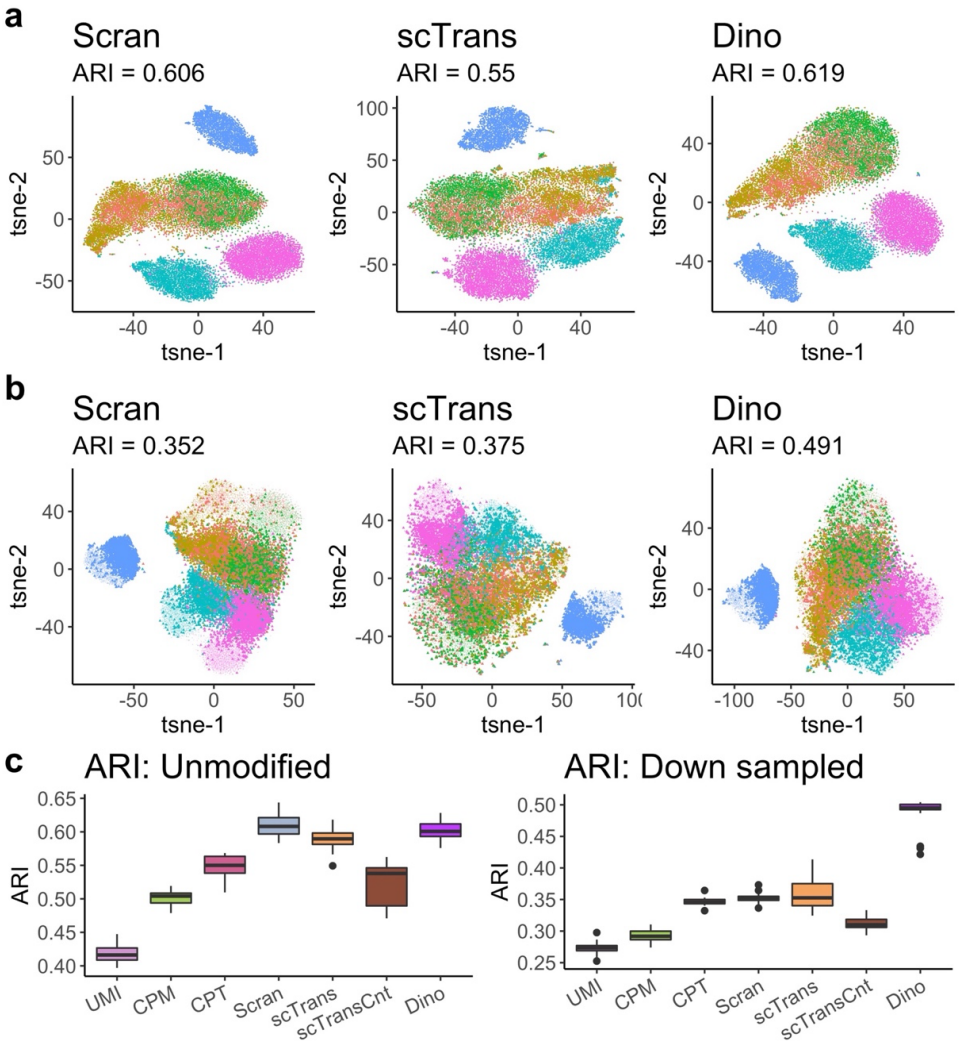


Figure 2.5: **The effects of normalization on clustering.**

a) tSNE plots of normalized PBMC68K_Pure data, colored by 6 cell-type annotations, show similarly high clustering accuracy across methods. b) The same clustering plots as in (a), but with half the data down-sampled (down-sampled cells in bold) prior to normalization to produce greater differences in *LS*. c) Boxplots of ARIs for multiple un-modified and down-sampled datasets across 24 samples of 25 thousand cells from the PBMC68K_Pure dataset.

2.4 Discussion

The development of novel sequencing platforms has dramatically increased the resolution with which transcriptional dynamics can be assayed. In particular, the parallel construction of cDNA sequencing libraries from microfluidics systems has allowed the simultaneous sequencing of many

thousands of individual cells from a common tissue sample. Simultaneously, the integration into such pipelines of UMIs has improved both the accuracy and precision with which such measurements can be taken, primarily by reducing or removing expression biases such as those introduced during PCR amplification. However, these technological advancements bring with them a fundamental change in the statistical properties of the data they generate, most notably by dramatically reducing the per-cell number of reads which can be sequenced; increasing the number of zeros present in the resulting dataset and rendering techniques which rely on (approximately) continuous distributions of counts less effective or even ineffective.

Our proposed method, Dino, addresses these challenges to effective normalization of technical variation in LS by adopting a model of count data based on the NB distribution. This modeling framework is common among existing techniques (Hafemeister and Satija, 2019; Love *et al.*, 2014; Risso *et al.*, 2018; Huang *et al.*, 2018). However, the incorporation of mixture components in Dino grants the method unique flexibility to accommodate the heterogeneity of expression present in many modern datasets. Dino further improves on existing methods by normalizing not through a deterministic transformation of observed UMI counts, but rather through a resampling procedure based on a flexible estimate of the true distribution of gene expression conditioned on observed UMI counts and LS for each gene and cell. This stochastic approach uniquely allows Dino to reduce or remove the dependency relationship between the distribution of normalized expression data and LS which remains after normalization with existing methods. In so doing, normalization by Dino results in more accurate and more powerful downstream analyses.

2.5 Future work

2.5.1 Restricted quantile sampling

The stochastic resampling procedure incorporated in the Dino method is successful at removing the dependency relationship between LS and the distribution of normalized expression. Further, the high power observed in simulated testing is both consistent across repeated simulations (Figure 2.4 and Supplemental Figures A.6-8) and our testing showed high concurrence (>90%) in true positive and false negative calls when testing on the same dataset normalized by Dino with different random seeds. Nonetheless, the stochastic nature of this procedure might be a concern to some researchers. For this reason, we are currently developing and testing an alternate resampling framework specifically aimed at preserving the relative *rank* or *order* of sequenced UMIs after normalization. While currently considered experimental, our procedure, Restricted Quantile Sampling (RQS) is an alternate sampling procedure available to the user as a function option in our R package.

RQS is based on the idea that, if two cells have the same LS but, for a given gene, the first cell has strictly more sequenced UMIs than the second cell, then, after normalization, the first cell should still have strictly greater normalized expression than the second cell. In other words, the ranks of UMI counts for cells of the same LS should not change, and the ranks of UMI counts for cells with similar LS should be unlikely to change, as is the case for deterministic methods such as the count scaling implemented in Scran. This approach is implemented by sampling not from the posterior distribution of the gene- and cell-specific latent means, λ_{ij}, but rather by sampling from a restricted range of the marginal distribution (mixture of Gammas) of λ with the range restriction being a function of the observed cell-specific LS and UMI counts.

Specifically, let $F(q|\delta_j)$ denote the CMF of the mixture of Poissons distribution estimated by the accelerated EM algorithm, conditioned on the cell specific (transformed) LS, δ_j. Then the sampling bounds for the RQS method can be defined as

$$L_j \coloneqq \begin{cases} 0 & if\ y_j = 0 \\ F(y_j - 1|\delta_j) & else \end{cases}$$
$$H_j \coloneqq F(y_j|\delta_j)$$

If we further define the inverse CDF of the mixture of Gammas distribution on λ as $G^{-1}(p)$ and let $U(L,H)$ denote the uniform distribution between lower-bound $L \geq 0$ and upper bound $H < 1$, then the RQS normalized expression can be sampled from:

$$\hat{y}_j \coloneqq G^{-1}(p_j)$$
$$p_j \sim U(L_j, H_j)$$

In current testing, RQS performs competitively with default Dino, and may in fact have slightly higher power in some situations. For this reason, we may make RQS the default method in future updates to the package. However, further validation is required before this step can be taken.

2.5.2 Alternate cell-specific size-factors

By default, Dino uses cell-specific LS as the measure of nuisance variation. In many cases this may be appropriate, or at least sufficient for successful analysis. However, in some situations it may be that LS is meaningfully correlated with (or otherwise dependent upon) true biological variability of scientific interest. This could occur when the majority of DE genes between two cell types are upregulated in one cell type compared to the other; i.e., a generally more transcriptionally active cell type across the full transcriptome. In such cases, normalization to remove the effects of nuisance variation measured by LS will, in the best case, additionally remove biological variation

of interest or, in the worst case, even induce spurious differences in average expression in the normalized data leading to high rates of false positives.

For this reason, Dino allows the user to supply alternate cell-specific measures of nuisance variation, such as the size-factors calculated by Scran. This observation and quick solution in Dino are also the foundation upon which the ongoing work on our new normalization method, Rhino, are based. Further details on this concern and on the current work on Rhino are the topics of Chapter 3.

2.6 Publication information

The methods, results, and figures described in this chapter are published in Brown, et al., 2021 (Brown, Ni, *et al.*, 2021).

The methods are further publicly available as an R package on Github (https://github.com/JBrownBiostat/Dino) and Zenodo (https://doi.org/10.5281/zenodo.4897558).

3 Improving nuisance variation estimation with Rhino

3.1 Background

A fundamental first step in the effective normalization of genomic data is the calculation of accurate measures of nuisance variability inherent in a given dataset. The normalization method then attempts to correct for gene expression differences characterized by these measures of nuisance variation. If any biases or systematic errors are introduced into the initial estimation of nuisance variation, such errors will be propagated to all downstream analysis.

Most frequently in modern RNA-sequencing analysis, nuisance variation is measured as cell-specific (or sample-specific) LS. LS is simple to define and quick to calculate – simply the column sums of the gene by cell expression matrix – and many studies demonstrate that LS is often at least sufficient for useful analysis. However, the use of LS has an in-built limitation in that the measure implicitly assumes that the sources of nuisance variation are typically technical in their origin, their influence being independent of the biological condition of any given cell. Specifically, one might suppose that technical sources of variation would alter the total per-cell sequenced expression, otherwise called LS. However, by using LS as the measure of nuisance variation, the scientist is implicitly also assuming that *all* such variations in LS are technical (specifically nuisance), an assumption which precludes the possibility that certain cell-types or conditions might express more or less in total across the genome for biological reasons of interest.

A more robust measure, by contrast, might be enabled by a more flexible assumption on the underlying biology being studied. For example, Scran, as well as MR upon which it is based (Lun *et al.*, 2016; Anders and Huber, 2010), both assume that some set of genes are equivalently

expressed (EE, having the same expected value after correcting for nuisance variation) between any individual cell and a reference measure of gene expression generated by taking the geometric mean of expression across cells, or between pooled cell expression and the reference in the case of Scran. Further, these methods suppose that a representative EE gene in this group can be identified by identifying the median of ratios of expression between the test cell and the constructed reference. Such an approach has the clear potential advantage of allowing the estimation of nuisance variation to be robust to biologically relevant differences in the underlying distribution of total expression across the genome. Of course, the downside of such an approach is one of precision; estimating size factors from the ratio of expression of just one gene is much less precise than estimating size factors by LS which, by definition, pools expression information across the entire genome.

Relative advantages or disadvantages aside, the potential relevance of this difference in approaches – basing the method for estimating nuisance variation on biological assumptions – can be demonstrated through an example of a (very) simple sequencing dataset. Suppose that we sequence two genes from four cells. Further suppose that the first two cells (columns 1 & 2 in the expression matrix) are of one cell-type and the second two cells (columns 3 & 4 in the expression matrix) are of another cell-type. Finally suppose that we have prior knowledge that the first gene (row 1 of the expression matrix) is DE between the two cell-types and that the second gene (row 2 of the expression matrix) is EE between the two cell-types. We observe the following raw expression:

$$\begin{bmatrix} 2 & 4 & 7 & 14 \\ 1 & 2 & 2 & 4 \end{bmatrix}$$

If we were to use LS as the measure of nuisance variation, then our scale factors would be some constant multiple of 3, 6, 9, and 18, those being the LS values for cells 1 through 4

respectively. Choosing a convenient multiplier such that, after normalization, the total expression for each cell is 9, our cell-specific scale factors are: 3/9, 6/9, 9/9, and 18/9. Dividing by these scale factors gives the LS normalized expression:

$$\begin{bmatrix} 6 & 6 & 7 & 7 \\ 3 & 3 & 2 & 2 \end{bmatrix}$$

In this case, the problem is clear. Gene 2, which we know to be an EE gene, is normalized to express 50% more in the first cell type compared to the second; a larger fold change than even exists in gene 1 between cell types, despite the implicit assumption that gene 1 could be DE. If instead we derive scale factors as a multiple of the expression in gene 2, known to be EE, we would calculate factors such as: 1/2, 2/2, 2/2, and 4/2 respectively (multiplying the expression in gene 2 by ½), giving normalized expression of:

$$\begin{bmatrix} 4 & 4 & 7 & 7 \\ 2 & 2 & 2 & 2 \end{bmatrix}$$

Here, not only is the EE status of gene 2 maintained across cell-types as desired, but also the fold change between cell-types for gene 1 has increased. This improvement further highlights the potential problem inherent in normalization to remove the effects of LS: doing so can not only remove true biological differences as in gene 1 above, but further, inappropriate correction for LS can in fact induce spurious DE between genes which should be expressed at similar rates between groups as in gene 2 above. This observation, that correcting for a biased measurement of nuisance variation can induce false positives is not new. As mentioned in the Introduction, at least as far back as 2010, Robinson and Oshlack observed that normalization for LS between cell-types, kidney and liver in their case, could induce false positives (Robinson and Oshlack, 2010). In their case, the effect was as dramatic as an approximately 30% apparent downregulation of the average gene in liver cells compared to kidney, including an average ~30% downregulation of expression in a select panel of housekeeping genes in liver compared to kidney. These genes, by definition of

being identified as housekeeping genes, were particularly expected to demonstrate constant expression following appropriate normalization. In their case, the culprit was that among truly DE genes between liver and kidney, upregulation was systematically higher in liver resulting in inflated LS measures of nuisance variation for that tissue and correspondingly overcorrected (divided by too large a scale factor) expression following normalization.

We observe a similar effect in modern sequencing data. In particular, we analyzed spatial sequencing data (10X Visium platform) from a sequential pair of sections from a skin biopsy with the lab of Dr. Drolet, Department of Dermatology at UW Madison. In these sections, the cell-type heterogeneity is clear from the H&E imaging, and this heterogeneity directly corresponds to region-specific variation in the total number of sequenced UMIs for each spot (the unit of measurement for this technology with spot-specific location metadata); each spot containing on the order of 10s of cells (Figure 3.1). High density regions such as the epidermal layer (far right in sample A, bottom in sample B) or highly active regions such as glandular tissue (center in both samples) show higher raw expression than low density/low activity regions such as the deeper adipose tissue (far left in sample A, top in sample B). Following successful normalization, we would expect that these effects, when driven by technical attributes such as the density of cells, would be removed. However, when driven by biological effects such as the cell-type-specific systematic upregulation of a set of genes, we would expect successful normalization to preserve these regional differences, similarly to the liver and kidney cells of Robinson and Oshlack.

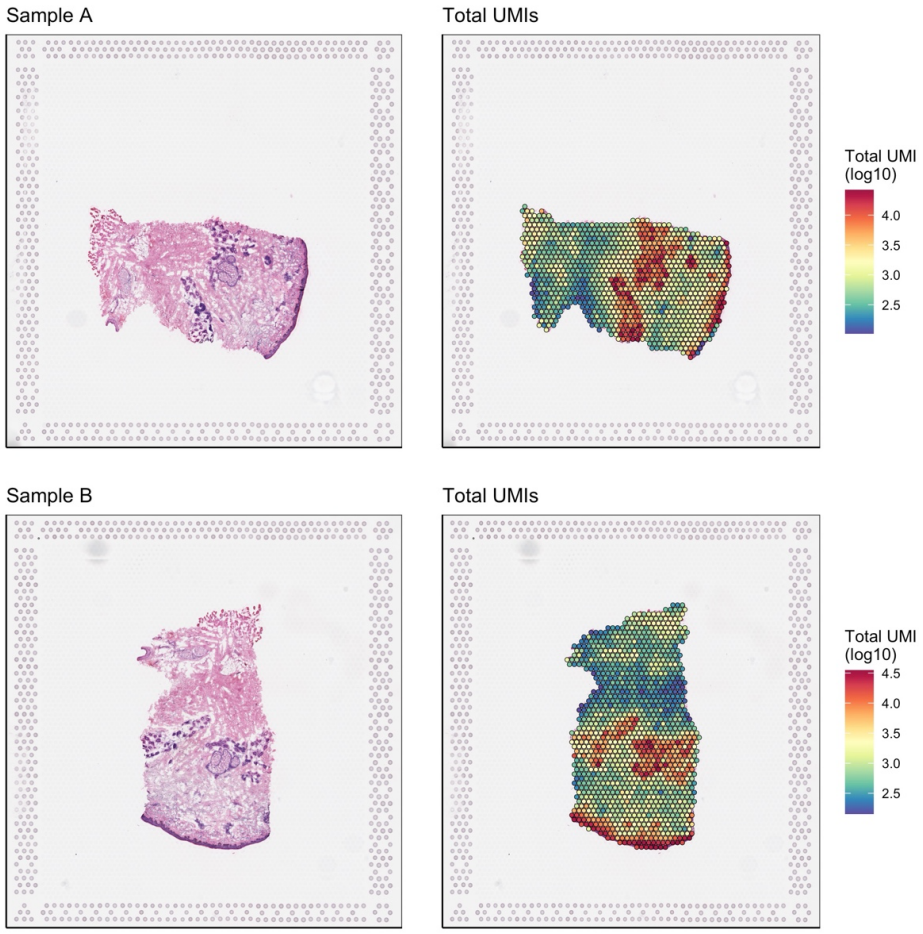


Figure 3.1: **Spatial patterns in total sequencing levels**

(top) H&E stain (left) and sum of total sequenced UMIs (right, log10 scale) from Sample A show correlation between cell-type/chromatin density in the H&E image and total sequenced UMIs. (bottom) Similar results hold for the Sample B section and sequencing taken from the same skin biopsy as Sample A.

However, following normalization for *LS* under the Dino method, we see that total normalized expression is constant up to sampling variation across the whole tissue section, as is expected for methods that correct for LS (Figure 3.2). Looking deeper, if we plot the spatial sum of normalized expression from a subset of the top 1,000 highest variance genes (variance of $\log(x+1)$ on normalized expression), we observe there are spatial patterns in the sum of normalized expression as expected. However, performing the same analysis on the following 4,000, low variance genes (all remaining genes after these first 5,000 are considered "non-expressed"), spatial

patterning persists despite the fact that this subset is expected to contain primarily, if not only, housekeeping and other EE genes across cell types.

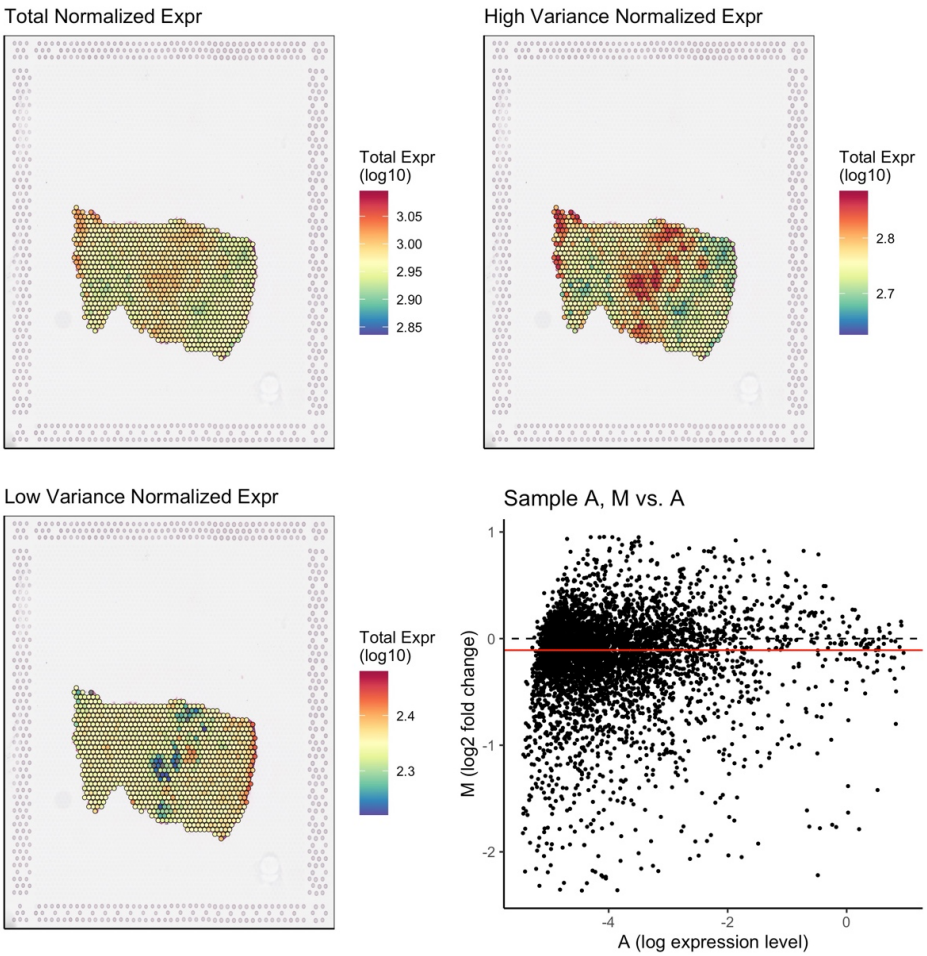


Figure 3.2: **LS normalization induces spurious differential expression**

(top-left) Normalization of expression by Dino for sample A results in roughly constant total expression across the tissue section. Sub-setting the top 1000 highest variance genes following normalization recapitulates the spatial dependency on expression levels as expected (top-right). However, the spatial dependency remains even when considering the next 4000 genes (low variance) (bottom-left) indicating a systematic bias in normalized expression for the typical gene. This systematic bias is measured between sweat glands and the dermis papillary with the more transcriptionally active sweat glands being systematically downregulated (median in red) after normalization as shown by an M vs. A plot of the full 5,000 genes previously considered (bottom-right).

Further, we can subset two sub-populations of cells by their morphology on the H&E image and known marker genes: transcriptionally active cells from the sweat glands and less

transcriptionally active cells from the structural dermis papillary. If we calculate the fold-changes in average (normalized) gene expression between these two tissues for the same first 5,000 genes and visualize the log-ratios as an M vs. A plot, we observe systematic bias away from 0 in the estimated fold-change values (Figure 3.2). In particular, the median log2 fold change across genes is about -0.1, corresponding to a roughly 7% systematic down-regulation of the median gene in the sweat gland spots compared to the dermis papillary spots, contrary to the assumption that only a minority of genes should be DE between the two populations.

It is with these motivating examples in mind that we are developing a new method: *Robust Heterogeneity Integration Normalization* or Rhino. Adapting the resampling techniques from Dino for the specific normalization component of the method, the main contribution of Rhino is the accurate and precise estimation of cell-specific size factors which are unaffected by the potential biases of LS factors when calculated on heterogenous samples. As single cell and spatial sequencing techniques become both increasingly powerful and economical, we expect the analysis of such heterogenous data only to increase, and thus the potential benefit of methods such as Rhino to be significant.

3.2 Methods

3.2.1 Model specification

Like Dino, Rhino approaches normalization as a modeling challenge; calculating parameters for an estimation of the underlying expression distribution and then using that estimate in an empirical Bayes framework to resample normalized expression values. Rhino differs, however, in basing the model of normalized expression on a generalized linear model (GLM). This sacrifices the per-gene flexibility of the approximately non-parametric estimate of gene expression which is present in

Dino. The benefit – similar to and inspired by previous work on GLMPCA (Townes *et al.*, 2019; Townes, 2019) and ZINB-WaVE (Risso *et al.*, 2018) – is that information can be pooled across genes for an improved global estimate as well as the potential to include further experimental or nuisance covariates (e.g., identifiers to correct for batch effects). The pooling of information across genes in Rhino is, further, critical to its primary purpose of estimating robust cell-specific size factors as will be demonstrated below. The model of gene/cell expression can be written for *G*-many genes and *J*-many cells as:

$$Y_{G\times J} \sim f^{P}\left(\mu_{G\times J}\right)$$
$$\mu_{G\times J} = exp\left(B_{G\times N}X_{N\times J} + W_{G\times L}Z_{L\times J} + 1_{G\times 1}\delta_{1\times J}\right)$$

Here we have abused notation slightly to indicate that the elements, y_{gj}, of the expression matrix Y are each independent samples from a Poisson distribution with corresponding mean μ_{gj}. The matrix of Poisson means further follows the usual log link function such that $exp(\cdot)$ is applied elementwise to the contained linear function. The purpose and format of the included matrices in the log-linear function can be described as follows.

1. $\boldsymbol{B_{G\times N}X_{N\times J}}$**: Intercept and experimental conditions.** This term encodes information on the log-average expression level for each gene insofar as the remaining terms are in some way centered. Further, this term encodes any additional information about experimental conditions, batch effects, or other covariates to be modeled.

    a. $\boldsymbol{B_{G\times N}}$ is the coefficient matrix for the *g*-many intercepts and experimental linear effects across the *N*-many effects. At a minimum, $N = 1$ and $B_{G\times 1}$ contains only gene intercept (expression level) information. At greater

values of N, $B_{G\times N}$ can encode, e.g., shifts in gene intercept to correct for batch effects.

b. $\boldsymbol{X_{N\times J}}$ is the design matrix for the dataset. In all cases, and in particular in the minimal case of $N = 1$, the first row of X is a row of 1s allowing for a common intercept for each gene g across all cells j. Further rows of X encode user supplied variables to explicitly model and/or regress out during the normalization procedure. For example, these further rows of X may contain indicator variables to encode shifts in average expression between batches. In the formulation of $\mu_{G\times J}$, X is the only parameter which is known; all others need to be estimated.

2. $\boldsymbol{W_{G\times L}Z_{L\times J}}$**: Latent space representation.** This term encodes a latent space representation of the observed expression in L-many dimensions. As such, this term contains the representation of all biological variability across cells, excepting only any experimental conditions the user may choose to incorporate into X, and is directly inspired by an analogous term in GLMPCA (Townes *et al.*, 2019; Townes, 2019).

   a. $\boldsymbol{W_{G\times L}}$ is the matrix of latent dimensions, directly analogous to the components in Principal Component Analysis (PCA). As such, the columns are constrained to be unit length and (approximately) orthogonal.

   b. $\boldsymbol{Z_{L\times J}}$ is the matrix of representations in the latent dimensions for each cell j.

3. $\boldsymbol{1_{G\times 1}\delta_{1\times J}}$**: Per-cell size factors.** This term, in particular the parameter row-vector, δ, encodes the per-cell size factors, the estimated measures of nuisance variation

between cells. Being log-additive, this term performs the same function as scaling the mean expression, μ_j, by a constant across all genes for a given cell as is the case for most normalization methods. Somewhat uniquely, however, δ encodes "relative" rather than absolute measures of the "size" of a particular cell. In particular, δ is constrained to be mean 0 such that $\Sigma_j \delta_j = 0$. This implies that if $\delta_1 - \delta_2 = 1$, then all that is known is that cell 1 contains *exp(1)* times as many UMIs as cell 2 for reasons of nuisance variation *after* accounting for any differences due to biological variability or linear effects encoded in the *WZ* and *BX* matrices respectively.

The observant reader may have already realized that the model as described is not estimable, even given the restricted solution space for W and δ. In particular, the rows of Z can translate arbitrarily while all and any differences can be accounted for by corresponding shifts in the intercept encoded by the first column of B. For this reason, as well as for reasons of computational stability and, not least, the robust estimation of size factors δ, parameters are estimated via penalized maximum likelihood estimation. For the likelihood function induced by the above model of sequenced UMIs, the following negative, penalized, log-likelihood function is minimized

$$-l\left(Y_{G\times J}\right) + \lambda_B \sum_{n>1} |b_n| + \sum_l \lambda_l |w_l| + \lambda_O \sum_{i>j} \left(acos\left(w_i^T w_j\right) - \frac{\pi}{2}\right)^2 + \lambda_z \sum_j |z_j| + \lambda_E U(Z)$$

$$s.t.\ w_l^T w_l = 1 \forall l\ \ and\ \ \sum_j \delta_j = 0$$

$$U(Z) = \sum_i \sum_j \frac{1}{2} c_i P_{ij} \left|z_i - z_j\right|_2^2$$

$$P_{ij} = \begin{cases} \frac{\left(1 + \left|z_i - z_j\right|_2^2\right)^{-1}}{\sum_{k\in\mathcal{N}_p(i)} (1 + |z_i - z_k|_2^2)^{-1}} & j \in \mathcal{N}_p(i) \\ 0 & else \end{cases}$$

$$c_i = \left(\frac{1}{2} \sum_j P_{ij} \left|z_i - z_j\right|_2^2\right)^{-1}$$

Several of the penalty terms serve critical or nuanced purposes, and so we describe each penalty term in detail in a dedicated section below. In brief, however:

1. **Batch effects:** the penalty term with the λ_B penalty coefficient is an *l-1* penalty on the non-intercept coefficients of B, that is, an *l-1* penalty on the batch/experimental effects.

2. **Sparse latent space:** the penalty terms with the λ_l penalty coefficients are *l-1* penalties on the columns of W, imposing sparsity on that matrix.

3. **Latent space orthogonality:** the penalty term with the λ_O penalty coefficient enforces approximate orthogonality on the columns of W.

4. **Sparse representations:** the penalty term with the λ_z penalty coefficient imposes sparsity and median-0 centering on the representations z_j of cells in the latent space.

5. **Latent space "gravity":** the penalty term with the λ_E penalty coefficient imposes an attractive force between latent representations of neighboring cells. Given the approximately inverse-square weighting imposed by P_{ij}, this penalty is roughly analogous to adding a gravitational force which attracts neighboring cells towards each other with little influence on more distant cells. In the above notation, $\mathcal{N}_p(i)$ denotes the set of p-many nearest neighbors to the point z_i.

3.2.2 Model penalties

The penalty terms imposed on the Rhino model optimization serve not only to regularize and render estimable the parameters, they also serve functional purposes. Some terms like the penalization on experimental effects aid in the interpretability of the resulting model. Others like

the attraction between neighboring cells are manifestations of underlying biological assumptions. In this section we detail the purpose and construction of the penalty terms in greater detail.

3.2.2.1 Batch effects

The *l-1* penalty on the non-intercept columns of the experimental effects matrix, *B*, serves two purposes. At the first level, assuming there is more than one column of *B* (otherwise this penalty term has no effect on the objective function being optimized), this penalty serves as a regularizer in the spirit of Lasso regression with the corresponding potential to reduce the error of the subsequent coefficient estimates. More importantly, however, the penalty on the columns of *B* serves to enforce the expectation that estimated coefficients in fact represent the experimental/batch effects being modeled. While it is still imperative upon the user to protect against confounding of biological and experimental/technical variability, the penalization on *B* preferences cell-to-cell variability being modeled in the *WZ* term as intended with the estimated coefficients in *B* representing global shifts between experimental conditions that occur across cell-types. For this reason, it is important that $\lambda_B > \lambda_z$ and that the user have some reasonable confidence that the cell-types across experimental conditions are both heterogenous and somewhat constant. The default value for λ_B is currently set to 1, but can be altered by the user.

3.2.2.2 Sparse latent space

The *l-1* penalty on the columns of *W* serves a similar, but actually more critical function to the penalization of the non-intercept columns of *B*. As before, the penalty imposes Lasso-type regularization. However, the key purpose of this penalty is the sparsity itself. As already stated, the fundamental goal of Rhino is the robust estimation of per-cell size factors. The biological assumption which underpins this approach is that, between any pair of cells, there is some set of

equivalently expressed genes. As with previous methods that made similar assumptions, like Scran, MR, and TMM, the goal is then to estimate size factors from this subset of EE genes. In the context of Rhino, an EE gene is one for which the latent space representation of expression is equal between two cells, that is, $w_g (z_1 - z_j) = 0$. This is trivially true for any row of W which is zero, i.e., for any $w_g = 0$. Such genes are thus considered globally EE, and serve as reference points by which the relative size factors between cell-types can be set.

The details of how this sparsity on W allows for robust estimation of the size factors δ are the subject of section 3.2.3. In the meantime, it is clear that careful consideration must be given to the selection of penalty coefficients $\lambda_1, \lambda_2, \ldots, \lambda_l$ for each column of W. For this reason, the selection of these coefficients is algorithmic, rather than set by the user. To perform coefficient selection, we first note that, for any element $w_{gl} = 0$, the minimization of the objective function will keep $w_{gl} = 0$ if and only if the magnitude of the un-penalized partial derivative with respect to w_{gl} is less than or equal to the corresponding l-1 penalty coefficient, λ_l.

We therefore calculate the un-penalized partial derivative with respect to each element w_{gl} at $w_{gl} = 0$. In practice, the ordering of absolute value of the un-penalized partial derivatives for each column of W results not just in a monotonically-decreasing trend, as must be the case by construction, but an L-shaped, visually convex curve ("visually" because local non-convexities may exist). This is consistent with the hypothesis that any given component (column of W) is representing the biological variation of a finite – and typically small – number of genes across cells, these genes being related or at least correlated in their activity and thus forming the vertical component of the L-curve. The remaining majority of genes, which form the horizontal component of the L-curve are thus considered to be noise; only correlated with the expression in the direction of the component by chance. This then suggests a rule by which to select the penalty coefficients

λ_l: set λ_l equal to the absolute un-penalized partial at the elbow-point – that is, the point of maximum curvature – in the ordered list of partials.

In practice, we allow the user to bias the algorithm in favor of modeling more genes by setting λ_l equal to some multiple of the elbow-point, by default this multiple is 0.95. Further, sanity controls are placed on the estimation procedure such that λ_l should be set between some bounds. By default, these bounds are set at the 99th and 50th quantiles of absolute, un-penalized partials. In practice, this imposes the restriction that no less than 1% and no greater than 50% of genes will be non-zero for any given column of *W*.

3.2.2.3 Latent space orthogonality

In order to make the representation of each cell in the latent space as useful as possible for downstream analysis – e.g., for graph-based clustering on the Rhino output rather than on secondary PCA of the normalized expression – we impose a soft orthogonality condition on the columns of *W* through the penalty:

$$\lambda_O \sum_{i>j} \left(acos\left(w_i^T w_j\right) - \frac{\pi}{2} \right)^2$$

Since the columns of *W* are constrained to the unit sphere, this reduces to a penalty on the angle between each pair of components, preferencing orthogonal components, i.e., $acos(w_i^T w_j) - \pi/2 = 0$. We note that this penalty, as opposed to other penalties which might impose a similar condition, is chosen as the penalty gradient has a particularly convenient form. In particular, some algebra will show that the gradient of a single term (fixed *j*) of the penalty with respect to w_i is

$$-2\lambda_O \left(acos\left(w_i^T w_j\right) - \frac{\pi}{2} \right) \left(1 - \left(w_i^T w_j\right)^2 \right)^{-\frac{1}{2}} w_j$$

In words, the gradient is a scaled vector in the direction of w_j. However, as we are performing constrained optimization on the unit sphere, it is the component of the gradient which is perpendicular to w_i which is important. Denoting the scalar values above by c, this perpendicular component can be calculated as

$$c\left(w_j - w_i w_i^T w_j\right)$$

which in turn has length

$$\begin{aligned}\left|c\left(w_j - w_i w_i^T w_j\right)\right|_2 &= |c|\left[\left(w_j^T - w_j^T w_i w_i^T\right)\left(w_j - w_i w_i^T w_j\right)\right]^{\frac{1}{2}} \\ &= |c|\left[1 - \left(w_i^T w_j\right)^2\right]^{\frac{1}{2}}\end{aligned}$$

which further implies that the gradient can be represented as

$$-2\lambda_O\left(acos\left(w_i^T w_j\right) - \frac{\pi}{2}\right) w_{ji}^{\perp}$$

where the representation follows from expanding c and where $w_{ji}^{\perp}$ denotes the unit vector in the direction of the component of w_j which is perpendicular to w_i. That is, the gradient in w_i is a unit vector pointing directly towards/away from w_j along the meridian of the unit sphere defined by w_j and w_i, scaled by the angle between the two components w_i and w_j.

In practice, the penalty coefficient, λ_O, is progressively increased during optimization such that $|W^TW\text{-}I|_F$ should be small, on the order of *1e-4*, where $|\cdot|_F$ is the Frobenius matrix norm.

3.2.2.4 Sparse representations

We impose an *l-1* penalty on the latent space representations, Z, of the cells primarily for the purpose of estimability. As has already been mentioned, the unpenalized likelihood function is inestimable as any translation of the rows of Z can be matched by corresponding updates to the intercept column of B. To solve this, we impose a functional restriction that the median in each row of Z should be zero through the *l-1* penalty.

Further, model testing has demonstrated that, in some cases, individual elements z_{lj} may diverge independent of the above mentioned estimability concerns. This occurs in sparse cells for which the non-zeros in the column y_j occur only in genes for which w_g is partially or completely non-zero. In such a case, optimization will attempt to minimize the size factor, δ_j so that the zero elements of y_j match the estimated non-zero mean for that gene. This minimization in δ_j is matched by corresponding inflation of z_j. A sufficient penalty on z_j likewise controls this divergent behavior.

In practice, we choose 0.01 as the default value for λ_z, being an empirical intermediate between a smaller value which allows greater freedom and a sufficiently large value to prevent divergent behavior.

3.2.2.5 Latent space "gravity"

The final penalty term is guided by first, the desire to perform soft-clustering within the Rhino framework, and second, the assumption that neighboring cells within the latent space are more likely than not derived from the same cell-type. This second component in particular suggests that differences between neighboring cells are more likely due to random noise than to true biological differences, and so differences in the estimated means for such pairs of cells should be penalized; a principle which provides the functional basis for the desired soft-clustering.

The specific formulation of this penalty term is driven in general by the theory of diffusion maps and in particular by the mathematics behind the t-SNE dimension reduction technique (van der Maaten and Hinton, 2008). The specific formulation of this penalty is, as noted above

$$\lambda_E U = \lambda_E \sum_i \sum_j \frac{1}{2} P_{ij} |z_i - z_j|_2^2$$

This formulation has a convenient interpretation: as with the similar term in t-SNE, it is exactly the formula for the spring potential energy (hence the notation $\lambda_E U$) of the system of cell

representations, z_j, if each cell is connected to all others by springs, and the strength (spring coefficient) of those connecting springs is P_{ij}. Minimizing the objective function thus includes a minimization of the potential energy of the latent representation; drawing nearby cells closer to each other.

The value P_{ij} has a dual interpretation, however. In the context of energy minimization as above, P_{ij} is the spring coefficient of the spring connecting cell i to cell j. In the context of diffusion mapping, however, P_{ij} is the conditional probability that cell i would choose cell j as its neighbor if this selection was a random process with selection probabilities being a function of Euclidean distance. For this reason, we follow the developments incorporated into t-SNE and use a Cauchy distribution (equivalent to a t-distribution with 1 degree of freedom) as the neighbor selection kernel. This gives the formula for P_{ij} as

$$P_{ij} = \begin{cases} \frac{\left(1 + \left|z_i - z_j\right|_2^2\right)^{-1}}{\sum_{k \in \mathcal{N}_p(i)} (1 + |z_i - z_k|_2^2)^{-1}} & j \in \mathcal{N}_p(i) \\ 0 \ \ else \end{cases}$$

For computational efficiency, we only calculate P_{ij} on a small set of p-many neighbors of cell i, $\mathcal{N}_p(i)$, setting values for other cells j to zero. As with t-SNE, the Cauchy distribution has a convenient density-preserving property when used in this way. This is because the Cauchy PDF is approximately an inverse-square function of the distance between cells. If two cells z_j and z_k are both in the neighborhood of cell z_i, but z_k is twice as far away as z_j, the conditional probability that cell z_k will be selected as the neighbor is approximately one fourth the probability that cell z_j will be selected regardless of the actual magnitude of those distances.

In the application to Rhino, we further extend this density-preserving property by scaling the values P_{ij} by a cell-specific constant, c_i. While the conditional likelihoods of neighbor selection might be approximately scale-invariant under the Cauchy distribution, the corresponding spring energies incorporated into the objective function are not through

the action of the squared distance component. For this reason, we normalize the penalty across all neighbors of cell z_i such that the expected value of the penalty, when initialized, is 1 for all cells. Specifically, we set

$$c_i = \left(\frac{1}{2} \sum_j P_{ij} \left| z_i - z_j \right|_2^2 \right)^{-1}$$

In practice it would be unstable, not to mention inconvenient, to have the values of P vary as the values of Z do. For this reason, the values of P are fixed following initialization for all future parameter updates. To ensure that the values of Z are reasonable at the time of initialization, the objective function is allowed to optimize without the influence of this energy penalty for several iterations. After this initial estimate of Z, P is initialized, and the optimization is continued under the full penalty. By default, the energy penalty coefficient is set to $\lambda_E = 1$.

3.2.3 Robust estimation of cell-specific size factors

Much of the introduction was dedicated to demonstrating why the use of LS as the measure of nuisance variation could lead to biased results, particularly in the context of cell-type heterogeneity. However, these estimates do have one particular benefit, their precision. By pooling information across all genes for each cell, LS factors are quite precise measures, even if they are measures of potentially the wrong thing. This is in contrast to some of the existing methods which attempt to provide unbiased size factor estimates.

Scran as the particular example improves upon previous techniques by pooling information across cells and then uses that pooled information to back-calculate cell-specific factors. This procedure gives Scran a certain robustness to the large numbers of zeros often present in modern data. However, at its core, Scran is still based on a ratio of the expression of one gene to a reference

expression level. Even when pooled, the expression level of the median ratio gene can still be in the single digits, resulting in an accurate, but imprecise size factor estimates.

Rhino aims to combine both approaches. Like Scran, Rhino aims to identify the set of EE genes and use those genes to calculate relative size factors between cells of potentially different types. However, like LS estimation, Rhino still uses information from all genes, particularly leveraging these data when calculating size factors between cells of the same or similar cell-types. The key to this approach is letting the *WZ* term identify first when two cells are likely to be of similar or different cell-types. If of different types, the relative difference in size factors will be driven by the EE genes (zeros, particularly zero rows in *W*). If of similar types, the *WZ* term drops out of the calculation of relative size factors and information across all genes can be leveraged. We sketch the inspiration for this in the context of maximum likelihood estimation below.

Consider two cells of the same cell type, *j* and *j'*. By this fact, they will have similar values of z_j and $z_{j'}$ since there is little if any biological variability between them. In the absence of extraneous cell-specific nuisance or experimental effects (e.g., *X* is absent from the model excepting the first intercept row), the cell specific mean vectors will be:

$$\mu_j = exp\{B_1 + Wz_j + 1_G\delta_j\} = exp\{B_1 + Wz_j\}\Delta_j = \mu_{RE,j}\Delta_j$$
$$\mu_{j'} = \mu_{RE,j'}\Delta_{j'}$$

where, as before, *exp{}* is applied elementwise. It can also be shown that the unpenalized derivative in δ_j is given by:

$$\sum_g \mu_{gj} - y_{gj} = \Delta_j \mu_{RE,\cdot j} - y_{\cdot j}$$

which implies that the maximum likelihood estimate of Δ_j is given by $\Delta_j = y_{\cdot j}/\mu_{RE,\cdot j}$. By extension, the ratio of exponentiated size factors is:

$$\frac{\Delta_j}{\Delta_{j'}} = \frac{y_{\cdot j}\mu_{RE,\cdot j'}}{y_{\cdot j'}\mu_{RE,\cdot j}}$$

However, as residual biological differences recede, that is, as $|z_j - z_{j'}| \to 0$, we observe that $\mu_{RE,\cdot j} - \mu_{RE,\cdot j} \to 0$ which further implies $\Delta_j / \Delta_{j'} \to y_{\cdot j} / y_{\cdot j'}$. In words, for two cells without estimated biological differences, the ratio of exponentiated RE size factors is simply the ratio of the LS of the two cells. In this way, a larger proportion of genomic information is used to improve the precision of size factor estimates $\delta_j = \log(\Delta_j)$.

The alternate case, where there are two new cells, j and j' which are from different cell types (still no experimental effects) is more complicated, so we suppose for this illustration that we are in the simplified situation where the entire dataset is comprised of just these two cells and proceed by contradiction. Suppose it is the case that we have obtained the maximum likelihood estimate for the unpenalized objective function. Further suppose that the set of estimated EE genes, i.e., those for which the rows of W are 0, is denoted by $\mathscr{E}$ and that

$$\frac{\Delta_j}{\Delta_{j'}} \neq \frac{\sum_{g \in \mathcal{E}} y_{gj}}{\sum_{g \in \mathcal{E}} y_{gj'}}$$

Since it is the case that $\mu_{RE,g,j} = \mu_{RE,g,j'} \; \forall g \in \mathscr{E}$, and since the gradient in B can be written as $(\mu_j - y_j) + (\mu_{j'} - y_{j'})$, we are left with two options.

First, if

$$\mu_{RE,g,j}(\Delta_j + \Delta_{j'}) - (y_{gj} + y_{gj'}) \neq 0 \; for \; any \; g \in \mathcal{E}$$

then the gradient in B is not zero and hence our parameters do not derive from a maximum likelihood estimate, a contradiction.

In the alternate case, we have that

$$(\Delta_j + \Delta_{j'}) = \frac{\sum_{g\in\mathcal{E}} y_{gj} + \sum_{g\in\mathcal{E}} y_{gj'}}{\sum_{g\in\mathcal{E}} \mu_{RE,g,j}}$$

and since

$$\frac{\Delta_j}{\Delta_{j'}} \neq \frac{\sum_{g\in\mathcal{E}} y_{gj}}{\sum_{g\in\mathcal{E}} y_{gj'}}$$

we have that both

$$\Delta_j \neq \frac{\sum_{g\in\mathcal{E}} y_{gj}}{\sum_{g\in\mathcal{E}} \mu_{RE,g,j}} \; and \; \Delta_{j'} \neq \frac{\sum_{g\in\mathcal{E}} y_{gj'}}{\sum_{g\in\mathcal{E}} \mu_{RE,g,j}}$$

Since the maximum likelihood estimate implies that $\Delta_j = \Sigma_g y_{gj} / \Sigma_g \mu_{RE,g}$, we have that

$$\sum_{g\notin\mathcal{E}} y_{gj} \neq \Delta_j \sum_{g\notin\mathcal{E}} \mu_{RE,g,j} = \sum_{g\notin\mathcal{E}} \mu_{g,j} \; and \; \sum_{g\notin\mathcal{E}} y_{gj'} \neq \Delta_{j'} \sum_{g\notin\mathcal{E}} \mu_{RE,g,j'} = \sum_{g\notin\mathcal{E}} \mu_{g,j'}$$

Since the unpenalized gradient with respect to z_j can be written as

$$W^T(\mu_j - y_j)$$

unless W is orthogonal to both differences in j and j', this implies that the gradient with respect to z_j or $z_{j'}$ is non-zero and again, our parameters cannot derive from a maximum likelihood estimate, a contradiction.

Finally, since the unpenalized gradient with respect to W can be written as

$$(\mu - Y)Z^T$$

optimization updates to W will move W into the span of at least some of the differences μ_j-y_j excepting only the pathological case wherein the parameters are initialized at a saddle point in the objective function. Given the dimensionality of the data and the fact that W and Z are initialized by PCA on $\log(Y + 1)$, such a pathological case is implausible.

Between these two situations – cells of the same cell-type or cells of differing cell-types – we have a worst-case scenario analogous to the method of TMM absent the heuristic in that method. That is, relative differences in scale-factors are minimally estimated by the sum of observed UMIs

across the set of genes estimated to be globally EE. In the best case, between cells of the same type, the sum of expression across all genes is used as all genes are definitionally EE between such cells. The reality is, of course, somewhere in-between as there will presumably be multiple cells in each of multiple cell-types in the dataset. Robust estimations will be derived across all genes for cells of the same type and information is then pooled across cells of different types through the action of WZ to estimate improved between cell-type size factors.

3.2.4 Model optimization

The Rhino model of gene expression is estimated, as has been previously mentioned, by the method of maximum likelihood. In particular, we proceed by minimizing the penalized, negative log-likelihood function. However, three challenges present themselves. First, many of the penalty terms take the form of *l-1* penalties which naturally do not lend themselves to standard calculus-based optimization techniques. Second, the objective function is easily shown to be non-convex, both in terms of the likelihood parameters as well as some of the penalty functions. Third, two parameters, δ and W are constrained in their solution space. To address all three of these challenges, we perform a modified quasi-newton optimization based on the L-BFGS method (Liu and Nocedal, 1989).

The first and simplest modification is to perform partitioned optimization, partially optimizing the objective function in Z and δ and then partially optimizing in B and W, alternating until convergence or iteration limit. This modification provides two benefits. First, it partially (though by no means completely) resolves the non-convexity issues. Second, and more importantly, updates in Z and δ require computing the gradient on all cells while updates in B and W do not. We therefore proceed, in the case of large datasets, to optimize B and W on rotating, random

subsets of the data, increasing the size of the subset when the observed increase in likelihood (calculated during updates to Z and δ) becomes a small fraction of the predicted increase. In this, we follow the method of Multi-Batch L-BFGS described by Berahas and Takáč (Berahas and Takáč, 2020).

Second, in order to perform optimization in the context of *l-1* penalties, we alter the calculated and stored gradients in accordance with the OWL-QN method (Andrew and Gao, 2007). In particular, Andrew and Gao observed that the Hessian was a function of only the component of the objective function which did not include the *l-1* penalty. As such, when using modified L-BFGS, we can and do derive accurate estimates of the approximate Hessian by taking the differences of the gradients calculated only on the negative log-likelihood and penalty terms that do not contain *l-1* regularization. Andrew and Gao further observed that the full gradient could be accurately calculated assuming that the initial and update point remain in the same quadrant, that is, that all parameter signs remain constant before and after the update, excepting only parameters that become zero or become non-zero. Replicating this procedure, we set to zero any parameters whose sign flips from positive to negative or visa-versa upon update.

The most important aspect of the fitting of the Rhino model is the constrained optimization of W and δ. Here, we once again adapt our implementation of sparse L-BFGS, relying this time on some of the work into optimization on manifolds (Qi *et al.*, 2010; Huang *et al.*, 2015). Applying the techniques of orthogonal projection of the gradient onto the tangent space and vector transport are straight forward in the case of updates to δ where the restriction is merely one of zero mean. The procedure is slightly more complex in the case of optimization of W. Fortunately, the theory of optimization on the unit sphere is well studied, and we are thus able to extend it to the context

of *l-1* optimization implemented through our application of OWL-QN. Generally, the tangent space at a given point, *x*, is:

$$T_x S^{n-1} = \{\xi \in \mathbb{R}^n : x^T \xi = 0\} = \{\xi \in \mathbb{R}^n : x^T \xi + \xi^T x = 0\}$$

The orthogonal projection onto the tangent space is:

$$P_x \xi_x = \xi - x x^T \xi_x$$

The retraction onto the unit sphere S^{n-1} is:

$$R_x(\eta_x) = \frac{x + \eta_x}{|x + \eta_x|_2}$$

The parallel transport of a vector in the tangent space, ξ, along the geodesic from a point x in the direction of another vector in the tangent space, η, is:

$$P_{\gamma_\eta}^{t \leftarrow 0} \xi = \left(I_n + (cos(|\eta|_2 t) - 1) \frac{\eta \eta^T}{|\eta|_2^2} - sin(|\eta|_2 t) \frac{x \eta^T}{|\eta|_2} \right) \xi$$

In our application, we calculate the update direction *d* by Riemannian L-BFGS (Huang *et al.*, 2015). However, the *l-1* penalty and OWL-QN approach mean that our updated point may not lie along the geodesic from our starting point in the direction of *d*. Specifically, our updated point may not be $R_x(d)$ for initial point *x* and step length 1. As such, we must re-calculate the "actual" update direction, *d'*; that is, calculate the updated direction such that $R_{x1}(d') = x_2$ for updated point, x_2, as:

$$d' = \frac{x_2}{x_1^T x_2} - x_1$$

When calculating and updating the stored vectors *s* and *y* for Riemannian L-BFGS, we can then use *d'* when performing parallel transport on the old space. In particular, the parallel transport from x_1 to x_2 in the direction of *d'* becomes

$$P_{d'}(\xi) = \left(I_n + (x_1^T x_2 - 1) \frac{d' d'^T}{|d'|_2^2} - \sqrt{1 - (x_1^T x_2)^2} \frac{x_1 d'^T}{|d'|_2^2} \right) \xi$$

Finally, when correcting the transported previous gradient in the calculation of the current y, the differential retraction can be calculated as:

$$\mathcal{T}_{R_{d'x}}(\xi_x) \coloneqq \frac{d}{dt} R_x(d' + t\xi_x)|_{t=0}$$

In our case, however, $\xi_x = d'$, and so the above reduces to:

$$\mathcal{T}_{R_{d'x}}(\xi_x) = \frac{d}{dt} \frac{x + d'(1+t)}{|x + d'(1+t)|_2} = \frac{|x+d'|_2 d' - (x+d')|x+d'|_2^{-1}(x+d')^T d'}{|x+d'|_2^2}$$

3.2.5 Normalization by resampling

As previously mentioned, the Rhino model identifies robust, cell-specific size factors through the unsupervised identification of candidate DE genes and leverages cell-type similarities to pool information across cells both in the construction of the latent space and through the action of the potential energy penalization. To use the now fitted Rhino model for normalization, we adopt the resampling techniques from Dino.

In particular, given maximum likelihood estimates of μ, we reformulate the model as a NB random variable to account for any potential overdispersion. Gene-specific overdispersion parameters, θ_g, are fit via maximum likelihood on the corresponding likelihood function for fixed mean μ:

$$y_{gj} \sim NB_{\mu_{gj},\theta_g}(y_{gj})$$

The NB distribution is parameterized under the mean-dispersion form such that the variance is represented as $\theta_g \mu_{gj}^2 + \mu_{gj}$. Estimates of overdispersion are then smoothed across gene expression (B_1), as in scTransform, to reduce overfitting.

Factoring μ into nuisance and biological components yields:

$$\mu = exp\{B_{-1}X_{-1} + 1_G \delta\} \circ exp\{B_1 1_J^T + WZ\} = \mu_N \circ \mu_B$$

where exponentiation is once again elementwise and ∘ denotes the (elementwise) Hadamard product. Normalized expression values are then computed as in Dino by sampling from the posterior distribution of the biological component of expected expression conditional on the observed counts *Y* and the estimated nuisance component of expression. This posterior distribution can then be concisely written as:

$$\hat{y}_{gj} \sim f^{G}\left(y_{gj} + \theta_g^{-1}, \frac{1}{\mu_{N,gj} + \frac{1}{\mu_{B,g}\theta_g}}\right)$$

3.3 Preliminary results

To test and validate the performance of Rhino, we first reconsider the dermal skin sections from the introduction. Performing Rhino normalization on these data resolves several of the concerns previously introduced in the context of LS-type normalization. Notably, we first observe that the per-spot sum of normalized expression is no longer roughly constant across the Sample A skin section (Figure 3.3). Of note, the region of glandular tissue, roughly half way between the left (largely adipose) and right (epidermis) edges of the section, retains its high level of expression first observed in the plot of unnormalized UMIs (Figure 3.1). This is consistent with the hypothesis that glandular tissue, by definition, upregulates a set of genes directly related to the glandular function. Since glandular tissue presumably still needs to express the same housekeeping genes at similar levels to surrounding tissue, properly normalized data should equalize to this second group of housekeeping genes, maintaining the upregulation of the gland-specific genes, and resulting in an overall up-regulation of glandular tissues compared to surrounding tissues.

By contrast, the epidermal layer on the far right of the tissue also demonstrated high un-normalized expression while the Rhino normalized sum is roughly equivalent to neighboring but

deeper tissues (Figure 3.3). From examination of the H&E stain (Figure 3.1), we know that there exists at least one source of nuisance variation that contributes to the high un-normalized counts. Specifically, the blue staining of chromatin on the surface of the tissue indicates a significantly higher density of cells than for deeper tissues which one might subsequently expect to result in higher-than-average total UMIs for that tissue region, as is indeed observed. However, the Rhino normalization which equalizes this surface layer with deeper tissues suggests that the entirety of this effect on the raw UMIs is artifactual.

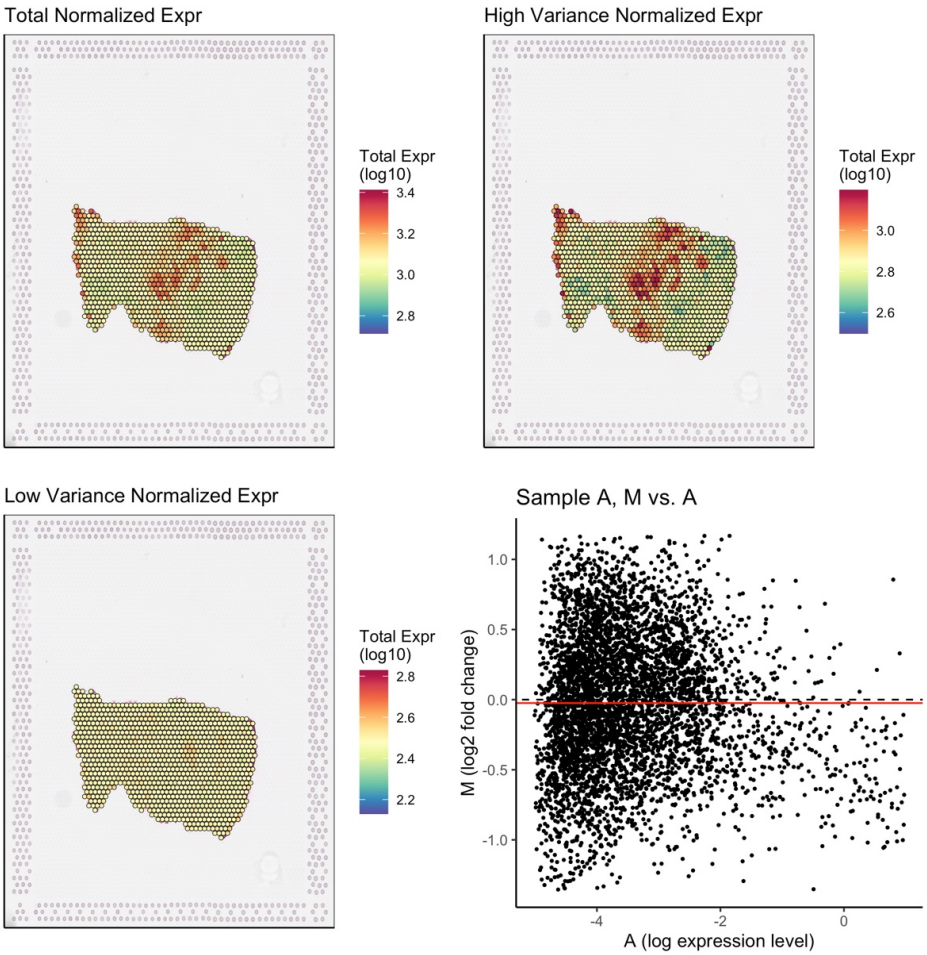


Figure 3.3: **Rhino normalization identifies nuisance variation**

(top-left) Normalization of expression by Rhino for sample A results in roughly spatially variable total expression across the tissue section. Sub-setting the top 1000 highest variance genes following normalization recapitulates the spatial dependency on expression levels as expected (top-right). The spatial dependency, however, is removed when considering the next 4000 genes (low variance), indicating the removal of the previous systematic bias in normalized expression from Dino for the typical gene (bottom-left). This correction is verified between sweat glands and the dermis papillary for the same panel of 5,000 genes with the more transcriptionally active sweat glands demonstrating equivalent expression across most genes (median in red) after normalization as shown by an M vs. A plot (bottom-right).

As before, we attempt to assess the normalization quality by subsetting the top 1,000 highest variance genes and the following 4,000 low variance genes (variance calculated on normalized data, genes beyond these first 5,000 are denoted as non-expressing for these purposes). Under Rhino, we observe clear spatial effects in the high variance subset, directly mirroring the

spatial effects which, to a lesser degree, exist in the sum of all normalized expression plot (Figure 3.3). Unlike the data normalized by Dino, however, the subset of low variance genes exhibits largely flat expression over the full surface of the tissue section, consistent with the assumption that these low-variance genes are expected to contain mostly if not entirely housekeeping and other EE genes whose expression *should* be constant across cells in well normalized data. To follow up on this point, we recapitulate the M vs. A plot from before on the Rhino normalized data between the subset of spots from the Sweat Glands and the subset of spots from the Dermis Papillary (Figure 3.3). Unlike previously, however, we largely fail to see any systematic bias in the estimated fold changes between cell-types with the median log2 fold change reduced to only about -0.02 (red line).

To further validate Rhino, we repeated the analysis of the EMT dataset from the Dino testing on normalized output from Rhino, Dino, and Scran. Similar to previous methods, Rhino is able to cleanly separate the predominantly epithelial cells from the inner region of the cell culture (purple) from the predominantly mesenchymal cells from the outer region of the cell culture (orange) (Figure 3.4). Further, when conducting an identical testing protocol to that used in the validation of Dino, the enrichment results on the curated list of 8 Hallmark terms from Rhino demonstrate similar or greater levels of significance than the corresponding analysis on Dino normalized data, both of which are generally more significant than results derived from Scran normalized data. As with Dino in Chapter 2, of particular note is the greater enrichment significance of the key Hallmark term defining the *epithelial to mesenchymal transition* for the Rhino normalized data (Figure 3.4).

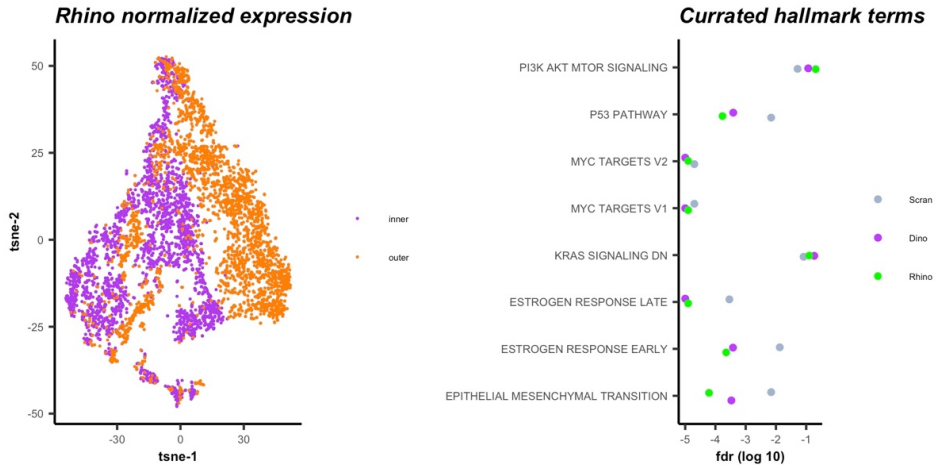


Figure 3.4: **Rhino improves power to detect differentiation pathways**

(left) tSNE plot of Rhino normalized EMT (same as from Dino validation) data cleanly separates the predominantly epithelial cells of the inner culture region (purple) from the predominantly mesenchymal cells of the outer culture region (orange). (right) Enrichment significance values for the curated subset of Hallmark gene sets.

3.4 Future work

The discussed preliminary results are more than encouraging, suggesting that Rhino is accomplishing its primary goal of computing unbiased estimates of nuisance variation. However, further validation is clearly necessary. In particular, it is necessary to run Rhino on a large set of experimental datasets from a variety of sources and containing a variety of levels of cell-type heterogeneity to ensure stability of the algorithm. Further, testing on simulated data is required to ensure that the high-level positive results so far seen extend, as with the enrichment testing on the EMT dataset, to improved sensitivity and specificity in detecting differential expression at the level of individual genes. Related, it is important to ensure that the soft clustering and implicit imputation of the Rhino method do not obscure the unique expression profiles of rare cell types, and so clustering and expression testing will need to be conducted on both simulated and experimental datasets which contain multiple small, unique sub-populations of cells.

However, prior to this in-depth validation work, a few technical challenges remain to be addressed. Of primary concern is the observed although slight instability in the optimization procedure. A likely candidate for this is the high degree of non-convexity imposed on the objective function by the orthogonality penalty, particularly in the later stages of optimization when λ_O has grown large to enforce stricter orthogonality in the columns of *W*:

$$\lambda_O \sum_{i>j} \left(acos(w_i^T w_j) - \frac{\pi}{2} \right)^2$$

To resolve this source of non-convexity and generally improve the efficiency of optimization, we observe that the desired form of *W* is, in fact, an element of the Stiefel manifold *St(G,L)*: the set of all $G \times L$ orthonormal matrices. Fortunately, the theory of Riemannian optimization on the Stiefel manifold is almost as robust as that on the unit sphere, with well-known orthogonal projections, retractions, and vector transport functions. Unfortunately, however, and unlike optimization on the unit sphere, the retraction onto the Stiefel manifold is not zero preserving (except for the first column of *W*), the specific function form for tranporting from element *X* in the Stiefel manifold in the direction of η_X being:

$$R_X(\eta_X) = qf(X + \eta_X)$$

where $qf(\cdot)$ returns the orthonormal matrix *Q* from the QR decomposition of the input. A similar problem exists for the orthogonal projection function:

$$P_X(\xi_X) = (I - XX^T)\xi_X + \frac{1}{2}X(X^T\xi_X - \xi_X^T X)$$

However, as all other optimization operations occur either on the gradient absent the *l-1* penalty or on the corrected update vector defined by the difference between the starting and the OWL-QN corrected updated matrix *W*, only this retraction and projection need to be specifically modified to zero preserving.

Following this method update and validation, we intend to release Rhino as a freely available R package on Bioconductor.

4 Accelerating developmental timing in chimeric stem cell cultures

4.1 Background

The potential benefits of regenerative medicine are great and, as such, have generated significant scientific interest in the field. With a shortage in general of organs available for transplant – never mind organs which are additionally immunologically compatible with the recipient – the ability to grow organs for transplant, as just one application, would represent immediate and dramatic improvements in patient care. In this pursuit, significant attention has been paid to understanding the biology of stems cells, whether embryological (ESCs) or induced pluripotent (IPSCs). Not only can ESCs/IPSCs be used to grow any tissue in theory, but IPSCs derived from a patient would, by definition, be perfectly compatible with the recipient's immune system.

Beyond the many outstanding technical challenges to such clinical application, practical considerations also exist. In particular, and perhaps surprisingly, it has been observed that stem cells grown in vitro develop at the same rate as the corresponding tissues in vivo (Barry *et al.*, 2017; Kanton *et al.*, 2019; Espuny-Camacho *et al.*, 2013; Maroof *et al.*, 2013; Gaspard *et al.*, 2008; Pollen *et al.*, 2019; Nicholas *et al.*, 2013). These slow developmental rates therefore necessitate equally long differentiation protocols, slowing both research into as well as future implementations of regenerative medicine (Saha and Jaenisch, 2009; Broccoli *et al.*, 2014).

The mirroring of developmental timing between in vitro and in vivo tissues, while a challenge, also suggests the existence of a developmental clock, intrinsic to species-specific genomes given the maintenance of the in vitro developmental timeline despite the lack of maternal signaling (Ebisuya and Briscoe, 2018). However, the presumed nature of this developmental clock

remains unknown and whether and to what degree the timing can be altered likewise remains unclear.

However, there is some indication that timings need not be fixed. For example, while human stem cells develop slowly as discussed, mouse stem cells develop in accordance with a 20-day gestation period (Gaspard *et al.*, 2008; Ying *et al.*, 2003; Shen *et al.*, 2006) despite the fact that many of these more rapidly developing tissues naturally perform analogous functions to their human counterparts. In particular, while mature neurons require several months to develop from human ESCs (hES), the same cells only take 5-14 days when derived from mouse ESCs (mES) (Nicholas *et al.*, 2013; Chuang *et al.*, 2013; Sun *et al.*, 2017; Shi *et al.*, 2012).

To expand our understanding of embryological developmental timing, we set out to test whether mature human neurons could be induced to develop according to an accelerated timeline similar to that observed in mouse cells. It was previously demonstrated that teratomas developed in a mouse host followed the original, slow developmental timeline, indicating that the mere presence of murine host factors was insufficient to accelerate development in human cells (Barry *et al.*, 2017). However, it remains unclear whether the transient signals expressed during murine development – which, by definition, were not present in the teratoma experiment – have the potential to accelerate the development of human cells. As such, we investigated whether human and mouse stem cells, grown in a chimeric co-culture, and driven towards neural differentiation would result in an accelerated developmental timeline for the human cells.

4.2 Results

4.2.1 Experimental design and data

Previously, a detailed RNA-sequencing (RNA-seq) time course of mouse and human pluripotent stem cells over three- or six-weeks of neural differentiation described, respectively, the drastically different species-specific rates of development *in vitro* (Barry *et al.*, 2017). Here, we set out to determine if co-differentiating human cells with mouse cells together could induce the human cells to differentiation at a quickened pace. Since hES cells are thought to more closely represent a post-implantation pluripotent stage, we used the similarly-staged mouse Epiblast stem (mEpiS) cells to compare with H9 hES cells (Brons *et al.*, 2007; Greber *et al.*, 2010; Tesar *et al.*, 2007). To identify cells from each species, we used mEpiS cells constitutively expressing cytoplasmic efficient green fluorescent protein (EGFP) and H9 cells expressing nuclear-localized H2B-mCherry (Figure 4.1).

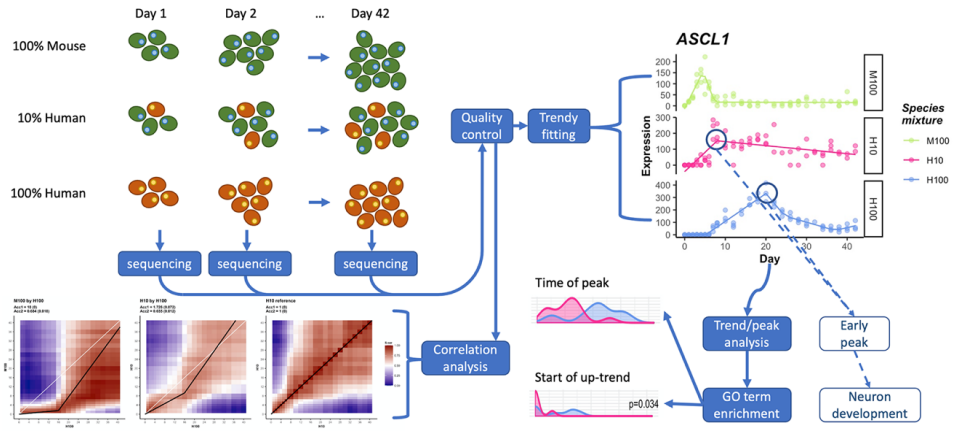


Figure 4.1: **Overview of data collection/analysis pipeline.**

(top left) Human (red) and mouse (green) cells are cultured at various mixing proportions over a 42 day neural differentiation time course; samples are harvested every 1-2 days for RNA-seq. Low quality biological replicates are removed from analysis and the data are normalized. (top right) Normalized data are fit to segmented regression built for RNA-seq data (Trendy) and temporal gene characteristics, such as peak times, are identified. (bottom right) Classified gene sets are further analyzed, including enrichment analysis for GO terms which are temporally accelerated or otherwise systematically altered in H10 compared to H100. (bottom left) In parallel to the previous analysis, normalized data are also correlated between time courses to identify transcriptome-wide effects.

To maximize any potential mouse-induced effects on human differentiation rate, we began by outnumbering human cells with the more quickly differentiating mouse cells in a ten-to-one ratio. 10% human co-cultured cells (H10), along with 100% mouse (M100) or 100% human (H100) control samples, were cultured under identical neural differentiation culture conditions (Brown, Barry, *et al.*, 2021) and samples in triplicate were collected for RNA-seq every 24 or 48 hours for six weeks (Figure 4.1). To minimize any confounding of results with known differences in cell cycle and cell fate choices due to differences in cell densities (Chetty *et al.*, 2013; D'Amour *et al.*, 2005; Bauwens *et al.*, 2008; Pauklin and Vallier, 2013; Roccio *et al.*, 2013), interspecies cell seeding confluencies were kept constant across species mixtures. After aligning transcripts to a combined human-mouse transcriptome to derive species-specific expression from the chimeric samples, samples passing quality control parameters (Supplemental Figure B.1) were processed for correlation analysis, fitted with gene expression patterns using the segmentation regression analysis R-package Trendy (Bacher *et al.*, 2018), and the timing of expression pattern changes was compared across samples (Figure 4.1).

Although mouse and human cells were singularized before seeding, time lapse microscopy revealed that, despite the clear occurrence of interspecies cell-cell interactions, cells preferentially clustered and proliferated with cells of their own species (Figure 4.2). Flow cytometry analysis revealed that while the intended starting cell ratios were seeded, as mouse cells differentiated quickly to become post-mitotic neurons, the still-proliferating human progenitor cells eventually overtook the culture. By day 12 of differentiation ~50% of H10 samples were of human composition, and by day 16 over 75% of samples were human cells (Supplemental Figure B.2).

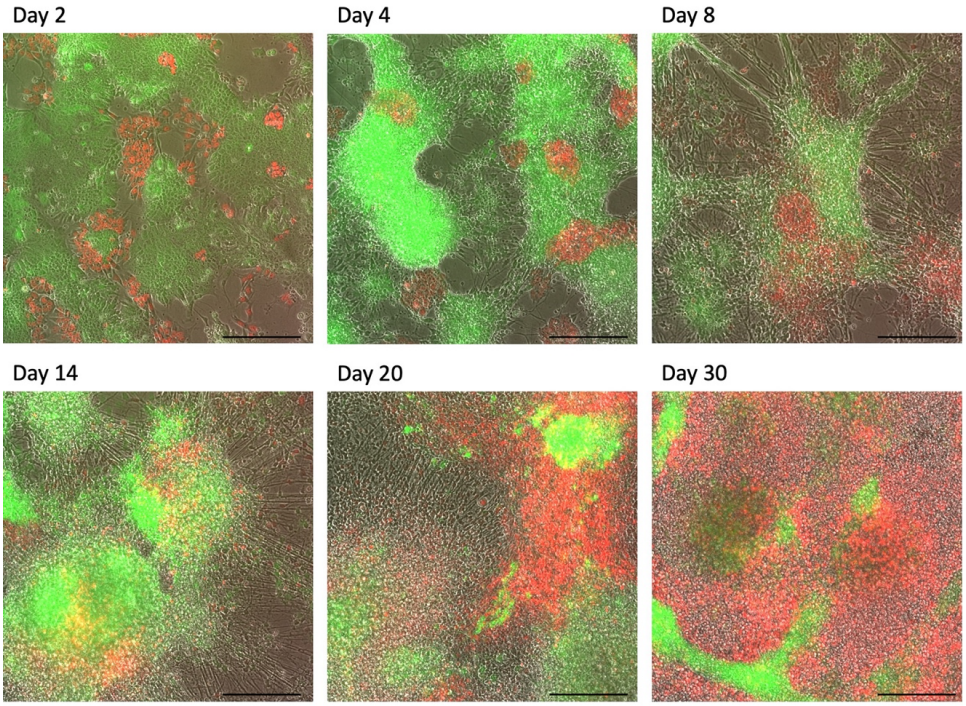


Figure 4.2: **Microscopy images of the H10 mixture across the time course.**

10% Human ES cells expressing nuclear-localized H2B-mCherry (red) were mixed with 90% mouse EpiS cells (expressing cytoplasmic GFP (green)) and co-cultured together under neural differentiation conditions for six weeks. Images captured at the time points indicated show clusters of associating human and mouse cells (red and green clusters respectively). Scale bars = 200μm.

4.2.2 Differential timing in the up-regulation of neural genes

To determine if gene expression patterns were accelerated in chimeric co-cultures, genes with fitted expression trends were compared between neural differentiation of human cells alone (H100) versus cells in a co-culture of 10% human cells mixed with 90% mouse cells (H10). We first asked if upregulated genes (genes trending up immediately or genes showing no change and then trending up) were upregulated earlier in mixed compared to control samples. Our bioinformatic analysis (See B.1 Statistical Methods) revealed that 929 genes were upregulated significantly earlier (begin up trending at least 2 days earlier) in H10 versus H100 samples, representing over 57% of all genes that trend up in both H10 and H100, excluding genes that begin to trend up on

day 0 in both cases (Figure 4.3A). We recognized several well-described neurogenic genes identified as accelerated in this early-upregulated category (Supplemental Figure B.3A), including genes involved in neural differentiation and migration (e.g., *STMN2, DCX, NEFL, NEUROG2, MYT1, MAPT*), neuronal signaling and synapse transmission (e.g., *SNAP25, SYT3, SYT4, SYN1*), neural stem cell identity (e.g., *FABP7, FGF10*), and glutamatergic and GABAergic neurons (e.g., *SLC1A3, GRIN2D, GABRA1*; Fig 3E). Therefore, genes from a seemingly wide range of neurodevelopmental functions were upregulated earlier under chimeric differentiation conditions.

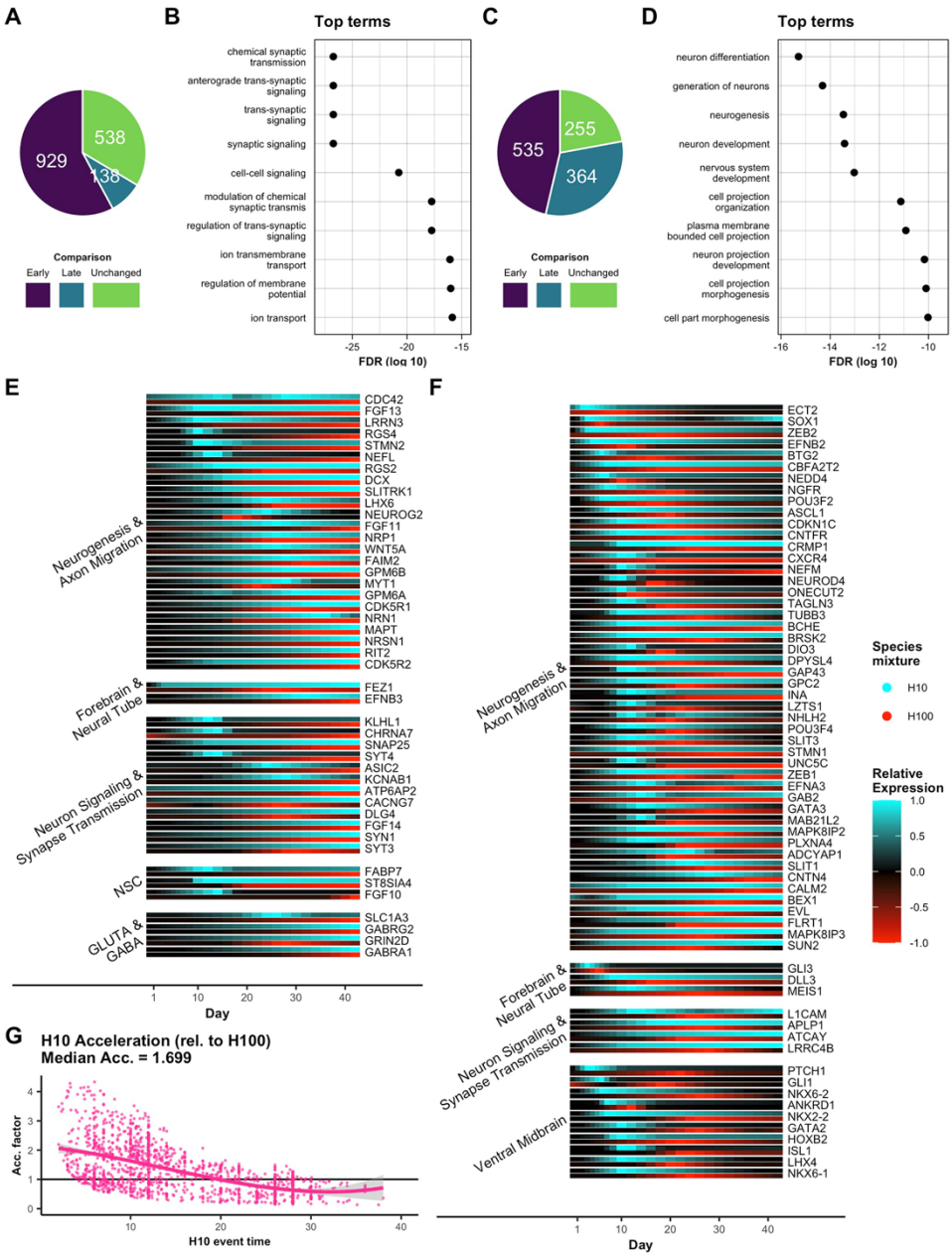


Figure 4.3: **Changes in neurodevelopmental gene expression are accelerated in human ES cells differentiated among mouse EpiS cells.**

(A) All genes which trend up in both H10 and H100 are classified as either early, late, or unchanged in H10 relative to H100 (omitting genes which already start up-trending at day 0 in both H10 and H100). (B) The top 10 most significant GO terms enriched for early upregulation in H10 demonstrated a clear pattern of acceleration in neuron and synaptic signaling-related genes (term enrichments shown as log10 adjusted p-values (FDR)). (C) All genes which peak in both H10 and H100 were classified as either early, late, or unchanged in H10 relative to H100. (D) The top 10 most significant GO terms enriched for early peaks in H10 showed acceleration of genes involved in neurogenesis and neuron development (term enrichments shown as log10 adjusted p-values (FDR)). (E) Relative expression plots of a curated subset of early-up (EU) genes collected into functional/regional groups. H10 (blue) and H100 (red) time courses were scaled such that 0 expression shows black and maximum expression between H10 and H100 shows as 1/-1 (within gene). (F) Relative expression plots of a curated subset of early-peak (EP) genes collected into functional/regional groups. (G) Genes with shared peaks or shared up-trends between H10 and H100 were used to compute acceleration rate point estimates as the ratio of H100 event times (e.g., time of peak in H100) to H10 event times. Point estimates were smoothed to give a continuous estimate of the H10 acceleration factor. The median fitted acceleration factor calculated over the first 16 days for H10 was given as 1.699.

Given that several recognizable neurogenic genes were among those identified as upregulated earlier in H10 compared to H100 samples (Supplemental Figure B.3A and Figure 4.3E), we set out to statistically test if early upregulated genes were specific to neural differentiation or a collection of genes within a random assortment of cellular processes. Functional GO-term enrichment (See B.1 Statistical Methods) of early-upregulated genes revealed that all of the ten most statistically significantly-enriched terms were associated with neuron and synaptic signaling (Figure 4.3B). In contrast, we did not observe neural-related GO term enrichment in genes upregulated later in H10 than in H100 (Supplemental Figure B.4A), confirming that neural genes were indeed specifically upregulated earlier in human cells co-differentiated with mouse cells.

In addition to the earlier upregulation of genes associated with neuron and synapse signaling, the duration of up-regulation was also significantly longer, often still trending upwards at the end point of the 6-week time course (Supplemental Figure B.5A). However, despite earlier onset of upregulation, their slopes were also significantly less steep than those of H100 samples (Supplemental Figure B.5A). These results indicate an earlier onset of synaptic signaling gene activation characterized by a more sustained, yet slower, rate of upregulation.

To ensure these results were not artifacts of transcript misalignment to the wrong species within the combined human and mouse reference transcriptome, we conducted an additional interspecies mixing time course experiment where intermixed cells were re-purified according to their species using Fluorescence-Activated Cell Sorting (FACS) (human-mCherry vs mouse-GFP) at each time point prior to RNA-isolation and sequencing. Importantly, post-sorted samples were aligned to the identical combined human and mouse transcriptome library. Sorted (s) sample

datasets are therefore labeled sH100, sH10, and sM100 to differentiate the sorted samples from the previously described data.

Computation of empirical misalignment rates showed low overall rates of misalignment across days (median 0.53% for sH100 and median 2.23% for sH10) (Supplemental Figure B.6A), and enrichment of misaligned transcripts failed to demonstrate any bias in neural-associated genes (Supplemental Figure B.6B). Although the sorted time course could not be performed in triplicate, nor at the same sampling frequency as the unsorted time course due to the extensive sort times necessary to collect enough cells to achieve sufficient read-depth, sorted sample expression analyses resulted in the same acceleration effects in sH10 relative to sH100 that we observed in unsorted samples (Supplemental Figure B.6C-E), confirming that our earlier detection of acceleration was not due to species-misaligned transcripts.

4.2.3 Differential timing in the peaks of transient expression of neural genes

During development, genes involved in neural differentiation are often not simply turned on, but rather are expressed in temporally-regulated dynamic patterns (Gurok *et al.*, 2004; van de Leemput *et al.*, 2014). To determine if genes with coordinated expression profiles were regulated more quickly, we next tested whether genes with peak expression profiles (consecutive up-down or up-flat segments) peaked earlier under chimeric versus human control conditions (See B.1 Statistical Methods).

Overall, we identified 535 genes that peaked earlier (at least two days) in chimeric culture conditions compared to control samples, representing over 46% of all peaking genes identified in the time course (Figure 4.3C). Similarly to early-upregulated genes, we recognized several peaking genes involved in neural development in the accelerated peak category (Figure 4.3F and

Supplemental Figure B.3B), including genes involved in neurogenesis (e.g., *ASCL1, NGFR, NEFM, TUBB3*), neural tube development (e.g., *MEIS1, GLI3, DLL3*), neuron signaling (e.g., *SNAP25, ATCAY*), and ventral midbrain differentiation (e.g., *ISL1, LHX4, NKX6-1*). We further validated that genes involved in neurodevelopment were specifically peaking early through GO-term enrichment analysis, and we found that all of the top ten most significantly enriched terms were associated with neural development (Figure 4.3D), whereas no obvious trend in neural-related GO-terms was found for genes with delayed peaks (Supplemental Figure B.4B). In contrast to early-upregulated genes that were enriched in neuron and synaptic signaling, early peaked genes were involved in neurogenesis, neuron projection development, and neuron differentiation (Figure 4.3D). Further, whereas early-upregulated genes had a slower rate of increase compared to control cells, early peaked genes exhibited an earlier time of start of upregulation towards the peak and a faster rate of upregulation to reach the peak (Supplemental Figure B.5B). Taken together, we report that the regulation of neurogenic genes was specifically accelerated in H10 compared to H100.

To quantify the degree of acceleration and investigate if acceleration was variable or uniform across the time course, we considered genes with shared peaks or shared up trends in both H10 and H100 and computed acceleration factors as the percent difference in time to peak or the start of up regulation in H10 compared to H100. Smooth regression of these point estimates provided a continuous estimate of the relative acceleration between H10 and H100 (Figure 4.3G, see B.1 Statistical Methods).

From this analysis, we uncovered that the majority of acceleration was in fact not constant over the course of co-differentiation. Rather, the majority of acceleration takes place during the first 16 days. While the median acceleration factor (reflecting fold-change acceleration of expression events) during this time was 1.699, acceleration varied from a maximum factor of 2.75

at the earliest stages of differentiation, converging eventually to a factor of 1 (non-accelerated) by day 20 (Figure 4.3G). It is notable that this gradual reduction in acceleration rate occurs concurrently as human cells begin out-proliferating post-mitotic mouse neurons (Figure 4.2 and Supplemental Figure B.2). Human cells start outnumbering mouse cells at day 12, the time at which acceleration effects dissipate, suggesting a correlation between mouse cell number and the acceleration effect they induce in co-culture (Figure 4.3G and Supplemental Figure B.2).

4.2.4 Chimeric co-culture affected the timing and expressions levels of some genes associated with neuron or brain region identity

Our neural differentiation protocol recapitulates a general neural developmental program and produces neurons of various regional identities (Barry *et al.*, 2017). To determine if chimeric co-culture of hES cells would affect cell lineage outcomes, we identified genes that were most differentially expressed (measured as fold change between maximum expression along the time course) in chimeric mixed samples compared to hES cell controls (Supplemental Figure B.8).

We observed some changes in the expression of transient signals as well as changes in sustained region-specific expression. For example, certain genes associated with the anterior dorsal neural tube showed earlier downregulation in H10 compared to H100, whereas genes linked to Gluta- and GABAergic neurons and neuron signal transduction showed patterns of downregulation at later times (Supplemental Figure B.8). Other genes broadly associated with neurogenesis show a mixture of these patterns. In contrast, some genes associated with the ventral midbrain showed transient upregulation in chimeric mixed samples compared to control samples (Supplemental Figure B.8). These effects would be consistent with an early exposure of *Shh* from mouse cells that could have triggered a cascade of downstream effects on gene expression, including *FOXA2*

(Bayly *et al.*, 2012), *NKX2.1*, and *PHOX2B* (Supplemental Figure B.8) (Dias *et al.*, 2020; Dessaud *et al.*, 2008). Our analysis therefore revealed that some genes associated with neuron cell type and regional identity were temporally and/or differentially expressed under chimeric conditions.

To verify that the acceleration effects described in this report were not largely due to a general shift towards neural cell types that appear earlier in development rather than a true acceleration, we performed a deconvolution analysis of H100 and H10 samples to monitor the appearance of various progenitor and intermediate cell stages and their differentiation over time (See B.1 Statistical Methods). This type of analysis estimates the relative proportions of cell types that may be present in a bulk sample by comparing bulk expression to a reference of purified or annotated single cell data. We compared our data to the CoDEx dataset of annotated single-cell sequencing from the developing human cortex (Polioudakis *et al.*, 2019) with the MuSiC R package (Wang *et al.*, 2019). Smoothed estimates of neural stage proportions indicated that co-cultured human cells mirrored the developmental progression of cell types of control samples, but at an accelerated pace (Supplemental Figure B.9). Specifically, similar progenitor-to-mature neural cell markers appeared in the same order in H10 and H100 (Supplemental Figure B.9), yet high proportions of excitatory neurons in H10 occurred earlier (days 12-16) compared to H100 (days 18-24). Taken together, although differential expression analyses identified changes in expression levels of some genes implicated in nervous system development, differentiation followed similar lineage pathways but at accelerated rates in chimeric compared to unmixed conditions.

4.2.5 Dose-dependency relationship between mixture proportions and human developmental timing

If the acceleration of hES cell differentiation was indeed mouse cell-induced, we reasoned that the rate of acceleration would be dose-dependent on the amount of mouse cells present in human co-cultures. Harnessing the data from multiple initial interspecies mixing proportions (0%, 10%, 85%, and 100% human vs mouse), we tested the dependence of the initial mixing proportions on the acceleration rate observed (Figure 4.4A).

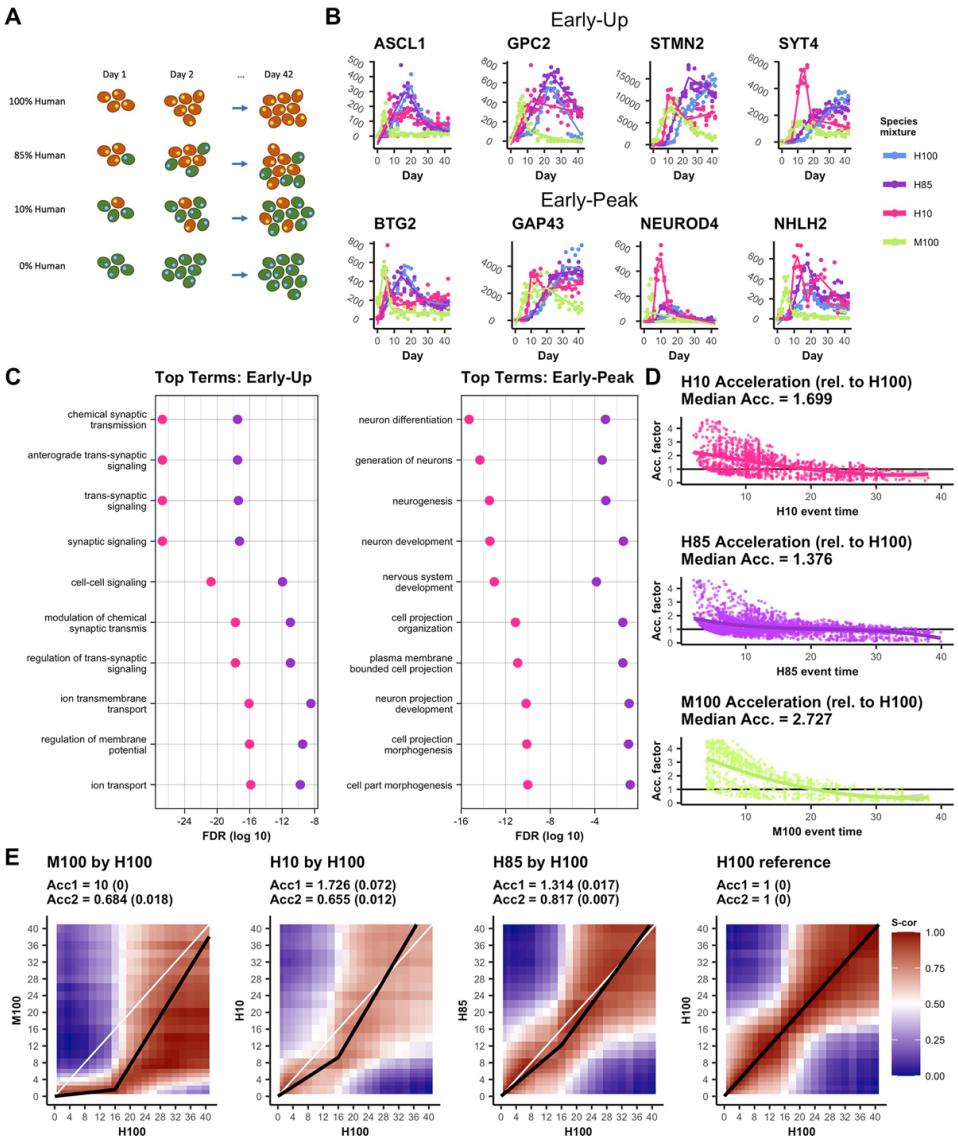


Figure 4.4: **Variable mixing proportions show a dose response of acceleration effects.**

(A) An additional, intermediate interspecies mixing proportion, H85(M15), was compared to H0(M100), H10, and H100 time courses. (B) Expression plots of curated EU and EP genes with fitted trend lines (solid) for H100 (blue), H85 (purple), H10 (red), and M100 (green). Observed, normalized data are also plotted (dots). (C) Top 10 EU and EP GO terms from H10 showing relative significance of term enrichment for H10 and H85. (D) Smoothed acceleration factors are calculated between each of H10, H85, and M100 (human orthologous genes) against H100 using the method in Figure 4.3 G (B.1 Statistical Methods). The median fitted acceleration from the first 16 days is reported. (E) Correlation (Spearman) heat maps where regions of high correlation (red) below the diagonal indicate accelerated activity where later days in H100 are correlated with earlier days in the comparison mixture. Correlations are calculated on a subset of highly dynamic genes (see Materials and Methods).

Overall, expression profiles of a selection of key neuronal genes with either early up-regulation or early peaks in H85 samples were chronologically intermediate between H10 and H100 expression profiles (Figure 4.4B). Overlaying these trends with the expression profiles of orthologous genes in the M100 sample reveals progressively later onsets of gene up-regulation/peaks with decreasing proportions of mouse cells among these genes (Figure 4.4B).

To determine whether these results extended to the broader set of neuron-associated genes, we replicated the GO-term enrichment analysis in the H85 sample. Testing term enrichment on those genes which either upregulated or peaked earlier in H85 relative to H100 resulted in a list of the most significant terms with the same patterns as in H10. However, comparing term significance levels between the top 10 most significant terms in the H10 analysis and their H85 counterparts shows that, while the H85 terms were still highly significant, they were less so than the H10 terms (Figure 4.4C). Further, direct computation of acceleration factors of the first 16 days, based on differences in shared peak times and the starts of up trends, resulted in progressively decreasing calculated accelerations with stepwise drops in percentage mouse cells: 2.727 (H0/M100), 1.699 (H10/M90), and 1.376 (H85/M15), consistent with a dose-response effect on acceleration (Figure 4.4D).

Pairwise correlations allowed us to further aggregate relative expression trends across genes. We took a subset of genes, targeting those with dynamic expression over time, and plotted correlations calculated between pairs of time points relative to H100 (Figure 4.4E). Mouse orthologs demonstrate a visually significant acceleration with day 2 expression being highly correlated with H100 out to day 16. The H10 and H85 time courses both showed visual acceleration with regions of high correlation below the diagonal, but with respectively lower magnitudes as the proportion of mouse cells decreases. Adapting a technique for estimating

acceleration factors from these correlation plots described in Rayon et al.(Rayon *et al.*, 2020) (see B.1 Statistical Methods) allowed us to compute average acceleration over the first 16 days independently of peaks or other expression events. We observed similar acceleration dynamics with correlation-based acceleration factors of 1.726 over the first 16 days for H10 and 1.314 over the first 16 days for H85 (Figure 4.4E). These correlation results, while dependent on specific geometries of the correlation plots, are themselves supported by comparing to the deconvolution of H10, H85, and H100, a procedure which is similarly based on a large panel of dynamic genes. In the deconvolution analysis, we observed higher proportions of mouse cells in mixtures resulting in the progressively earlier sequential maturation of progenitor and intermediate cell types (Supplemental Figure B.9).

4.2.6 Differential correlation with in vivo control tissues

We compared our data with the *Brain Span* human fetal sample references to assess if our *in vitro* acceleration is consistent with sample maturity *in utero* (Sunkin *et al.*, 2013; Miller *et al.*, 2014; Allan Human Brain Atlas: BrainSpan (Atlas of the Developing Brain)). We calculated correlations between our observed in vitro data and five tissue regions from the *Brain Span* database across weeks 8, 9, and 12 of development (See B.1 Statistical Methods). Across all time points and tissues, our mixed H10 and H85 samples increased correlation with the *Brain Span* reference earlier than the H100 control in a manner that was dose-dependent (Figure 4.5A).

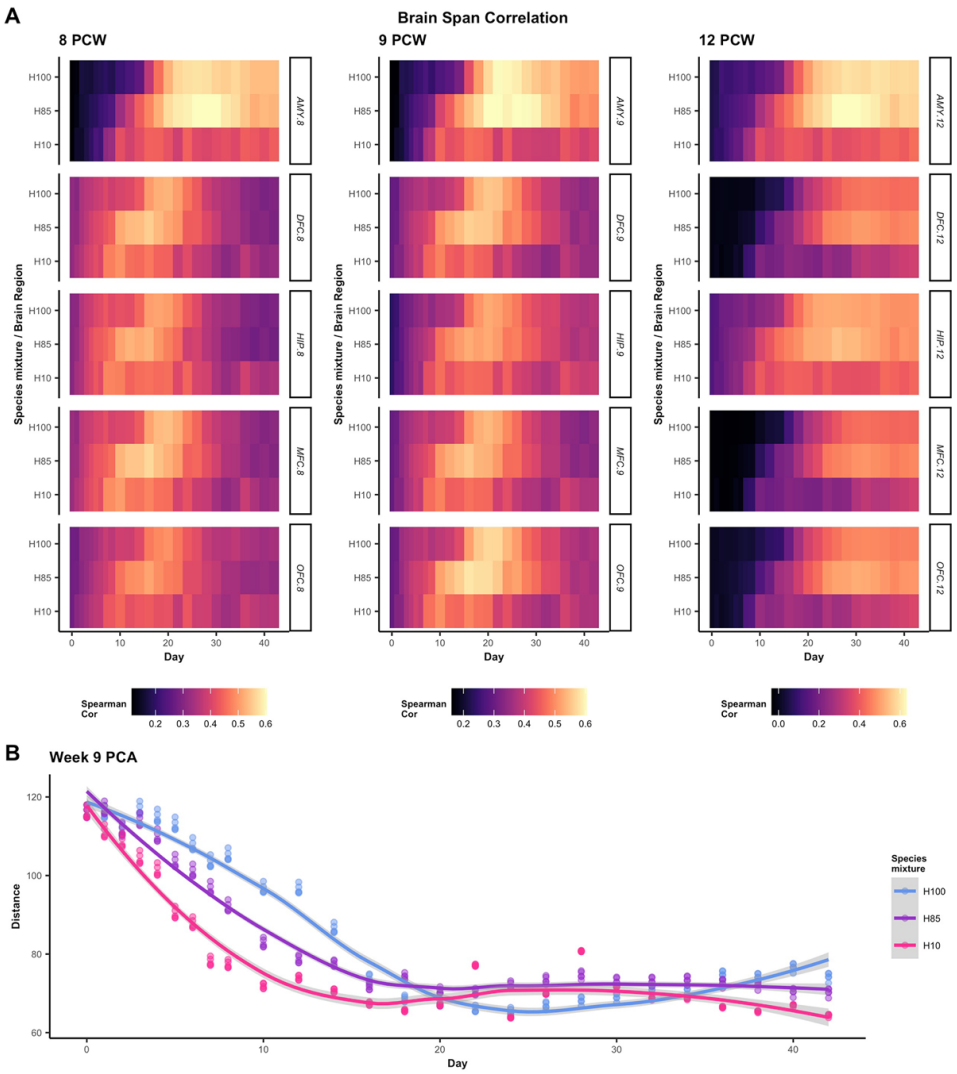


Figure 4.5: **Comparison with Brain-Span regions further demonstrates a dose-response in acceleration effects.**

(A) Correlations (Spearman) between fitted trends and Brain-Span data are calculated at three Brain-Span time points and across the five brain regions represented at all three times. Calculations are performed on a subset of highly dynamic genes (see Materials and Methods). (B) Dissimilarity (PCA-based distance, see Materials and Methods) between species mixtures and each of the 5 reference brain regions are computed for each day and smoothed to estimate a continuous dissimilarity metric over time.

A complementary analysis based on a variation of principal component analysis (PCA) (Townes, 2019; Townes *et al.*, 2019) replicates these findings. Dimension reduction of the gene expression data allows the distance between the representation of Brain Span references and the representations of our experimental data to be interpreted as a dissimilarity metric (See B.1

Statistical Methods). Smoothing across regions for the week 9 reference, we observed that H10 minimizes dissimilarity between days 12-16, which is before H85 (days 16-20), which is further before H100 (days 20-24) (Figure 4.5B). Accelerated correlation to *in vivo* data was also confirmed through a similar analysis of annotated brain tissue from the *Human Protein Atlas* (Yu *et al.*, 2015; RNA FANTOM brain region gene data) (Supplemental Figure B.10), consistent with a genome-wide neural program that is activated earliest in M100, then significantly accelerated in H10, followed by moderately earlier in H85, and latest in H100 samples.

While we leave the determination of mechanisms responsible for regulating the developmental clock to future work, comparisons of accelerated genes with curated gene sets allowed us to speculate on candidate pathways and transcription factors/miRNAs that may be involved (Subramanian *et al.*, 2005; Liberzon *et al.*, 2015; Kanehisa, 2000, 2019; Kanehisa *et al.*, 2021; Nishimura, 2001; Jassal *et al.*, 2020; Chen and Wang, 2020; Yevshin *et al.*, 2019). Enrichment of the differences between up-trend/peak times (See B.1 Statistical Methods) identified signaling pathways activated earlier in both H10 and in M100 compared to H100 samples, including G-protein coupled receptor (GPCR) signaling pathways and miRNA-regulated pathways MAPK/ERK (MIR4801, MIR4731) (Gholizadeh *et al.*, 2020; Alahverdi *et al.*, 2020), and PI3K/AKT (Gao *et al.*, 2020; Zhihong *et al.*, 2020), which may play roles driving developmental rates (Supplemental Figure B.11). We also identified developmental regulators of interest, such as *NSRF,* a master neural developmental regulator essential for gastrulation that may also influence the expression of thousands of genes during development (Wang *et al.*, 2012; Schoenherr and Anderson, 1995; Thompson and Chan, 2018; Bruce *et al.*, 2004) and *OCT1*, an essential regulator of development that plays crucial roles in the earliest cell fate decisions during

embryonic development (Sebastiano *et al.*, 2010; Shen *et al.*, 2017; Perovanovic *et al.*, 2020), that may warrant further investigation.

4.3 Discussion

In this study, we report for the first time multifaceted effects of interspecies mixing on the differentiation of hES cells. Through comprehensive RNA-seq time courses, we uncover that co-differentiation of hES cells intermixed with mEpiS cells was sufficient to accelerate components of neural gene regulatory programs, and identified genes with roles in neural lineage and regional identities that were both temporally and differentially expressed. We went on to demonstrate that the acceleration effect was dose-dependent on the starting ratio of interspecies cells (Figure 4.4), and that the chimeric samples correlated to *in vivo* tissue samples earlier in the differentiation time course than human samples alone (Figure 4.5 and Supplemental Figure B.10).

Previously, it was reported that the faster differentiation of mouse cells compared to human cells may be in part caused by increased speed of transcriptional upregulation of genes, indicated by steeper slopes in gene expression over time (Barry *et al.*, 2019). Consistent with a mouse cell-induced acceleration of human cell neural differentiation, here we found that the slopes of peaked genes in human cells co-differentiated with mouse cells were also significantly increased in accelerated genes compared to control samples (Supplemental Figure B.5B). However, non-peaking, mostly monotonic, genes whose upregulation began earlier showed lesser slopes in chimeric samples, despite starting their upward trend significantly earlier and often continuing upwards for the duration of the time course (Supplemental Figure B.5A). These results may suggest different functional roles of early-upregulated monotonic genes compared to genes with peak expression profiles. Indeed, genes with increased slopes and earlier peaks were significantly

enriched in processes of generation of neurons and neuron cell projections, whereas earlier upregulated monotonic gene trends with lesser slopes were enriched in neuron and synaptic signaling events (Figure 4.3). Although we identify differences in gene expression profiles in our time course in this report, the functional maturity of resulting neurons in control versus chimeric co-differentiation conditions remains to be determined.

The mechanisms regulating developmental tempos and how interspecies co-culture might affect the differentiation speed of another species remain unknown. Although cells from different species exhibit different cell cycle rates, and counting rounds of cell division has been proposed as a possible mechanism for a cell's ability to track developmental time (Temple and Raff, 1986), multiple reports also suggest that cell division is not required for differentiation in a number of systems(Gao *et al.*, 1997; Burton *et al.*, 1999; Harris and Hartenstein, 1991). Cell size is also unlikely to regulate developmental speeds as many cell types are of similar sizes across species with drastically different developmental rates (Savage *et al.*, 2007). Another intriguing possibility is that metabolic rates, sometimes related to cell size, cell cycle, and mammalian body mass (Savage *et al.*, 2007), could directly modulate species-specific developmental timing (Brown *et al.*, 2004; Hamilton *et al.*, 2011; Miyazawa and Aulehla, 2018). However, when removed from the body and placed into tissue culture, cells from different species exhibit similar metabolic rates, indicating variable metabolic rates are unlikely to account for the species-specific developmental speeds retained *in vitro* (Brown *et al.*, 2007; Wheatley and Clegg, 1994). Genome size similarly does not seem well-correlated to developmental time across mammalian species (Kasai *et al.*, 2013; Árnason *et al.*, 2018).

Recently, elegant *in vitro* models of mouse and human segmentation clocks with species-specific timing have been reported and are being used to study factors affecting developmental

time (Matsumiya *et al.*, 2018; Chu *et al.*, 2019; Matsuda, Yamanaka, *et al.*, 2020; Diaz-Cuadros *et al.*, 2020). Two recent papers have identified a correlation between some biochemical reaction rates (e.g. protein stability and turnover rates) and developmental tempos (Matsuda, Hayashi, *et al.*, 2020; Rayon *et al.*, 2020), although if, or to what extent, intrinsic developmental clocks could be altered was not determined (Matsuda, Hayashi, *et al.*, 2020). Here, we show that cell-cell signaling alone is sufficient to affect the developmental clock. Further, we identified candidate signaling pathways and regulators activated earlier in both H10 and in M100 compared to H100 samples that may warrant future investigations (Supplemental Figure B.11).

Previously, several studies suggested that the intrinsic species-specific developmental timer was faithfully retained under various conditions, including 2D vs 3D culture methods (Nicholas *et al.*, 2013; Marchetto *et al.*, 2019; Pollen *et al.*, 2019; Lancaster *et al.*, 2013; Kelava and Lancaster, 2016) and interspecies transplant/implantation studies into adult hosts (Barry *et al.*, 2017; Espuny-Camacho *et al.*, 2013; Maroof *et al.*, 2013; Nicholas *et al.*, 2013). While these studies revealed that non-embryonic interspecies conditions were insufficient to alter developmental time, in this study we demonstrate that factors actively driving an embryonic developmental program from pluripotency, rather than a mature host environment, can be sufficient to affect components of the developmental clock of cells from another species.

The ability of stem cells of different species to resolve conflicting developmental speeds has significant implications in the development of chimeric embryos for human organ formation (De Los Angeles *et al.*, 2018). With a widespread shortage of immunologically-matched organs for patients in need of organ transplants, the ability to grow transplantable human organs through human stem cell chimeric contributions to embryos remains an interesting potential therapeutic approach (J. Wu *et al.*, 2017; Das *et al.*, 2020). However, many barriers remain, including poor

human chimeric contributions, possibly in part due to the vastly different developmental rates between neighboring cells of different species (De Los Angeles *et al.*, 2018; Ebisuya and Briscoe, 2018; Masaki *et al.*, 2015). In this study, we demonstrate that it is possible for mouse cells to influence developmental rates and outcomes of neighboring human cells.

Previous reports of successful human cell contributions to chimeric mammalian embryos (Mascetti and Pedersen, 2016; J. Wu *et al.*, 2017; Yang *et al.*, 2017), including a recent report of the highest contribution (4%) of human cells in mouse-human chimeric embryos (Hu *et al.*, 2020), could imply that human pluripotent stem cells may be induced to accelerate their developmental rate to match that of their embryonic host species. However, maturation rates of human cells in interspecies chimeras have not been well characterized. Our comprehensive time course results in this study indicate that human developmental time could be accelerated by co-differentiating cells within chimeric embryos, although collateral impacts in cell lineage outcomes may occur. In the case of neural differentiation in this study, we did find genes involved in dorsal forebrain development, for example, that were temporally downregulated in interspecies samples while genes involved in ventral midbrain development were upregulated, likely, at least in part, due to an earlier and increased exposure to *Shh* (Figuress 4.3-4.5) (Placzek and Furley, 1996; Dale *et al.*, 1997; Lupo *et al.*, 2006). Importantly, mouse and human brains do not share identical brain physiologies, cell type compositions, nor brain region proportions (Hodge *et al.*, 2019; Sjöstedt *et al.*, 2020), so it is perhaps not surprising that some altered cell fate choices are made when cells are exposed to signals intended to created divergent outcomes. Thus, it will be important to monitor cell outcomes in chimeric embryos for human organ growth to verify that cell type contributions and organ functions are not affected.

Although the protocol described here will not have clinical applications due to the xenotropic nature of the conditions, it does suggest that the human developmental clock can be accelerated; and while specific factors involved and clock mechanism itself remain to be dissected, our proof-of-concept study provides evidence that the species-specific developmental clock may be amenable to acceleration for clinically-relevant benefit.

4.4 Publication information

The methods and results described in this chapter are published in Brown, Barry, et al., 2021 (Brown, Barry, *et al.*, 2021).

A Appendix to Chapter 2

A.1 Estimation of gene-specific CDFs by cLAD regression

This section is edited for clarity from supplement S1.4 from Brown *et al.* 2021 (Brown, Ni, *et al.*, 2021).

Chapter 2.2.2 mentions that the initial values of the means, $\lambda_{k,}$ are taken from equal spacings along an estimate of the eCDF of the counts y_j at $LS\ \delta=1$. Unfortunately, estimation of this eCDF is non-trivial; standard formulas for the eCDF are confounded by the need to remove the effect of LS prior to computation and, more importantly, are further confounded by the truncation of y_j at 0.

To solve this problem, we demonstrate that this eCDF can, in fact, be efficiently calculated for each gene by a modified application of cLAD regression (Powell, 1984, 1986) at a carefully chosen grid of given intercepts. cLAD regression has the benefit here of solving for linear functions of quantiles in data – hence its alternate name, quantile regression. The pairing of fitted intercepts (quantiles) and corresponding percentiles then define points along the desired eCDF.

Regression calculations are conducted on the log-log scale, so the both the *LS* values, δ_i, and the counts, y_j, are log transformed with a floor of $\log(0.999)$ for the observed counts; $d_j = \log(\delta_j)$, $z_j=max\{log(y_j), \log(0.999)\}$. The implicit addition of a ≈ 1 pseudo count to the observed zeros prior to log transformation is motivated by the fact that computations are performed using censored quantile regression, which can be written as standard quantile (LAD) regression on the subset of data to the right of the intersection of the regression curve and censoring threshold (Powell, 1986). In particular, the regression model for the observed data is:

$$z_j = max\{log(0.999), \beta_0 + d_j + \epsilon_j\}$$

where β_0 is some gene-specific intercept and ϵ_j is a random error term following some, possibly complex distribution. For example, in the presence of population heterogeneity, the ϵ_j might follow a bi-modal distribution. Placing the zeros at log*(0.999)* allows the regression solutions to be defined in terms of the strictly positive data as the censoring threshold is placed just below log*(1)*. The choice of a constant slope term (the implicit coefficient of 1 on the *LS* term), while mathematically convenient in the following, is also not unreasonable. On the multiplicative (log) scale, expression should have about a slope one relationship with *LS* across all genes, recalling the above comment that both counts and *LS* are log-transformed for the cLAD regression.

A.1.1 **Estimation of expression distribution quantiles**

The gene-specific eCDFs to be computed are denoted by vectors of quantiles and the associated estimated percentiles. To calculate the sample quantiles and percentiles of the eCDF, and as a natural extension of the censored linear model of the data distribution, the mathematics of censored least absolute deviations regression (cLAD) are adopted. The cLAD minimization problem is

$$\min_{\beta_0} \left\{ \sum_j p_\pi [z_j - (\beta_0 + z_j)^*] \right\}$$

where $p_\pi(x)=(\pi - I(x<0))(x)$, $(x)^*=max\{\log(0.999), x)$, and π is some percentile (eg. $\pi=0.5$ for median quantile regression). It has previously been shown that cLAD regression has beneficial properties, particularly consistency in the presence of only weakly defined residual distributions (Powell, 1984, 1986). Additionally, study of the solution set to cLAD regression has shown that solution coefficients define lines which pass through at least as many data points as there are free

parameters (Branham, R. L., 1982). Given this, the regression problem can be significantly reduced. Given a fixed percentile, π, the set of possible solutions is confined to a set of parallel lines – one intersecting each point in the data set – which is finite for finite data. The regression problem is then to determine which line, uniquely defined by its intercept β_0, minimizes the loss function.

As such, the set of quantiles defining the eCDF is easily defined. In particular, the eCDF quantiles are the set of intercepts, β_{0j}, defining the lines passing through each of the data points. The intercepts/quantiles refer to the previously mentioned "grid of given intercepts." Since this set of intercepts contains the minimizing solution regardless of the choice of π in the regression problem, any additional quantiles would by definition be redundant. Specifically, the set of quantiles $\{q_j\}$ is defined as:

$$\{q_j\} = \{\beta_{0j}\} = \{z_j - d_j\}$$

A.1.2 Estimation of expression distribution percentiles

For simplicity, consider a highly expressed gene for which there are no observed zero counts prior to log transformation and suppose without loss of generality that the quantiles $\{q_i\}$ are all unique and that the indices j are of decreasing order such that $q_j > q_{j+1}\ \forall j$, $q_1 = max\{q_j\}$. In this case, computing the percentiles associated with each quantile is trivial and follows the standard eCDF formula:

$$p_j = \frac{1}{n}\sum_i \mathbb{I}(q_i \le q_j) = \frac{n-j+1}{n}$$

where $n = |\{q_j\}|$. The more general case where a gene may contain zeros, possibly many zeros, is more complicated. However, given the assumption that the residual distribution (ϵ_j) is constant in *LS*, a simple and analogous solution exists:

$$p_j = \frac{n_j - \left|\{q_i\}_j\right|}{n_j}$$

where

$$\begin{aligned} \Lambda_j &= \{i: \delta_i \geq \delta_j - (z_j - log(0.999))\} \\ n_j &= |\Lambda_j| \\ \{q_i\}_j &= \{q_i: q_i > q_j, i \in \Lambda_j\} \end{aligned}$$

The linear interpolation of the set $\{(\mathrm{q_j}, \mathrm{p_j})\}$ defines an eCDF from which the π_j can be initialized.

A.1.3 Monotonicity correction

In practice, this estimate of the percentiles p_j can become unstable for high j (when n_j becomes small). Additionally, some forms of population heterogeneity can cause estimation problems, especially as they can violate the assumption of constant residual distribution. Here we derive the above formulation for the percentiles, p_j, as well as corrections for these situations.

Recall the example of the highly expressing gene mentioned above. As noted, it is trivial to estimate p_j, and guaranteed that the p_j will be both unique and monotone. To facilitate the discussion of the more general case (where there are zero counts), consider this ideal problem (no zeros) in the context of standard LAD regression. For notational convenience, make the following definitions:

$$\begin{aligned} SUD_j'(\beta_{0j}) &\coloneqq \sum_{\{i: z_i > (\beta_{0j} + d_i)\}} [z_i - (\beta_{0j} + d_i)] \\ SLD_j'(\beta_{0j}) &\coloneqq \sum_{\{i: z_i \leq (\beta_{0j} + d_i)\}} [(\beta_{0j} + d_i) - z_i] \end{aligned}$$

where the abbreviations denote “sum of upper deviations” and “sum of lower deviations” respectively across the collection of cells indexed by i.

To solve for a percentile given a quantile, the previous regression problem is inverted – and expanded to remove the $p_\pi(\cdot)$ notation – to find the set

$$\left\{\pi: \beta_{0j} = \min_{\beta_0} \left(\pi SUD_j'(\beta_0) + (1-\pi) SLD_j'(\beta_0)\right)\right\}$$

Note that $SUD_1'=0$ and $SLD_n'=0$ and that SUD_j' (SLD_j') are increasing (decreasing) in j. Additionally, a linear interpolation of SUD_j' (SLD_j') would have positive (negative) derivatives. Thus, the surface of the convex hull of the set $\{(SUD_j', SLD_j')\}$ contains all points within the set. This means that for each β_{0j}, there exists some π_j for which β_{0j} is the unique minimizer of the usual, non-inverted, LAD problem. Specifically, a minimizing π_j is one such that the line $\pi_j SUD_j'-(1-\pi_j)SLD_j'=c$, for some constant c, is a sub-tangent of the linear interpolation of $\{(SUD_j', SLD_j')\}$ at the relevant point.

This can be demonstrated as follows: suppose $-\pi_j/(1-\pi_j)$ is a slope in the sub-derivative of the linear interpolation at point (SUD_j', SLD_j') so that the convex combination of sums of deviations takes some value, $\pi_j SUD_j'+(1-\pi_j)SLD_j'=b$. Consider then the point (SUD_{j+1}', SLD_{j+1}') and define $(dU,dL) := (SUD_{j+1}', SLD_{j+1}') - (SUD_j', SLD_j')$ where $dU,-dL>0$ so the convex combination at $j+1$ can be written as $\pi_j(SUD_j'+dU)+(1-\pi_j)(SLD_j'+dL)=b+\pi_j dU+(1-\pi_j)dL$. Since $-\pi_j/(1-\pi_j)$ is in the sub-derivative, it is the case that

$$-\frac{\pi_j}{(1-\pi_j)} \leq \frac{dL}{dU} \Rightarrow \pi_j dU + (1-\pi_j) dL \geq 0$$

with equality only if $-\pi_j/(1-\pi_j)$ is the maximal sub-derivative, showing $\beta_{0(j+1)}$ is not a solution for the LAD problem given weight π_j excepting only the minimal π_j allowed by the sub-derivative. A similar result holds for point $j-1$.

Therefore, to solve for a percentile p_j, one can consider the derivatives of the sums of deviations parameterized by the intercept β_{0j}

$$dSUD'_j = \frac{d}{d(\beta_{0j})} SUD'_j = -\left|\{j: z_j > (\beta_{0j} + d_j)\}\right| = -(j-1)$$
$$dSLD'_j = \frac{d}{d(\beta_{0j})} SLD'_j = \left|\{j: z_j \leq (\beta_{0j} + d_j)\}\right| = (n-j+1)$$

where the second equalities follow from the uniqueness and ordering of the indices *j*.

Then the slope of one possible subtangent in terms of π_j is

$$-\frac{\pi_j}{(1-\pi_j)} = \frac{dSLD'_j}{dSUD'_j} = -\frac{(n-j+1)}{(j-1)}$$

so

$$p_j = \pi_j = \frac{-\frac{dSLD'_j}{dSUD'_j}}{1-\frac{dSLD'_j}{dSUD'_j}} = \frac{(n-j+1)}{(j-1)+(n-j+1)} = \frac{n-j+1}{n}$$

which is the same result as above.

To generalize solving for percentiles $\{p_j\}$ in the context of cLAD regression, one need only make a few modifications to the above results. First, define

$$SUD_j(\beta_{0j}) = \sum_{z_i \geq (\beta_{0j}+d_i)^*} \left[z_i - (\beta_{0j} + d_i)^*\right]$$
$$SLD_j(\beta_{0j}) = \sum_{z_i < (\beta_{0j}+d_i)^*} \left[(\beta_{0j} + d_i)^* - z_i\right]$$

for the censoring function $(x)^*$. Then the percentile problem seeks to find sets of a familiar form:

$$\left\{\pi: \beta_{0j} = \min_{\beta_0} \left(\pi SUD_j(\beta_0) + (1-\pi) SLD_j(\beta_0)\right)\right\}$$

with a familiar solution,

$$-\frac{\pi_j}{(1-\pi_j)} = \frac{dSLD_j}{dSUD_j}$$

assuming the same convexity conditions hold for the set $\{(SUD_j, SLD_j)\}$.

In the presence of censoring, the convexity conditions may not hold. A correction to enforce convexity is discussed later in the section. The main difference between the general result

which accommodates censoring and that for LAD regression is in the precise formulation of the derivatives of the sums of deviations.

$$\begin{aligned} dSUD_j &= \frac{d}{d(\beta_{0j})} SUD_j \\ &= -\left|\{j: z_j > (\beta_{0j} + d_j)^*, d_j \geq \beta_{0j} - log(0.999)\}\right| \\ dSLD_j &= \frac{d}{d(\beta_{0j})} SLD_j \\ &= \left|\{j: z_j \leq (\beta_{0j} + d_j)^*, d_j \geq \beta_{0j} - log(0.999)\}\right| \end{aligned}$$

These derivatives do have a similar interpretation to those of the previous section, however. Specifically, up to a sign change, they are the number of observations above/below the regression line under consideration which *also* have *LS* above the point where that regression line hits the censoring threshold of log*(0.999)*.

This gives a convenient interpretation to the solution for p_j as well. The solution itself is

$$p_j = \frac{dSLD_j}{dSLD_j - dSUD_j}$$

which is simply the empirical percentile from before, but computed on the subset of observations with LS above the point where the regression line becomes censored. This is consistent with the result from Powell that cLAD regression is equivalent to LAD regression performed on the subset of data for which the probability of censoring is uniformly no greater than the regression percentile π and at some covariates the probability of censoring is strictly less than π.

It was previously noted that the censored data do not guarantee the convexity conditions on the set of upper and lower deviations as is the case in traditional LAD regression. This can occur stochastically in the lower quantiles when there are few data points from which to estimate the percentiles. This can also occur systematically when the observed expression is correlated with *LS* as may occur when sub-populations of cells express in aggregate at different levels.

To correct for both of these issues simultaneously, a monotonicity condition is imposed on the estimated p_j. First, the p_j are computed only on the subset of upper/lower deviations that exist on the edge of the convex hull of $\{SUD_j, SLD_j\}$. Following computation of percentiles on this subset of quantiles, percentiles are adjusted such that differences between adjacent percentiles are bounded above and below. The bounds are as follows:

$$p_j - p_{j+1} \geq \frac{\sum_k z_k > \left(\beta_{0j} + d_k\right)^* - \sum_k z_k > \left(\beta_{0j+1} + d_k\right)^*}{n}$$
$$p_j - p_{j+1} \leq p_j - \frac{\sum_k z_k < \left(\beta_{0j+1} + d_k\right)^*}{n}$$

A.2 Datasets

This section is edited for clarity from supplement S2 from Brown *et al.* 2021 (Brown, Ni, *et al.*, 2021).

A.2.1 PBMC_Pure

PBMC68K_Pure is a partner dataset to PBMC68K (Zheng et al., 2017) produced by purifying peripheral blood mononuclear cells (PBMCs) into 10 cell types through the use of cell-type specific isolation kits and separately sequencing each group. One group was then computationally separated into two resulting in 11 annotated cell-types. These cell-type annotations are considered here as ground truth when evaluating the effects of normalization on downstream clustering. For increased accuracy, the six cell-types for which tSNE plots do not separate into sub-groups (van der Maaten and Hinton, 2008; Van Der Maaten, 2014) were subset: CD4+ T Helper2, CD4+/CD25 T Reg, CD4+/CD45RA+/CD25- Naive T, CD4+/CD45RO+ Memory, CD56+ NK, and CD8+/CD45RA+ Naive Cytotoxic. Zheng *et al.* identify these particular cell-types as demonstrating little sub-structure (Zheng *et al.*, 2017).

UMI count matrices and barcode (cell) metadata are available from the GitHub repository associated with the publication: https://github.com/10XGenomics/single-cell-3prime-paper/tree/master/pbmc68k_analysis.

A.2.2 PBMC5K_Prot

PBMC5K_Prot is a dataset of approximately 5 thousand PBMCs sequenced by and available from 10X genomics under the name "5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry)" and processed under cell ranger version 3.1.0. A panel of 31 surface proteins were sequenced in parallel with the cDNA libraries. We perform unsupervised clustering on the protein abundance estimates to generate pseudo-annotations independently from RNA expression measurements.

A.2.3 MaltTumor10K

MaltTumor10K is a dataset of approximately 10 thousand cells from a MALT tumor sequenced by and available from 10X genomics under the name "10k Cells from a MALT Tumor - Gene Expression and Cell Surface Protein" and processed under cell ranger version 3.0.0. A panel of 17 surface proteins were sequenced in parallel with the cDNA libraries. We perform unsupervised clustering on the protein abundance estimates to generate pseudo-annotations independently from RNA expression measurements.

A.2.4 MouseBrain

MouseBrain is a dataset of approximately 9 thousand mouse brain cells sequenced by and available from 10X genomics under the name "9k Brain Cells from an E18 Mouse" and processed under cell ranger version 1.3.0.

A.2.5 PBMC68K

PBMC68K is a partner dataset to PBMC68K_Pure (Zheng *et al.*, 2017) produced by sequencing approximately 68 thousand PBMCs. In the original paper, pseudo-annotations were generated by computational matching of these cells to the purified lines of PBMC68K_Pure. We, however, treat these as unannotated cells. UMI count matrices are available from 10X genomics under the name "Fresh 68k PBMCs (Donor A)" and processed under cell ranger version 1.1.0.

A.2.6 EMT

EMT is a dataset of 5,004 sequenced MCF10A mammary epithelial cells induced to undergo spontaneous epithelial to mesenchymal transitions (EMTs) through the cellular detection of neighboring unoccupied space (McFaline-Figueroa *et al.*, 2019). This spatial effect allowed the authors to dissect an inner region a-priori expected to be primarily epithelial cells and an outer region a-priori expected to be primarily mesenchymal cells which were then sequenced separately. The authors produced another dataset of cells activated by TGF-β (denoted TGFB in the barcode metadata), but we consider only the first dataset (denoted Mock in the barcode metadata). Included in the initial publication, the authors describe eight gene sets from the Hallmark collection (Liberzon *et al.*, 2015) which they consider to be significantly enriched for activity during EMT. We take this set of terms as a ground truth for assessing power under a range of normalization techniques: ESTROGEN RESPONSE LATE, ESTROGEN RESPONSE EARLY, P53 PATHWAY, KRAS SIGNALING DN, MYC TARGETS V1, MYC TARGETS V2, PI3K AKT MTOR SIGNALING, and EPITHELIAL MESENCHYMAL TRANSITION. The UMI count matrices and barcode metadata are available on GEO under accession number GSE114687.

A.2.7 Dataset processing

For the un-annotated datasets published by 10X (PBMC5K_Prot, MaltTumor10K, MouseBrain, PBMC68K), the UMI count matrices analyzed in this paper were derived from un-filtered gene-barcode matrices. *emptyDrops* (R package: *DropletUtils*; parameters: lower = 20, niters = 160000, test.ambient = TRUE) was used to differentiate empty droplets from barcodes associated with cells (Lun *et al.*, 2019). Cellular barcodes were then defined as those with FDR corrected p-values less than 1e-3. For the datasets with surface protein expression (PBMC5K_Prot and MaltTumor10K), cells were additionally filtered to retain only those with a minimum of 100 protein-specific UMIs. This procedure resulted in datasets with 4978 cells (PBMC5K_Prot), 8670 cells (MaltTumor10K), 3756 cells (MouseBrain), and 77249 cells (PBMC68K). Where applicable, surface protein expression was omitted from the rows of UMI count matrices for downstream analysis and testing.

Pseudo-annotations were generated from the datasets with surface protein expression (PBMC5K_Prot and MaltTumor10K) in a manner similar to the unsupervised clustering of all datasets. Surface protein expression was normalized by the method of median ratio (Anders and Huber, 2010). Note: as all surface proteins are generally expected to have cell-type-specific abundances, this normalization is only expected to equalize counts within cell-types meaning that between cell-type calculations of relative abundance are rendered inapplicable; as our purpose is clustering, this is not a problem. Using the Seurat pipeline, protein expression is reduced to 20 dimensions for PBMC5K_Prot (from 29 distinct proteins) and 15 dimensions for MaltTumor10K (from 17 distinct proteins). Graph based clustering is then performed using *FindNeighbors* and *FindClusters* (additional parameters: algorithm = 3, n.start = 100, n.iter = 100) from the Seurat package.

A.3 Data simulation

This section is edited for clarity from supplement S3 from Brown *et al.* 2021 (Brown, Ni, *et al.*, 2021).

A.3.1 Initial grouping

We generate our simulated datasets from experimentally derived UMIs with the purpose of making our simulations as representative of the characteristics of individual, experimentally derived cells as possible. The first step is to normalize experimental data using Dino and then perform unsupervised clustering on the normalized data. This results in (relatively) homogenous subsets of cells. Using these cluster annotations, we then generate individual clusters of simulated data from the raw (unnormalized) UMI counts from each of these cluster annotations. These simulated clusters, mirroring the cell-type heterogeneity in the original data, are then merged into a test dataset.

A.3.2 Group filtering

We vary the size of simulated clusters by powers of 2 (40 cells, 80 cells, 160 cells, etc.). To this end, we calculate the largest k such that we have k calculated clusters with at least 40×2^k cells and discard the remaining experimental data. If $k=6$, then we have exactly 6 clusters of experimental data with at least $40 \times 2^6 = 2560$ cells in each group, and we discard cells from any smaller clusters.

A.3.3 Cluster pair simulation

The simulated dataset is constructed of pairs of simulated clusters for which the EE and DE genes are known by design. In the case were $k=6$ as above, the simulated datasets then consist of 6 cluster pairs, or 12 simulated clusters total. Within a cluster pair, there is an induced difference in *LS*

between the pairs and DE genes are randomly selected. Between cluster pairs, there may also be systematic differences in *LS*, but only to the extent that there are systematic differences in the *LS*s of the experimental cells these cluster pairs are based on.

To construct one cluster pair, an experimental cluster, denoted by C_k, is randomly sampled. Each of the two simulated clusters in the pair will consist of 40 cells if this is the first cluster pair, 80 cells if this is the second cluster pair, and so on increasing by factors of 2. Denote the number of simulated cells in each cluster by *n*. To generate the first *n/2* cells in each cluster, we sample *n* cells from the experimental data. In order of increasing *LS*, we sum pairs of experimental cells to create *n/2* pseudo-cells with roughly double the *LS* of either of the cells they are comprised of.

Denote the simulated clusters in the pair by *A* and *B*, each to consist of *n* simulated cells. To induce a difference in *LS* between *A* and *B*, we sample a *LS* fold change, δ_{fc}, from the range *3/2* to *4*, and for convenience assign *A* to be the group with higher average *LS*. The first *n/2* cells in each group will be generated by binomial sampling from the *n/2* pseudo-cells, with the induced fold change in *LS* arising from differences in the binomial probability parameter, *p*. Some algebra shows that the choice of

$$p = 0.5 \pm \frac{(\delta_{fc} - 1)}{2(\delta_{fc} + 1)}$$

for *A* and *B* respectively will produce clusters with all EE genes once *LS* is accounted for under normalization. Specifically, for pseudo-cell s_j, simulated cells a_j and b_j from *A* and *B* respectively are generated as:

$$a_j \sim Binom(s_j, p_+)$$
$$b_j \sim Binom(s_j, p_-)$$

where p_+ and p_- are the two variants of *p* respectively.

However, this approach only generates EE genes. To simulate known DE genes, we subset those genes in C_k with at least 25% non-zeros. From this set of genes, we sample 10 to be induced DE genes, with sampling weighted by the inverse density of log gene expression, calculated simply as the log of the mean UMIs in C_k. As with the fold change in *LS*, we sample 10 DE fold changes from the range *3/2* to *6*, denoted by $\gamma_{fc,g}$ with the subscript *g* indexing the 10 gene-specific DE fold changes. As we do not want all DE genes to be upregulated in *A*, we invert each of the $\gamma_{fc,g}$ with probability 0.5. If we now consider the binomial probability, p, to be a vector of length equal to the number of genes, and p_{DE} to denote the subset of elements which are DE after correcting for *LS*, some similar algebra to the above shows that

$$p_{DE,g} = 0.5 \pm \frac{(\delta_{fc}\gamma_{fc,g} - 1)}{2(\delta_{fc}\gamma_{fc,g} + 1)}$$

where this formulation also allows the definition of a_j and b_j to be defined as the same binomial random variable parameterized by p_+ or p_-, where p now includes information about DE sampling.

Two problems remain to be addressed; that we have only discussed the generation of *n/2* of the cells in each group and that correcting for *LS* as defined here can induce slight but systematic differential expression in the EE genes. Take, for example, the extreme case where all the DE genes are upregulated in *A* relative to *B* after correcting for *LS*. In this case, calculating *LS* from the sum of simulated UMIs within a cell, and correcting for that *LS*, will induce a slight but consistent down-regulation in the EE genes in *A* relative to *B*. We address this by adding a correction factor to the remaining *n/2* simulated cells in each cluster.

The degree of this induced bias can be simply calculated as the ratio of expected total *LS* (total meaning summed across cells as well as genes) under the above DE model and the model where all genes are simulated EE. Let p_+ be, as above, a vector of binomial probabilities which

includes DE information for the simulation of A and let p_{EE+} be a corresponding binomial probability vector for which all genes are EE, that is, suppose all elements of p_{EE+} are equal to the original, scalar, definition of p_{+}. Let C_{k-} denote a vector of gene-wise UMIs, summed across all cells in C_k. Then, the total expected LS for the EE case is $p_{EE+}{}^{T}C_{k-}$ and the degree of bias in the above simulated cells, inducing DE in simulated EE genes, is

$$\alpha_{+bias} = \frac{p_{+}^{T}C_{k-}}{p_{EE+}^{T}C_{k-}}$$

This can be interpreted as implying that normalized EE genes in A will, on average, be a factor of $1/\alpha_{+bias}$ different from what would have been the case had all genes had been simulated as EE. If $\alpha_{-bias}=\alpha_{+bias}$, then this wouldn't be a problem, but such is not the case. Unfortunately, it is also the case that $\alpha_{-bias} \neq 1/\alpha_{+bias}$, as can be shown by simple counter examples. Therefore, we compute separate corrective factors for A and B under the principle that expression of EE genes between A and B should, when averaged across the first $n/2$ cells and the second, corrected $n/2$ cells demonstrate the desired fold change in LS. This leads to the corrective factor, c

$$\begin{aligned}\left(\frac{c_{+}p_{+}^{T}C_{k-}}{p_{EE}^{T}C_{k-}}\right)^{-1} &= 1-\left(\frac{1}{\alpha_{+bias}}-1\right)\\ \Rightarrow c_{+} &= \frac{1}{2\alpha_{+bias}-1}\\ \Rightarrow c_{-} &= \frac{1}{2\alpha_{-bias}-1}\end{aligned}$$

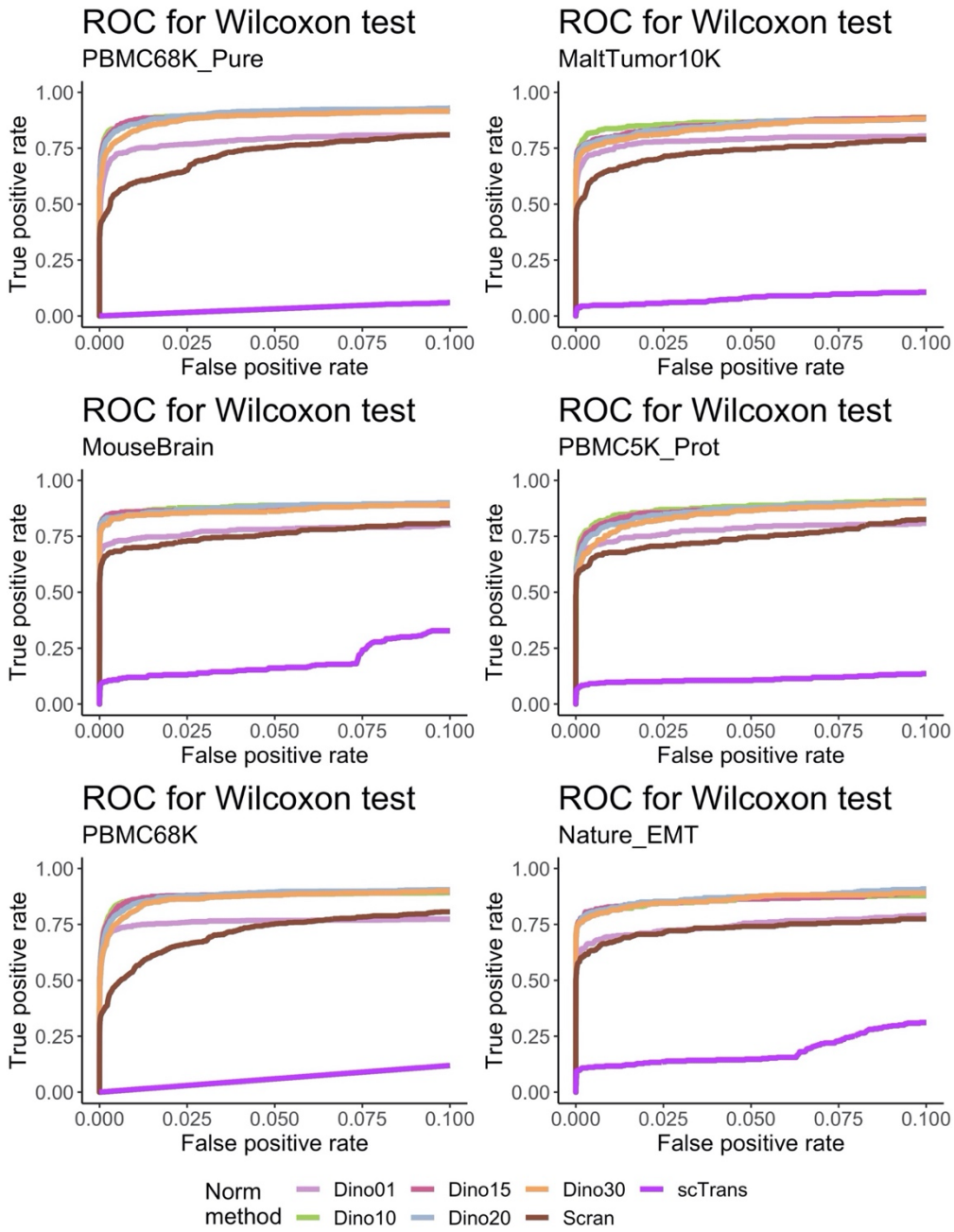
This then fully defines the simulated cells:

$$a_j \sim \begin{cases} Binom(s_j, p_{+}), j \leq n/2 \\ Binom\left(s_{j-\frac{n}{2}}, c_{+}p_{+}\right), j > n/2 \end{cases}$$

$$b_j \sim \begin{cases} Binom(s_j, p_{-}), j \leq n/2 \\ Binom\left(s_{j-\frac{n}{2}}, c_{-}p_{-}\right), j > n/2 \end{cases}$$

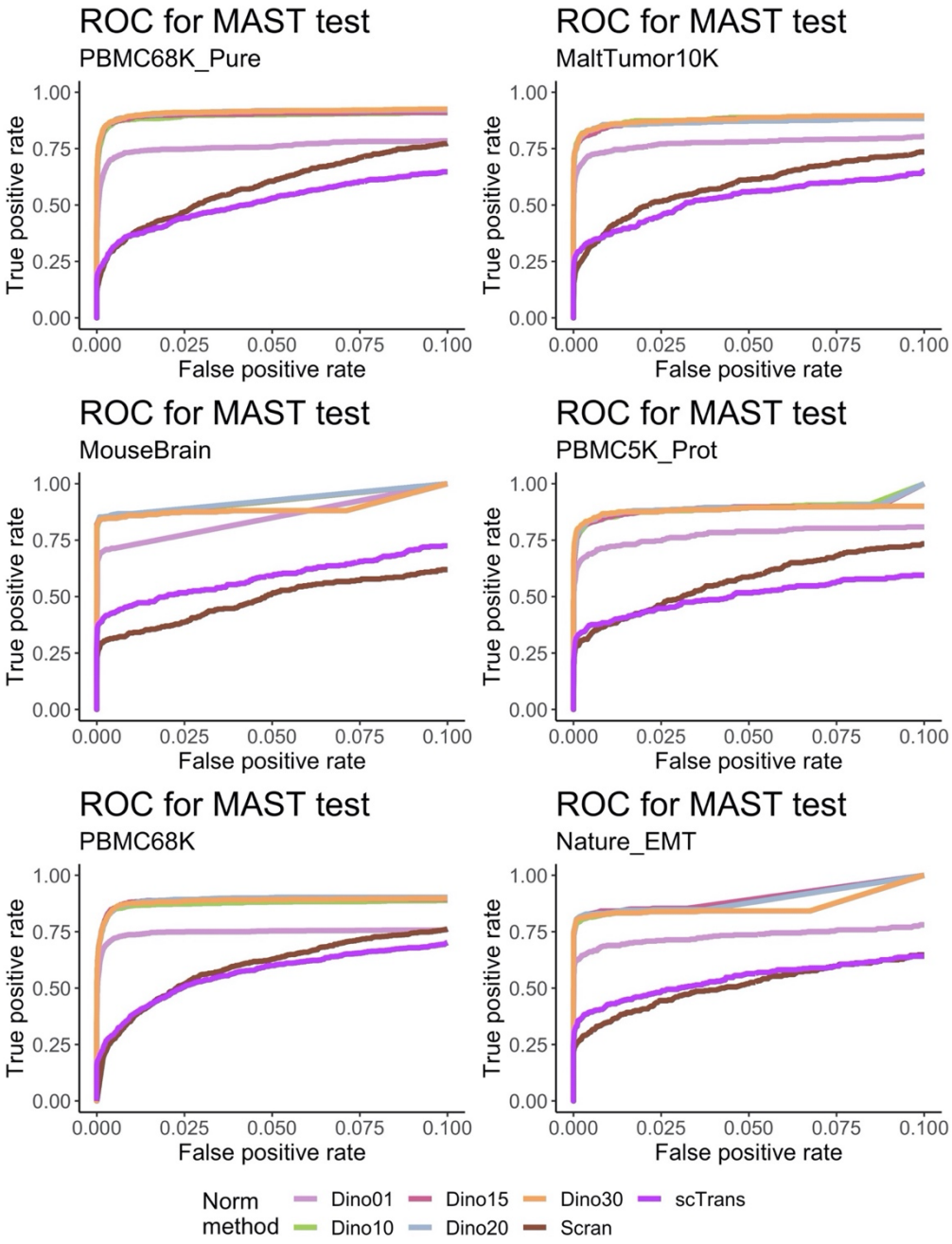
To complete a simulated dataset, the above steps for generating the cluster pair *A* and *B* are repeated for the remaining experimental clusters, generating a heterogenous samples of simulated data for which pairs of simulated clusters have known EE and DE genes.

A.4 Supplemental Figures



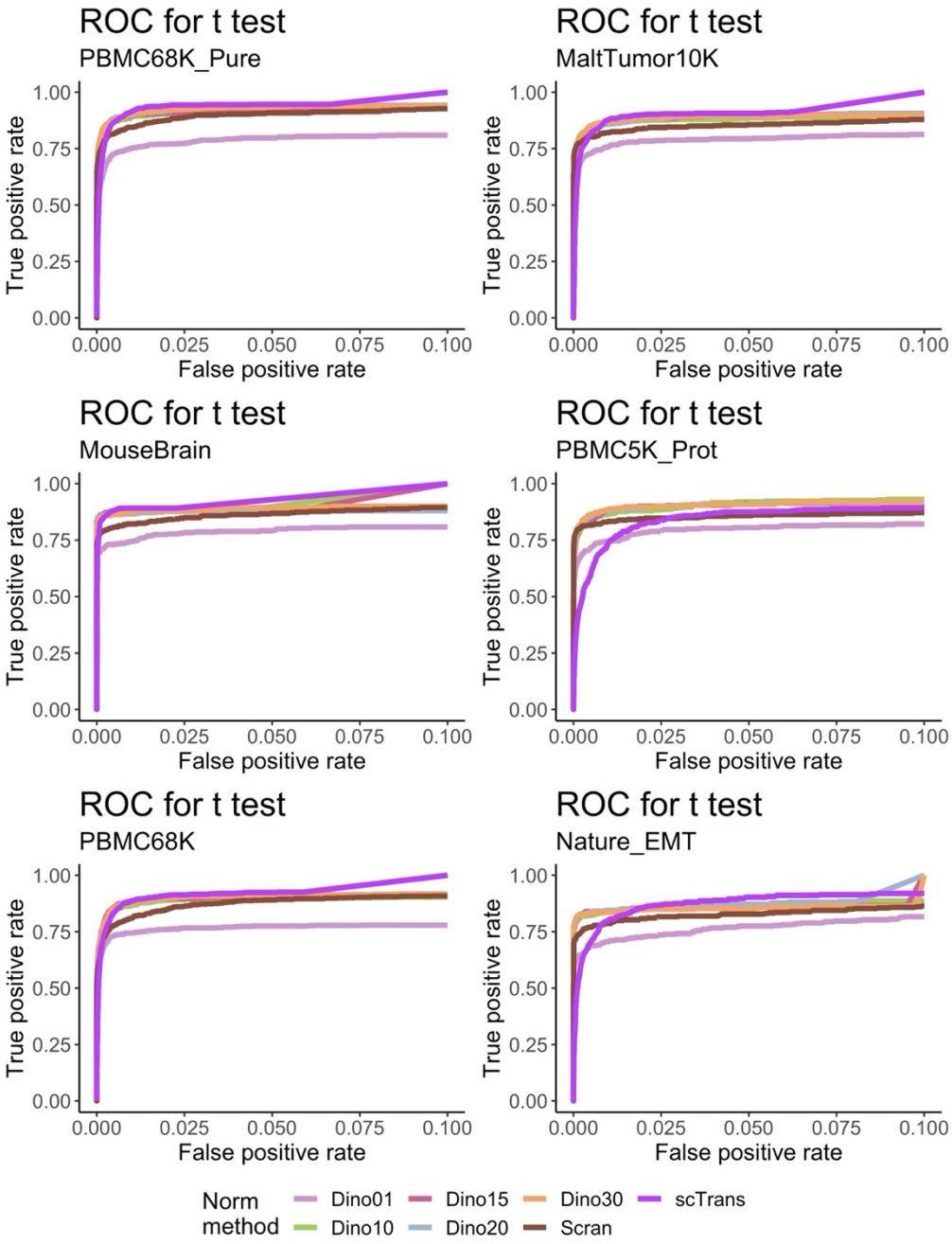
Supplemental Figure A.1: **Dino concentration parameter variation: Wilcoxon test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a Wilcoxon rank sum test, where the average is calculated across 12 simulations from each dataset. The concentration parameter (default 15) used for each run of Dino is indicated in the numeric suffix.



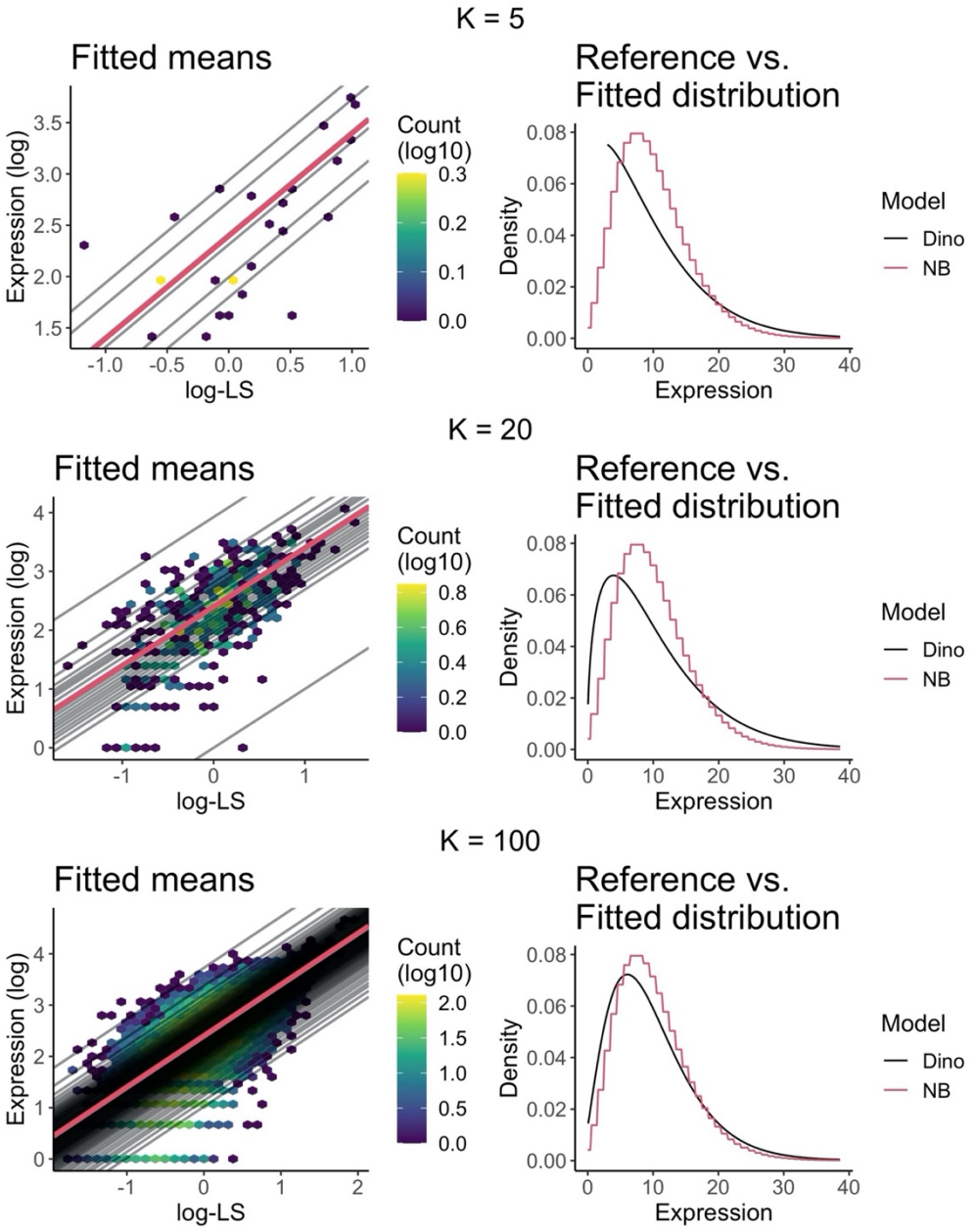
Supplemental Figure A.2: **Dino concentration parameter variation: MAST test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a MAST test, where the average is calculated across 12 simulations from each dataset. The concentration parameter (default 15) used for each run of Dino is indicated in the numeric suffix.



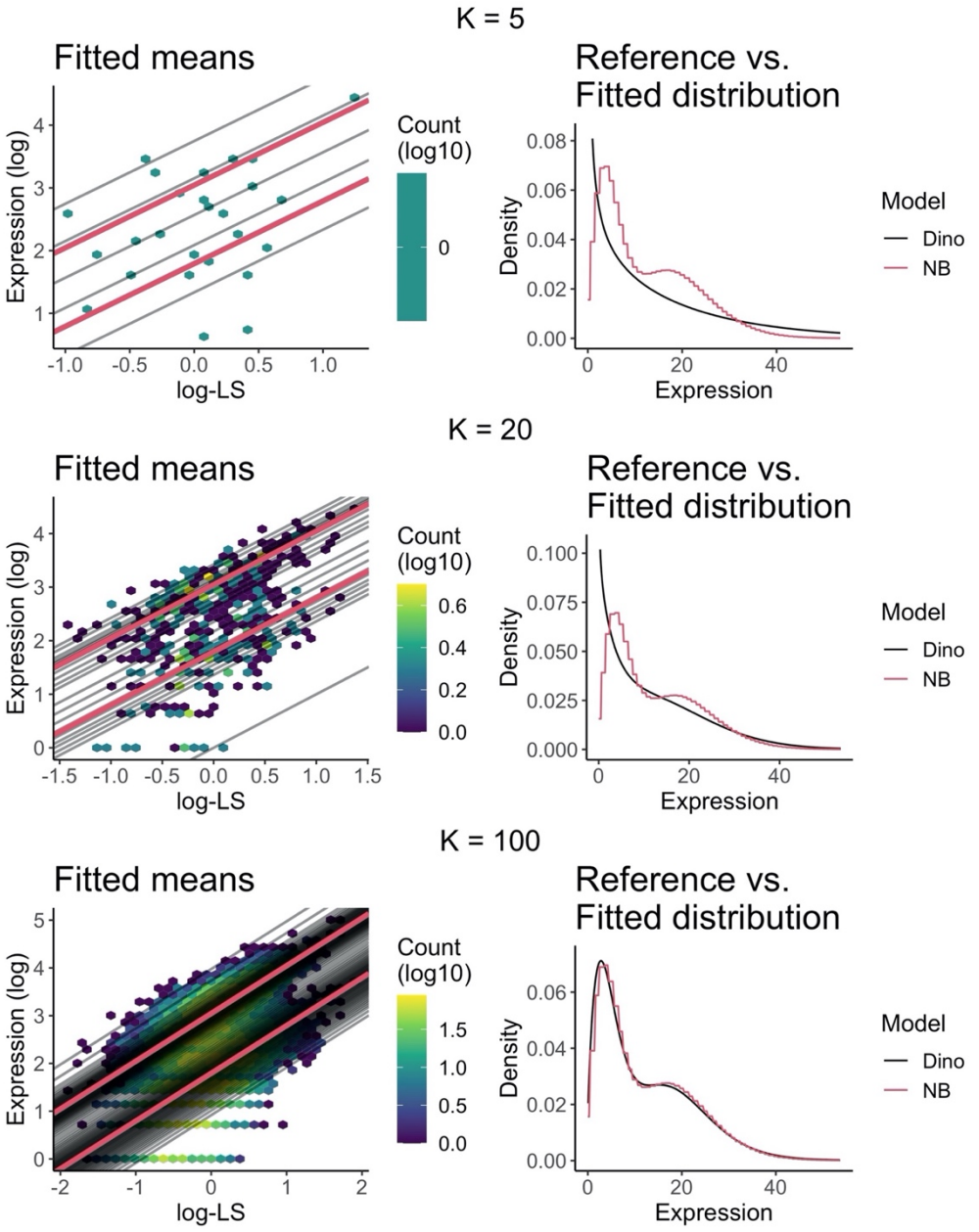
Supplemental Figure A.3: **Dino concentration parameter variation: t-test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a t-test, where the average is calculated across 12 simulations from each dataset. The concentration parameter (default 15) used for each run of Dino is indicated in the numeric suffix.



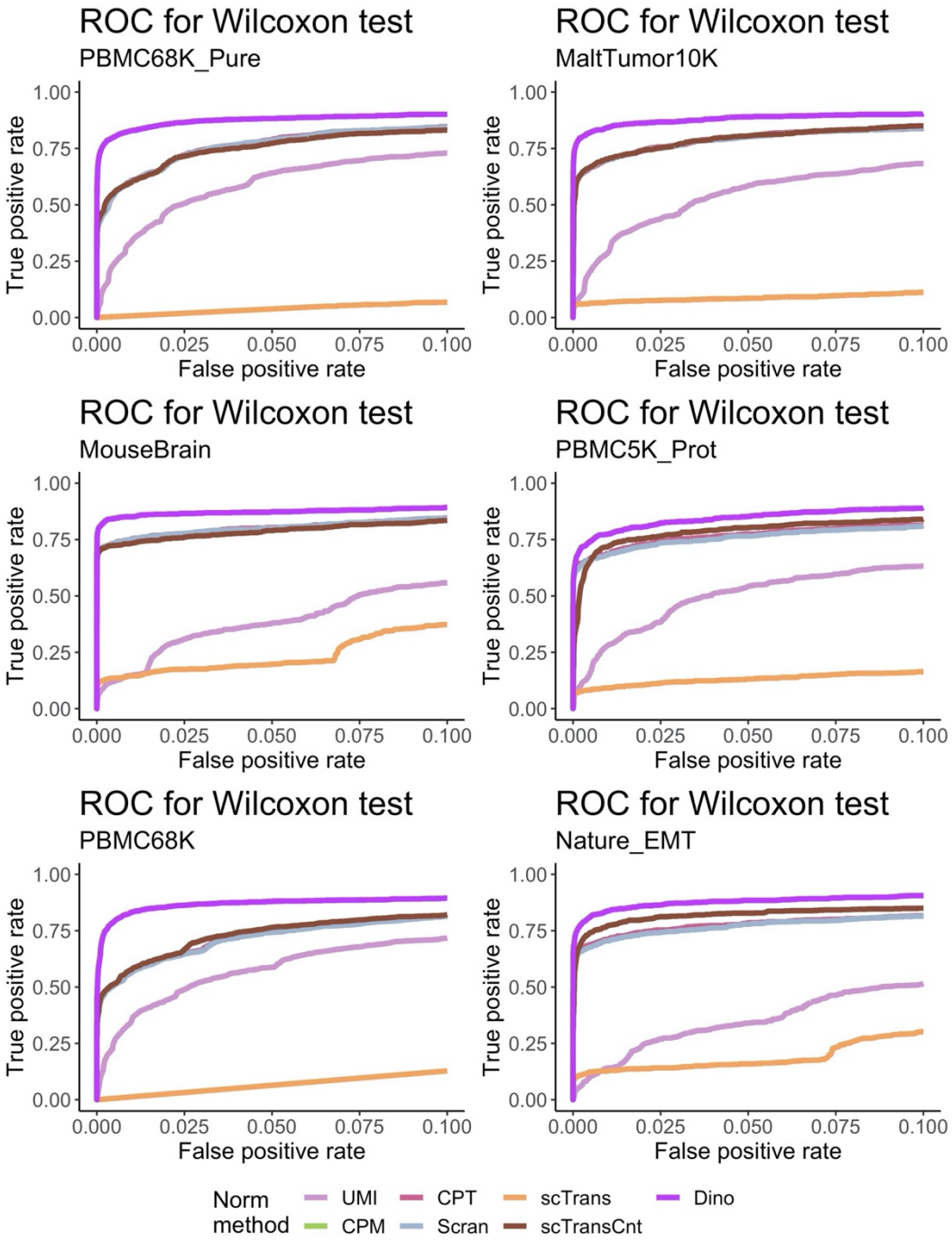
Supplemental Figure A.4: **Larger values of *K* improve model fit for unimodal distributions.**

We simulated UMI count data under a unimodal Negative Binomial distribution. Each row of the figure indicates a new simulation with sequentially larger values of *K* for the Dino fit. The background of the left column plots the simulated UMI counts as a heatmap (log expression against log library size). Overlaying this heatmap are the mean trend lines fitted from the Dino mixture model (gray) and the true mean of the sampling distribution (red). The right column plots the true distribution of the simulated data at LS=0 (red) and the Dino estimate of the prior distribution which is subsequently used to generate normalized expression values. In practice *K* is chosen algorithmically for each gene, and so small values of *K* relate to correspondingly fewer data points, hence the sparsity of the heatmap of simulated data for *K=5*.



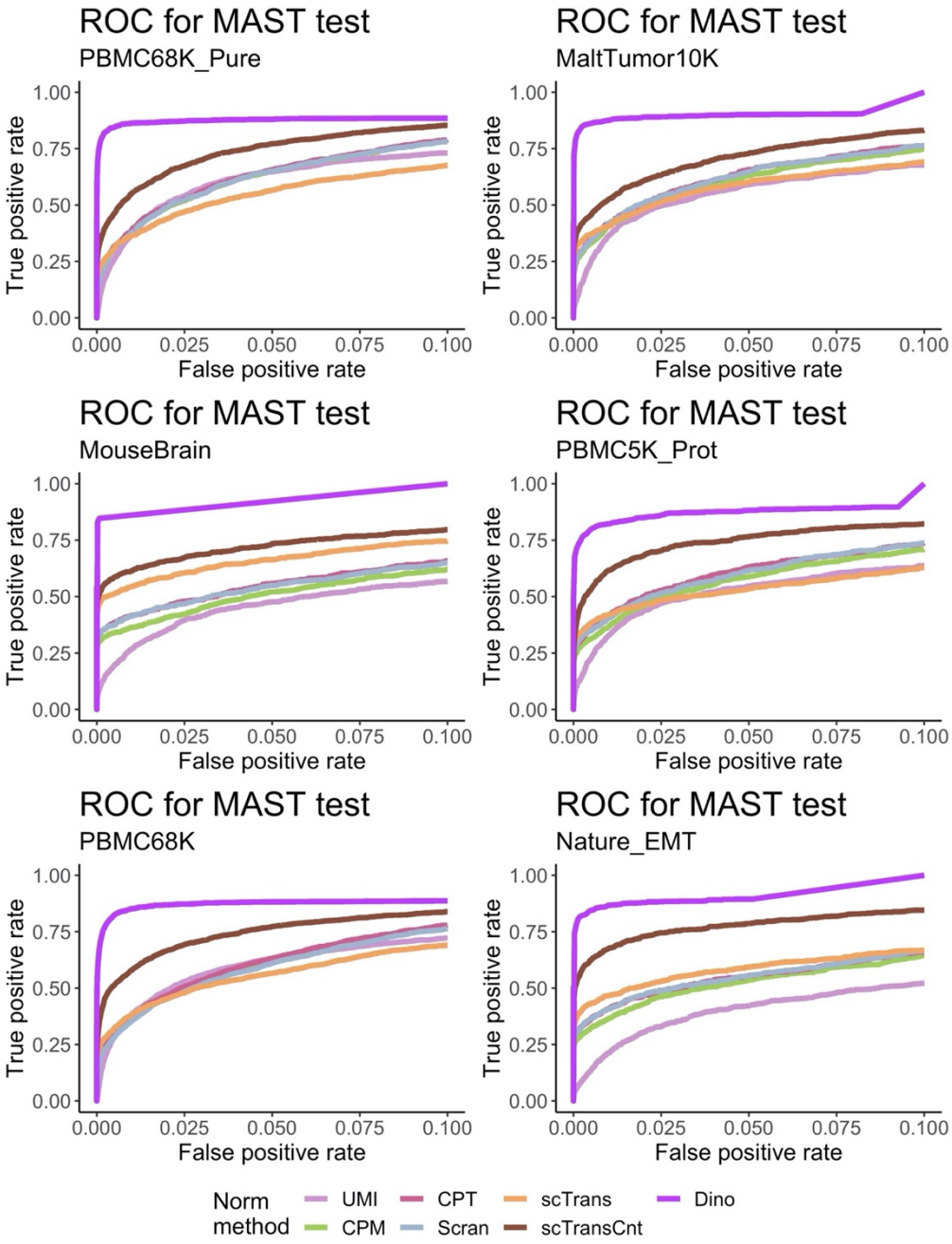
Supplemental Figure A.5: **Larger values of *K* improve model fit for multimodal distributions.**

We simulated equal numbers of UMI counts from a mixture of two NB distributions, each with different means and dispersion parameters. Each row of the figure indicates a new simulation with sequentially larger values of *K* for the Dino fit. The background of the left column plots the simulated UMI counts as a heatmap (log expression against log library size). Overlaying this heatmap are the mean trend lines fitted from the Dino mixture model (gray) and the true means of the sampling distribution (red). The right column plots the true distribution of the simulated data at LS=0 (red) and the Dino estimate of the prior distribution which is subsequently used to generate normalized expression values. In practice *K* is chosen algorithmically for each gene, and so small values of *K* relate to correspondingly fewer data points, hence the sparsity of the heatmap of simulated data for *K=5*.



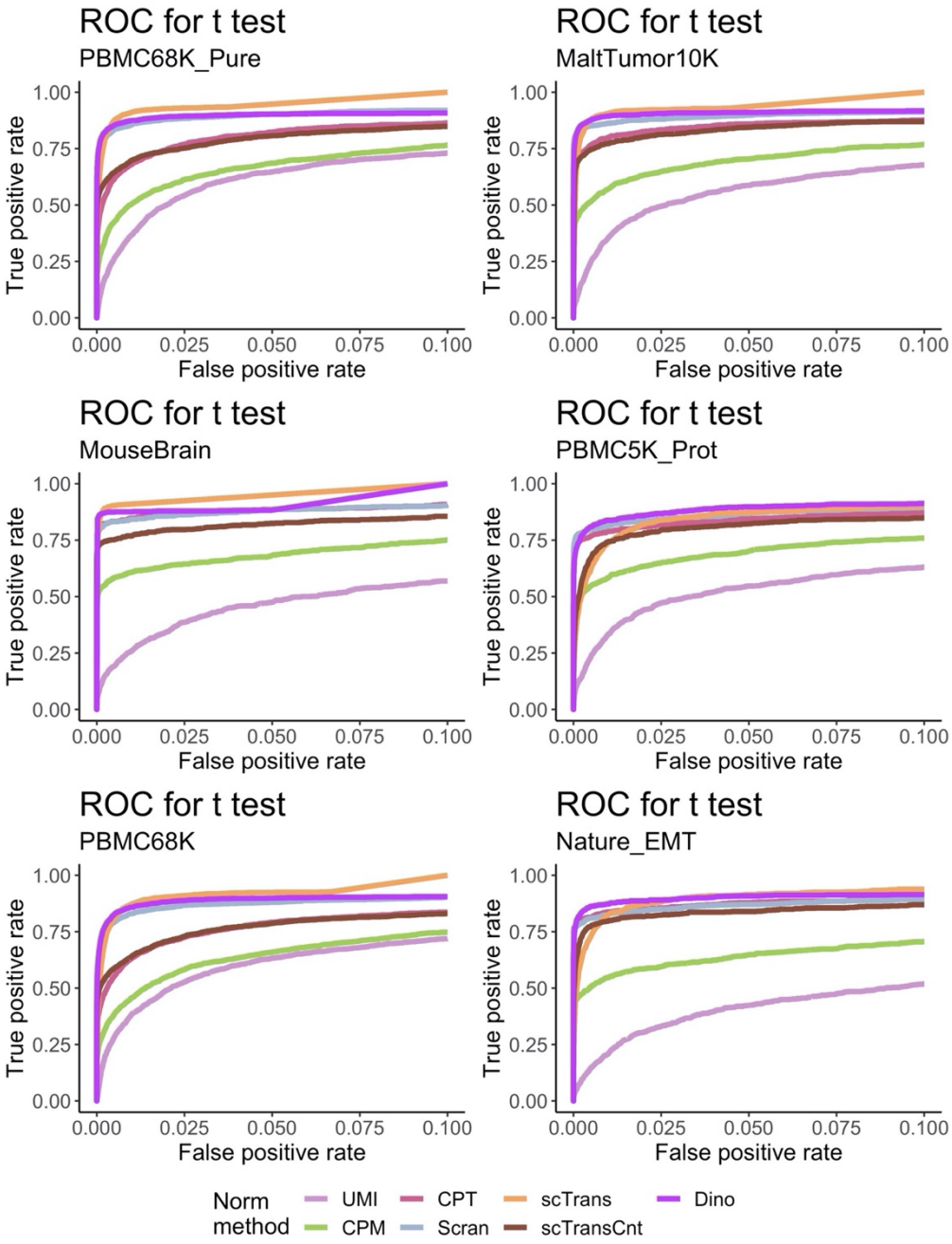
Supplemental Figure A.6: **Normalized DE testing comparison: Wilcoxon test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a Wilcoxon rank sum test, where the average is calculated across 30 simulations from each dataset.



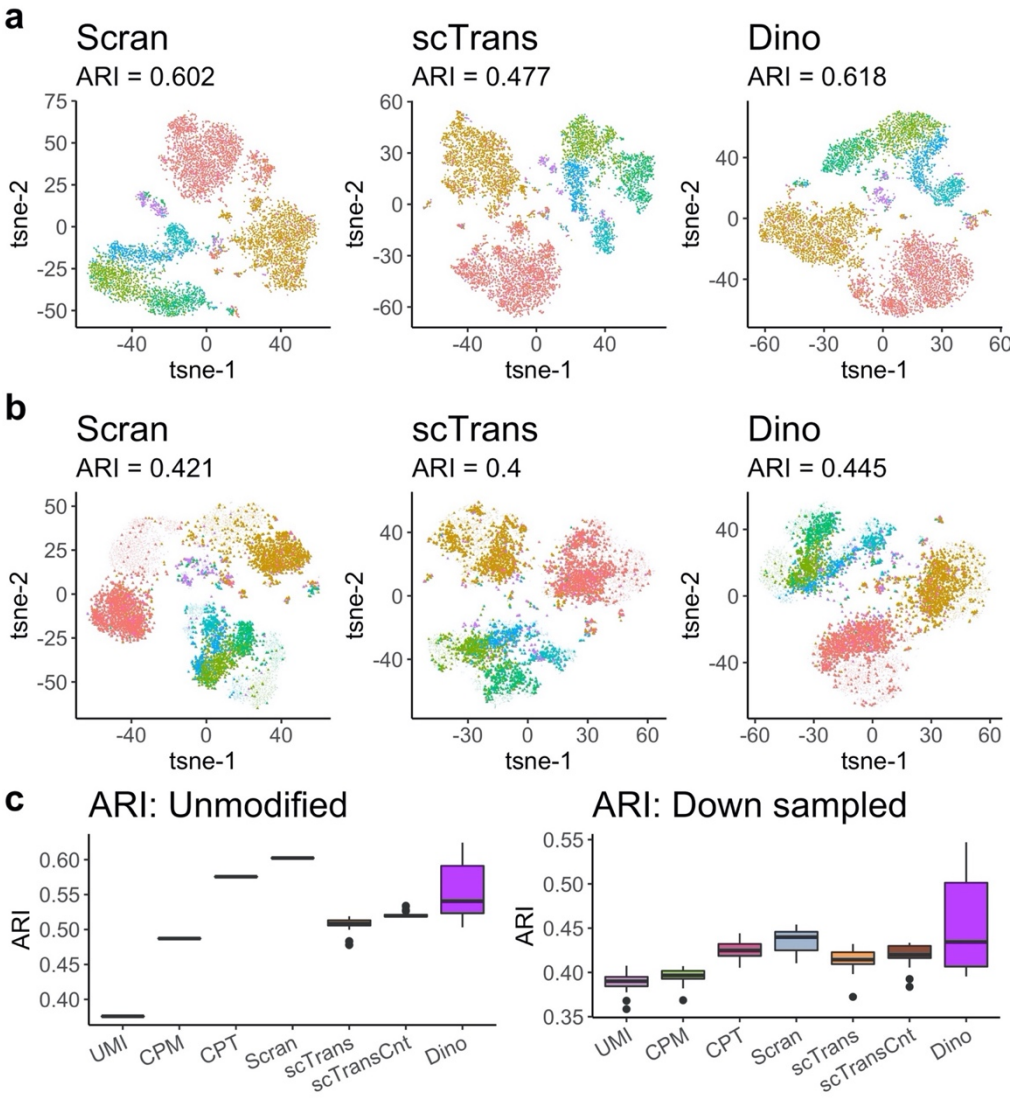
Supplemental Figure A.7: **Normalized DE testing comparison: MAST test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a MAST test, where the average is calculated across 30 simulations from each dataset.



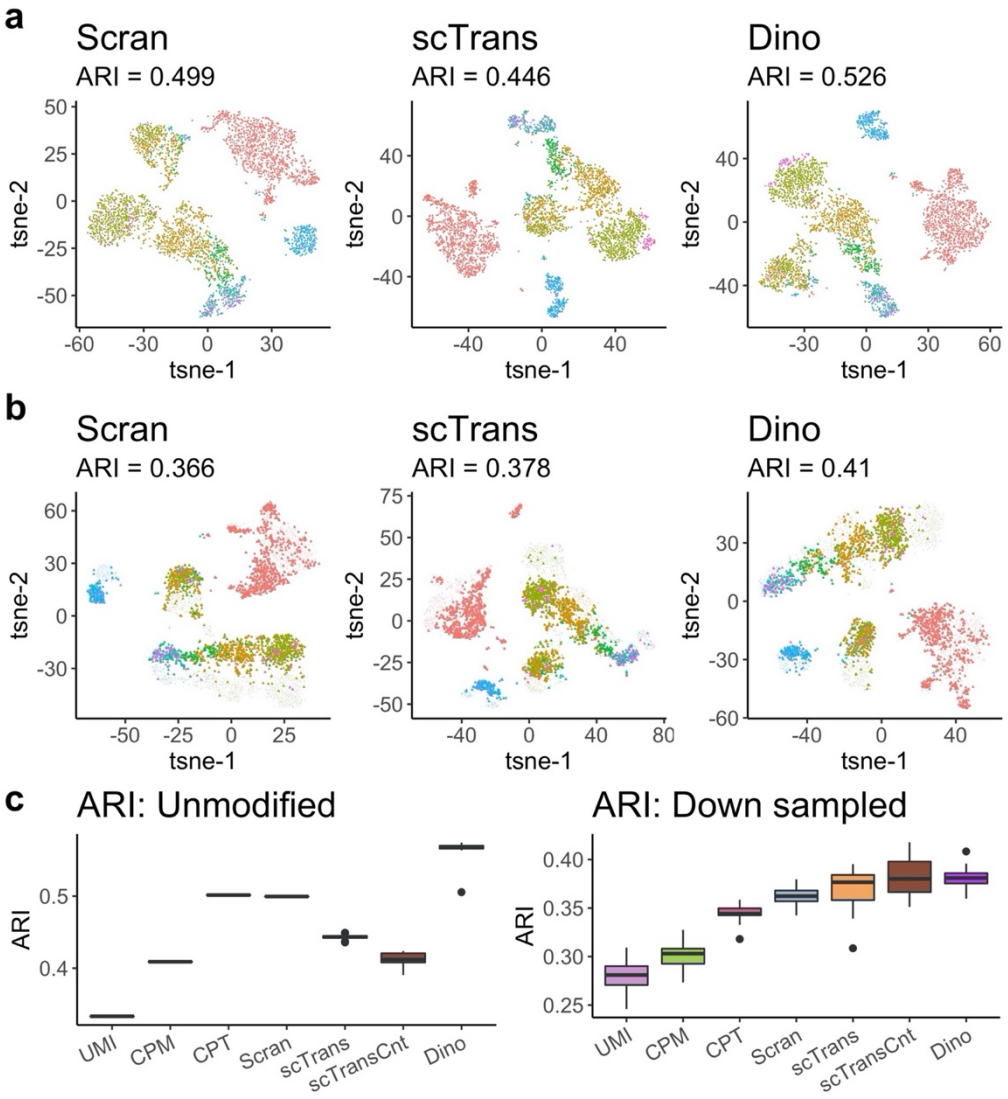
Supplemental Figure A.8: **Normalized DE testing comparison: t-test.**

Simulated data based on each of the considered datasets were normalized using each method. ROC curves colored by normalization method define the relationship between average TPR (Power) and average FPR for a t-test, where the average is calculated across 30 simulations from each dataset.



Supplemental Figure A.9: **Normalized clustering comparison: MaltTumor10K.**

a) tSNE plots of normalized MaltTumor10K data, colored by 11 pseudo-annotations, show similarly high accuracy across methods. b) The same clustering plots as in (a), but with half the data down-sampled prior to normalization to produce greater differences in *LS*. c) Boxplots of ARIs for multiple un-modified and down-sampled datasets across 24 replications of the normalization procedures and, for the down-sampled data, 24 applications of the down-sampling.



Supplemental Figure A.10: **Normalized clustering comparison: PBMC5K_Prot.**

a) tSNE plots of normalized PBMC5K_Prot data, colored by 11 pseudo-annotations, show similarly high accuracy across methods. b) The same clustering plots as in (a), but with half the data down-sampled prior to normalization to produce greater differences in *LS*. c) Boxplots of ARIs for multiple un-modified and down-sampled datasets across 24 replications of the normalization procedures and, for the down-sampled data, 24 applications of the down-sampling.

Norm. Method	*UMI*	*CPM*	*CPT*	*Scran*	*scTrans*	*scTransCnt*	*Dino*	
PBMC68K_Pure	0.797 (0.008)	0.871 (0.006)	0.871 (0.006)	0.868 (0.006)	0.894 (0.007)	0.859 (0.008)	0.818 (0.010)	**Power**
	0.196 (0.009)	0.153 (0.010)	0.153 (0.010)	0.152 (0.010)	0.622 (0.003)	0.152 (0.008)	0.007 (0.001)	**FPR**
PBMC5K_Prot	0.738 (0.014)	0.796 (0.013)	0.796 (0.013)	0.789 (0.013)	0.838 (0.014)	0.764 (0.014)	0.729 (0.017)	**Power**
	0.244 (0.016)	0.068 (0.014)	0.068 (0.014)	0.068 (0.014)	0.630 (0.007)	0.025 (0.005)	0.004 (0.001)	**FPR**
MaltTumor10K	0.774 (0.011)	0.853 (0.009)	0.853 (0.009)	0.854 (0.009)	0.923 (0.008)	0.854 (0.009)	0.808 (0.012)	**Power**
	0.216 (0.013)	0.121 (0.013)	0.121 (0.013)	0.120 (0.013)	0.704 (0.003)	0.107 (0.010)	0.003 (0.001)	**FPR**
MouseBrain	0.788 (0.014)	0.818 (0.011)	0.818 (0.011)	0.820 (0.011)	0.896 (0.009)	0.831 (0.009)	0.751 (0.012)	**Power**
	0.440 (0.021)	0.069 (0.014)	0.069 (0.014)	0.068 (0.014)	0.516 (0.003)	0.095 (0.013)	0.000 (0.000)	**FPR**
PBMC68K	0.799 (0.009)	0.859 (0.006)	0.859 (0.006)	0.852 (0.007)	0.896 (0.007)	0.856 (0.007)	0.823 (0.007)	**Power**
	0.202 (0.007)	0.174 (0.008)	0.174 (0.008)	0.173 (0.008)	0.595 (0.003)	0.139 (0.012)	0.009 (0.001)	**FPR**
EMT	0.736 (0.013)	0.812 (0.014)	0.812 (0.014)	0.812 (0.014)	0.883 (0.014)	0.824 (0.011)	0.753 (0.015)	**Power**
	0.384 (0.017)	0.092 (0.012)	0.092 (0.012)	0.094 (0.012)	0.562 (0.007)	0.040 (0.007)	0.001 (0.000)	**FPR**

Supplemental Table A.1: **Average power and FPR statistics: Wilcoxon test.**

30 simulated datasets are produced from each case study dataset. In each simulation, the data are normalized by the panel of methods and significantly DE genes are identified using the Wilcoxon rank sum test. DE genes are defined as those with a Benjamini and Hochberg adjusted p-value less than 0.01. Average power and FPR is calculated over the 30 simulated datasets (standard error computed across simulations).

Norm. Method	*UMI*	*CPM*	*CPT*	*Scran*	*scTrans*	*scTransCnt*	*Dino*	
PBMC68K_Pure	0.788 (0.008)	0.956 (0.006)	0.956 (0.006)	0.963 (0.004)	0.917 (0.006)	0.906 (0.007)	0.819 (0.010)	**Power**
	0.172 (0.008)	0.274 (0.006)	0.270 (0.005)	0.299 (0.005)	0.427 (0.012)	0.179 (0.011)	0.002 (0.000)	**FPR**
PBMC5K_Prot	0.720 (0.014)	0.943 (0.009)	0.950 (0.009)	0.966 (0.007)	0.843 (0.011)	0.799 (0.012)	0.730 (0.016)	**Power**
	0.213 (0.015)	0.346 (0.018)	0.345 (0.016)	0.372 (0.016)	0.386 (0.021)	0.071 (0.009)	0.001 (0.000)	**FPR**
MaltTumor10K	0.758 (0.010)	0.953 (0.008)	0.953 (0.007)	0.962 (0.007)	0.897 (0.008)	0.905 (0.009)	0.804 (0.011)	**Power**
	0.188 (0.012)	0.213 (0.012)	0.305 (0.010)	0.327 (0.010)	0.372 (0.020)	0.182 (0.016)	0.001 (0.000)	**FPR**
MouseBrain	0.769 (0.015)	0.962 (0.007)	0.972 (0.007)	0.969 (0.007)	0.881 (0.010)	0.880 (0.010)	0.754 (0.012)	**Power**
	0.400 (0.020)	0.521 (0.022)	0.571 (0.021)	0.561 (0.021)	0.255 (0.021)	0.208 (0.022)	0.000 (0.000)	**FPR**
PBMC68K	0.789 (0.008)	0.946 (0.007)	0.946 (0.006)	0.963 (0.005)	0.914 (0.008)	0.870 (0.009)	0.824 (0.007)	**Power**
	0.180 (0.006)	0.268 (0.005)	0.267 (0.004)	0.304 (0.004)	0.392 (0.007)	0.144 (0.012)	0.005 (0.000)	**FPR**
EMT	0.715 (0.013)	0.951 (0.010)	0.961 (0.009)	0.955 (0.009)	0.889 (0.012)	0.853 (0.011)	0.744 (0.015)	**Power**
	0.349 (0.017)	0.499 (0.012)	0.508 (0.012)	0.493 (0.012)	0.349 (0.007)	0.111 (0.007)	0.000 (0.000)	**FPR**

Supplemental Table A.2: **Average power and FPR statistics: MAST test.**

30 simulated datasets are produced from each case study dataset. In each simulation, the data are normalized by the panel of methods and significantly DE genes are identified using the MAST test. DE genes are defined as those with a Benjamini and Hochberg adjusted p-value less than 0.01. Average power and FPR is calculated over the 30 simulated datasets (standard error computed across simulations).

Norm. Method	*UMI*	*CPM*	*CPT*	*Scran*	*scTrans*	*scTransCnt*	*Dino*	
PBMC68K_Pure	0.802 (0.008)	0.810 (0.007)	0.853 (0.007)	0.868 (0.007)	0.864 (0.009)	0.864 (0.008)	0.827 (0.010)	**Power**
	0.200 (0.009)	0.149 (0.008)	0.080 (0.006)	0.014 (0.002)	0.004 (0.000)	0.130 (0.007)	0.003 (0.000)	**FPR**
PBMC5K_Prot	0.741 (0.014)	0.776 (0.012)	0.778 (0.013)	0.776 (0.012)	0.768 (0.013)	0.772 (0.014)	0.746 (0.014)	**Power**
	0.255 (0.016)	0.120 (0.015)	0.007 (0.002)	0.001 (0.000)	0.013 (0.001)	0.017 (0.003)	0.002 (0.000)	**FPR**
MaltTumor10K	0.769 (0.011)	0.792 (0.010)	0.838 (0.009)	0.847 (0.010)	0.846 (0.012)	0.859 (0.009)	0.823 (0.011)	**Power**
	0.224 (0.013)	0.134 (0.012)	0.029 (0.005)	0.004 (0.001)	0.003 (0.000)	0.070 (0.007)	0.001 (0.000)	**FPR**
MouseBrain	0.790 (0.013)	0.810 (0.010)	0.809 (0.011)	0.808 (0.011)	0.809 (0.012)	0.834 (0.011)	0.772 (0.012)	**Power**
	0.455 (0.022)	0.184 (0.019)	0.001 (0.001)	0.002 (0.001)	0.001 (0.000)	0.063 (0.010)	0.000 (0.000)	**FPR**
PBMC68K	0.800 (0.009)	0.802 (0.008)	0.835 (0.006)	0.853 (0.007)	0.861 (0.007)	0.852 (0.007)	0.837 (0.007)	**Power**
	0.206 (0.006)	0.161 (0.006)	0.096 (0.005)	0.018 (0.002)	0.008 (0.001)	0.124 (0.012)	0.006 (0.000)	**FPR**
EMT	0.732 (0.013)	0.765 (0.015)	0.788 (0.015)	0.782 (0.015)	0.832 (0.010)	0.823 (0.011)	0.778 (0.013)	**Power**
	0.397 (0.017)	0.175 (0.016)	0.001 (0.000)	0.002 (0.001)	0.011 (0.001)	0.024 (0.004)	0.001 (0.000)	**FPR**

Supplemental Table A.3: **Average power and FPR statistics: t-test.**

30 simulated datasets are produced from each case study dataset. In each simulation, the data are normalized by the panel of methods and significantly DE genes are identified using the t-test. DE genes are defined as those with a Benjamini and Hochberg adjusted p-value less than 0.01. Average power and FPR is calculated over the 30 simulated datasets (standard error computed across simulations).

B Appendix to Chapter 4

B.1 Statistical Methods

B.1.1 Mixed species sample quality control

To assess the quality of alignment to the combined human-mouse transcriptome, misalignment rates were quantified in the H100 (pure human) and M100 (pure mouse) samples. In these cases, transcripts which align to the mouse and human subset of the transcriptome respectively represent errors of misalignment. Typical misalignment rates across samples appeared to be well controlled as the majority of H100 samples aligned less than 0.5% of transcripts to mouse genes (median ~0.35%, third quartile ~0.37%). The majority of M100 samples similarly aligned less than 1.5% of transcripts to human genes (median ~0.53%, third quartile ~1.42%) (S1 Fig).

A few samples (~5%) exhibited high misalignment rates (>5%). For this reason, samples with unusually low sequencing depth were removed. The filtering criteria considered log10 transformed sequencing depth (within sample sum of total expression) and removed samples with depth below the median minus 1.5 times the IQR. This procedure removed the majority of individual samples in H100 and M100 with high alignment error rates. Therefore, misalignment is believed to be primarily a function of, or at least well predicted by, low sequencing depth (Supplemental Figure B.1).

A second filter was implemented to remove samples with expression profiles significantly different from biological replicates of the same time point and temporally neighboring samples. Normalized data (see below for details) from the top 1000 highest variance genes across samples within each mixture was reduced to 10 principal components. This number roughly accounts for

the majority of temporal variability based on the variance explained by each component. Loadings for each component were expected to follow a smooth curve in time, following the portion of the developmental trajectory defined by the principal component. For this reason, loadings were fitted with a 4^{th} degree spline regressed against time. Studentized residuals were tested for being significantly different than the regression curve. A sample level p-value was derived by testing against the null distribution that the maximum residual across the 10 components (in absolute value) was t-distributed. The method of Benjamini and Hochberg(Benjamini and Hochberg, 1995) was used to provide adjusted p-values. A backward elimination and forward selection procedure was then applied. Specifically, the sample with the smallest adjusted p-value below 1e-05 was removed and the process repeated until no samples had an adjusted p-value below 1e-05 (if a sample is the last remaining observation from a particular time point, it was not considered for removal regardless of its adjusted p-value). Samples were then added back in one-at-a-time in the order of removal. Any with adjusted p-values above 1e-05 were retained for further analysis, and otherwise were rejected permanently. The filtered dataset was renormalized prior to analysis.

Empirically, this procedure was shown to remove several remaining high-error samples from M100 without removing high sequencing depth samples across species mixture groups (Supplemental Figure B.1).

B.1.2 Normalization of mixed species samples

We used a modified application of the scran(Lun *et al.*, 2016) method for normalization of the expected count data. Human and mouse aligned transcripts were normalized separately, and so relative levels of normalized expression were not directly comparable between species. Consider the human mixtures (H10, H85, or H100); mouse mixtures were normalized identically. When

biological replicates existed for a time point, scran was first applied to normalize these samples. Average normalized expression of biological replicates was then normalized, again via scran, across both time points and mixtures.

B.1.3 Segmented regression and gene-trend classification

The dynamics of gene expression through time were defined by a segmented regression implemented using the Trendy(Bacher *et al.*, 2018) package. Trendy automatically selects the optimal number of segments (up to a maximum of 5 in this application) and requires that each segment contain a minimum number of samples (5 in this application). Additionally, an automatic significance test on segment slopes classifies segments as increasing, decreasing, or flat. As the test is itself somewhat conservative, we used a significance threshold of 0.1 (default) to determine these slope classifications. Trendy was then applied to all genes for which the 80% quantile of normalized expression is above 20 for at least one mixture.

Following regression, the segment trend classifications were used to define sets of genes by patterns of behavior relative to a reference dataset (H100 in the majority of the published analysis). Genes were classified into subsets of accelerated or differentially expressed (DE) relative to the reference dataset according to the following criteria:

1. Accelerated by Early Up (EU):
    a. Both the test gene and the reference gene contain an increasing segment which is not preceded by a decreasing segment. If multiple such segments exist, only the first is considered.
    b. The increasing segment in the test gene must start at least 2 days before the increasing segment in the reference gene.

c. The slope of the increasing segment in the test gene must be at least 5 times the slope of the (non-increasing) reference segment which contains the start time of the test increasing segment (typically the segment just prior to the increasing reference segment). This filter removes genes for which the reference segment containing the start time is labeled as flat by Trendy (slope is not significantly different from 0), but is fitted with an up-trending slope. This can happen in instances where the reference segment is short and so does not contain enough sample points for the up-trend to be labeled as significant.

2. Accelerated by Early Peak (EP):

    a. Both the test gene and the reference gene contain a peak defined by an increasing segment followed by a flat or decreasing segment. The peak itself is defined by the time of the breakpoint between these two segments.

    b. The peak in the test gene must be at least 2 days before the peak in the reference gene.

3. DE Up:

    a. The maximum fitted value of the test gene plus 1 must be at least 3 times the maximum fitted value of the reference gene plus 1. The inclusion of the plus 1 bias to each side prevents very lowly expressing genes from appearing DE due to small differences in fitted values which are only multiplicatively large due to the low overall expression.

Genes in H10 or H85 matching these acceleration/up-regulation criteria were denoted as "Early" or "Up" respectively.

We also ran this classification denoting H100 as the test datasets. When genes matched the criteria in this case, we denoted the corresponding gene in the reference dataset, H10 or H85, “Late” or “Down” according to the specific criteria met.

B.1.4 Acceleration factor estimation

Point estimates of the relative acceleration of one dataset compared to another were computed from genes which either peak in both datasets or trend up in both datasets. For simplicity, consider the case of H10 relative to H100. From peaking/up-trending genes, the event time was calculated: time of peak or time of the start of up-trend respectively. When a gene both peaks and trends up, the peak was preferred as it was assumed to be a more accurate estimate of regulatory changes. Up-trending genes (without peaks) which start up-trending in either H10 or H100 on day 0 were discarded. Point estimates were then calculated as the ratio of the event time in H100 to the event time in H10. In this way, a ratio of 2 would indicate that, at the time of the even in H10, that gene is accelerated to be 2x as fast as the gene in H100. Point estimates are computed across pairs of datasets.

To compute a continuous estimate of acceleration factors, the above point estimates are smoothed using spline regression (linear model in R with a basis spline under default parameters) against event time in the test (e.g., H10) dataset. It should be noted that these acceleration factors are best interpreted as estimates of the relative acceleration of the genes which are active at that time point. Acceleration factors of 1 therefore identify time points, and thereby sets of genes active at that time point, which are relatively unchanged between the conditions.

B.1.5 Gene set enrichment

Accelerated and DE gene sets were further characterized through testing for GO term enrichment. The topGO(Alexa *et al.*, 2006) package and org.Hs.eg.db(Carlson, 2019) dataset were used to perform enrichment testing on GO terms belonging to the biological processes (BP) ontology. The set of all genes on which Trendy segmented regression was run was used as the background set (see above for subset definition). Significant p-values were then FDR corrected(Benjamini and Hochberg, 1995) prior to analysis.

Pathway and transcription factor/miRNA enrichment was performed in a similar manner. In these cases, the piano(Väremo *et al.*, 2013) package was used to accommodate non-binary statistics. Specifically, enrichment was performed on the difference between up-trend or peak events between a test dataset (e.g., H10) and a reference dataset (e.g., H100). When available, the difference was calculated from the time of peaks in each dataset. Absent peaks, the difference was calculated from the time of the start of up-trends. Genes without either shared peaks or shared up-trends were given a difference of 0.

Enrichment for these differences were performed against two collections of gene sets from the MSigDB database(Subramanian *et al.*, 2005; Liberzon *et al.*, 2015). The first was a curated collection of pathways, including KEGG(Kanehisa, 2000, 2019; Kanehisa *et al.*, 2021), Biocarta(Nishimura, 2001), and Reactome(Jassal *et al.*, 2020) sets of gene pathways. The second was a collection of miRNAs(Chen and Wang, 2020) and transcription factors(Yevshin *et al.*, 2019) (TFs) and downstream regulated genes. Enrichment was performed with the runGSA function from the piano package (4e6 permutations, minimum gene set size of 1, maximum gene set size of 250).

B.1.6 Sorted sample quality control validation

Sorted samples, sH100, sH10, sM90, and sM100 were similarly aligned to a combined transcriptome (as described above) to provide a validation dataset. One data point was removed for low sequencing depth (day 29 from sH10, fewer than 1e3 expected counts where typical sorted samples had greater than 1e6 expected counts) and all others were retained.

Empirical misalignment rates were computed for sH100 and sH10 as the fraction of expected counts aligned to the mouse portion of the transcriptome; median values across days were 0.53% and 2.23% respectively.

Active misaligned genes were identified as genes in the off-target portion of the reference transcriptome (e.g., mouse genes for sH10) with an 80% quantile of expected counts ≥ 20. Enrichment following the above-described procedure was performed on these gene sets.

Normalization was performed using the calculateSumFactors function in scran (default parameters) to compute scale factors which expected counts were then divided by. As with the other data, Trendy was used to perform segmented regression (maximum 4 breakpoints, minimum 2 points per segment, p-value threshold 0.1). Output from Trendy was used to classify genes as EU/LU. Peak analysis was omitted as the lower resolution of the data prevent robust identification of peaks (e.g., visually identifiable peaks are not significant under Trendy regression). Acceleration factor estimation was computed from these shared up-trend genes in the above-described manner, and enrichment was performed on EU genes, again as above.

B.1.7 Correlation analysis

Expression similarity across time points, species mixtures, and external reference datasets was assessed through gene expression correlations. To ensure that computed correlations were

representative of the temporal gene dynamics being studied, correlations were computed on only a subset of genes. Highly dynamic genes were subset from all Trendy-fit genes by calculating the coefficient of variation of fitted values. The highest CV across species mixtures was then retained as a measure of each gene's level of temporal dynamics, and the top 2000 most dynamic (highest CV) genes were subset for analysis.

Relative similarity of species-mixtures was computed as the correlation matrix (spearman type) between time points where within-day biological replicates were averaged together to obtain a single day expression value.

Similar calculations of correlations between the species-mixture data and two outside datasets, the BrainSpan atlas of the developing human brain(Miller *et al.*, 2014; Allan Human Brain Atlas: BrainSpan (Atlas of the Developing Brain)) and the Human protein atlas(Yu *et al.*, 2015; RNA FANTOM brain region gene data), were conducted. In these cases, the genes used to calculate correlations were the union of the top 1500 most dynamic genes from H10/H85/H100 and the top 1500 most dynamic genes (highest CV across cell-types) from the relevant in vivo reference dataset.

B.1.8 Correlation-based acceleration

To use in vitro correlation heatmaps to estimate acceleration factors, we adapted a technique described in Rayon et al. 2020(Rayon *et al.*, 2020). Specifically, we performed a version of weighted regression whereby the weights derive from the correlation values. However, as the in vitro data was observed to not have a constant acceleration factor, we performed segmented regression with a fixed breakpoint at day 16. The specific function to minimize was then:

$$\min_{\theta^*}\left\{\sum_{i,j} cor\left(EC_{t_i}, EC_{t_j}\right)^2 dist_{\perp}\left(\left(t_i, t_j\right)|\theta^*\right)^2\right\}$$

where *EC* denotes expected counts (correlation is spearman type, so normalization in unnecessary), t_i and t_j denote days in the reference and test datasets respectively (e.g., H100 and H10), and $dist_{\perp}()$ denotes the perpendicular distance to the current estimate of the segmented regression given regression coefficients θ* from the provided time pair (coordinates on the correlation heatmap). Minimization was conducted in R using the optim function (L-BFGS-B method, upper and lower bounds of 10 and 1/10 respectively, segmented regression fixed to pass through (0, 0), initial slopes set to 1 in each segment). Standard errors for coefficient estimates were generated by bootstrapping solutions from random samples (with replacement) of the input genes. Regression slopes then defined the desired acceleration factor up to an inversion.

B.1.9 In vivo dissimilarity

Dissimilarity between in vitro data (average across biological replicates for a given day and species mixture) and in vivo references was computed from highly dynamic genes (see criteria above in Correlation analysis) using a variation of principal component analysis (PCA). To accommodate the distributional properties of these sequencing data, as well as the properties of the reference data, a variation on PCA, glmpca(Townes, 2019; Townes *et al.*, 2019), which uses a negative binomial model residuals was used to perform dimension reduction to 6 dimensions (6 principal components). Dissimilarity was then computed as the distance (Euclidian) between an in vitro data point in the low dimensional space and the corresponding low dimensional representation of a reference in vivo data point. glmpca was run with the negative binomial family, fisher optimizer, penalty of 10, minimum iterations of 400, and was parameterized by size factors derived from

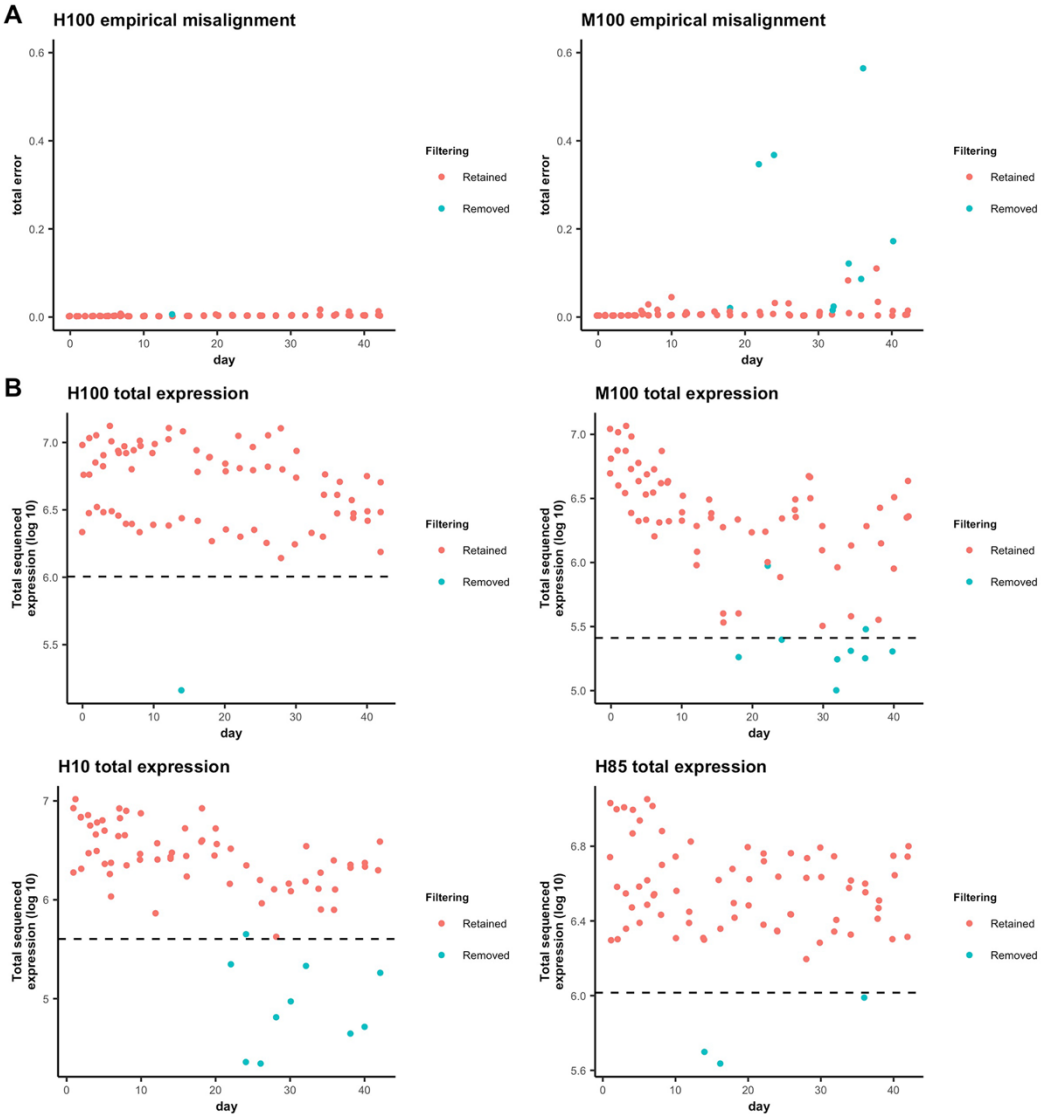
Scran to normalize the (unnormalized) expected counts from the in vitro data and the in vivo reference.

Note that the BrainsSpan data were available as reads per kilobse million (rpkm) rather than the expected counts (EC) used in this analysis. For this reason, analysis on the BrainSpan data was conducted using fpkm from the in vitro data as the best available analog.

B.1.10 Deconvolution analysis

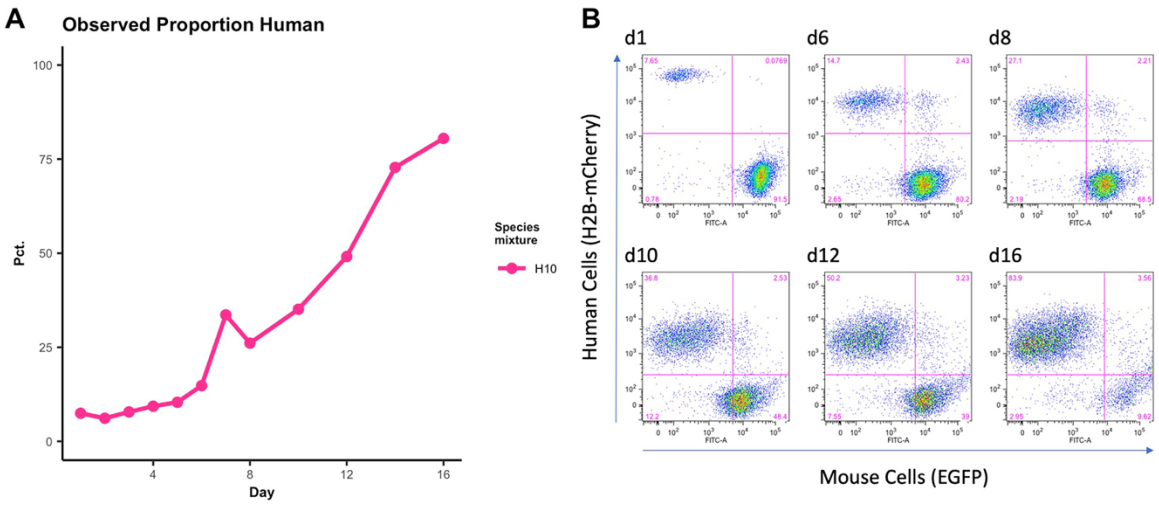
Deconvolution analyses to estimate proportions of cell-types in the observed bulk sequencing data were performed using the music_prop function (default parameters) from the MuSiC(Wang *et al.*, 2019) package with the CoDEx database of annotated developing brain cells(Polioudakis *et al.*, 2019) as reference. Within species-mixture, estimated proportions were smoothed across biological replicates and days using the DirichletReg function from the DirichletReg package(Maier, 2020) and a basis spline (df = 4).

B.2 Supplemental Figures



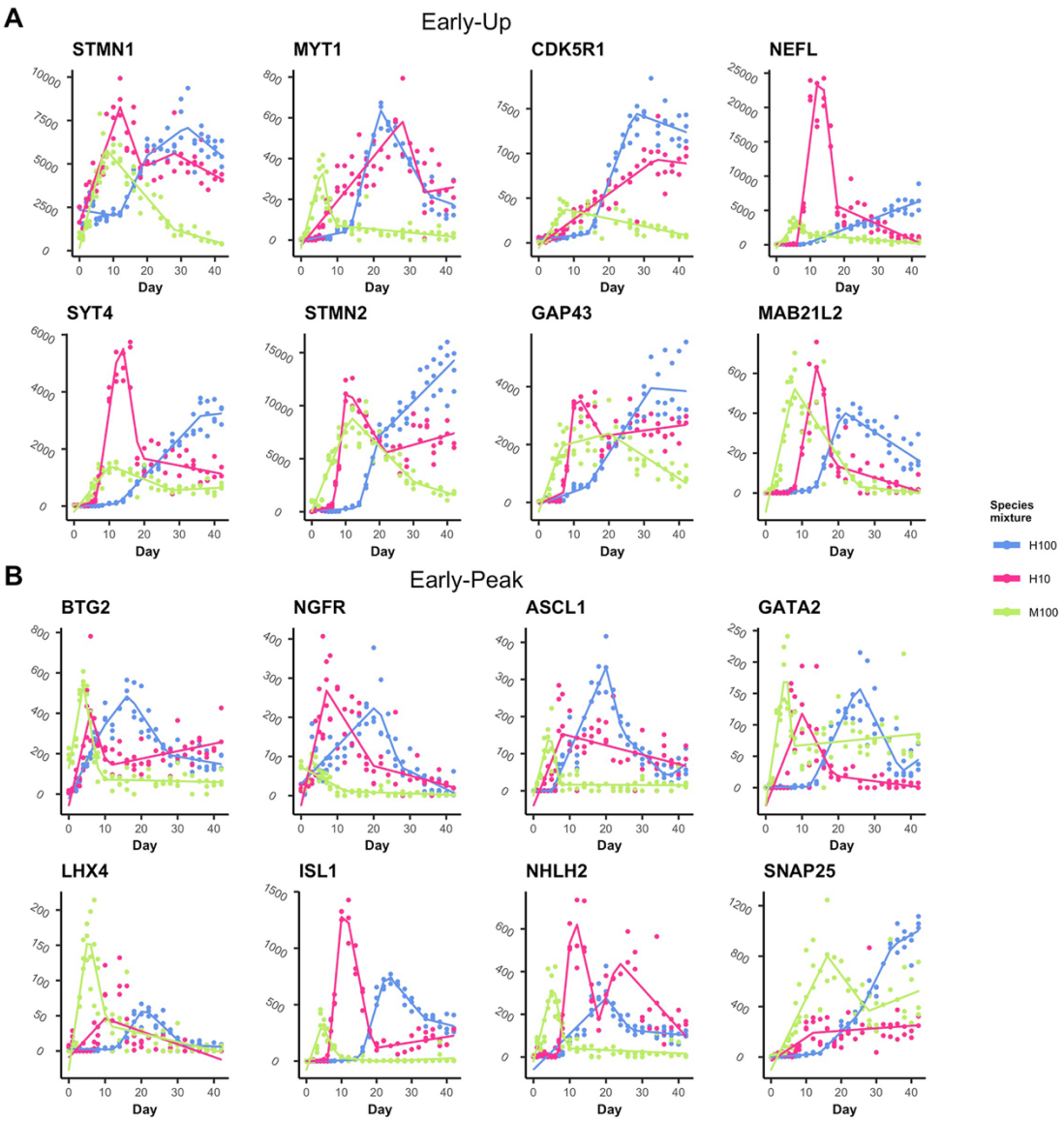
Supplemental Figure B.1: **Quality control filtering removes samples with uncharacteristically low sequencing depth.**

(A) Observed per-sample misalignment rates for pure human (H100)/pure mouse (M100) mixtures. (B) Observed log10 total sequencing depth summed across sequences aligned to either human or mouse. Most samples removed from analysis (blue) are below the depth filtering threshold (dashed line) (see Materials and Methods). Otherwise, the M100 results suggest that the higher-depth removed samples are those with higher rates of misalignment (top/middle, right column).



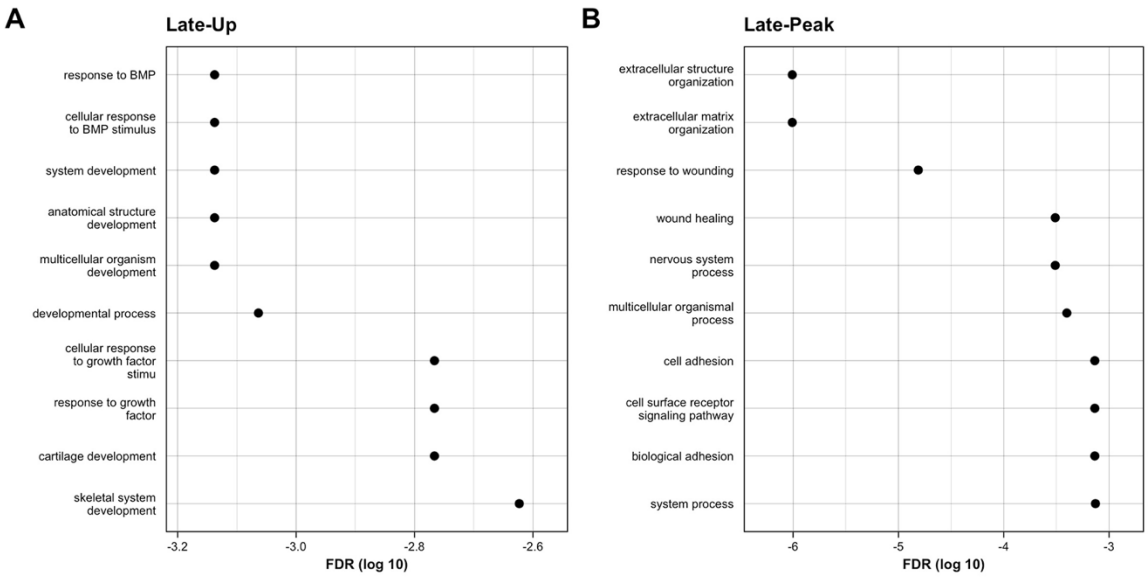
Supplemental Figure B.2: **Seeded human cell proportions increase over time.**

(A) Observed percent of human cells in H10 mixture out to 16 days. (B) FACS plots intensities used to compute relative proportions of human and mouse cells in H10 mixture.



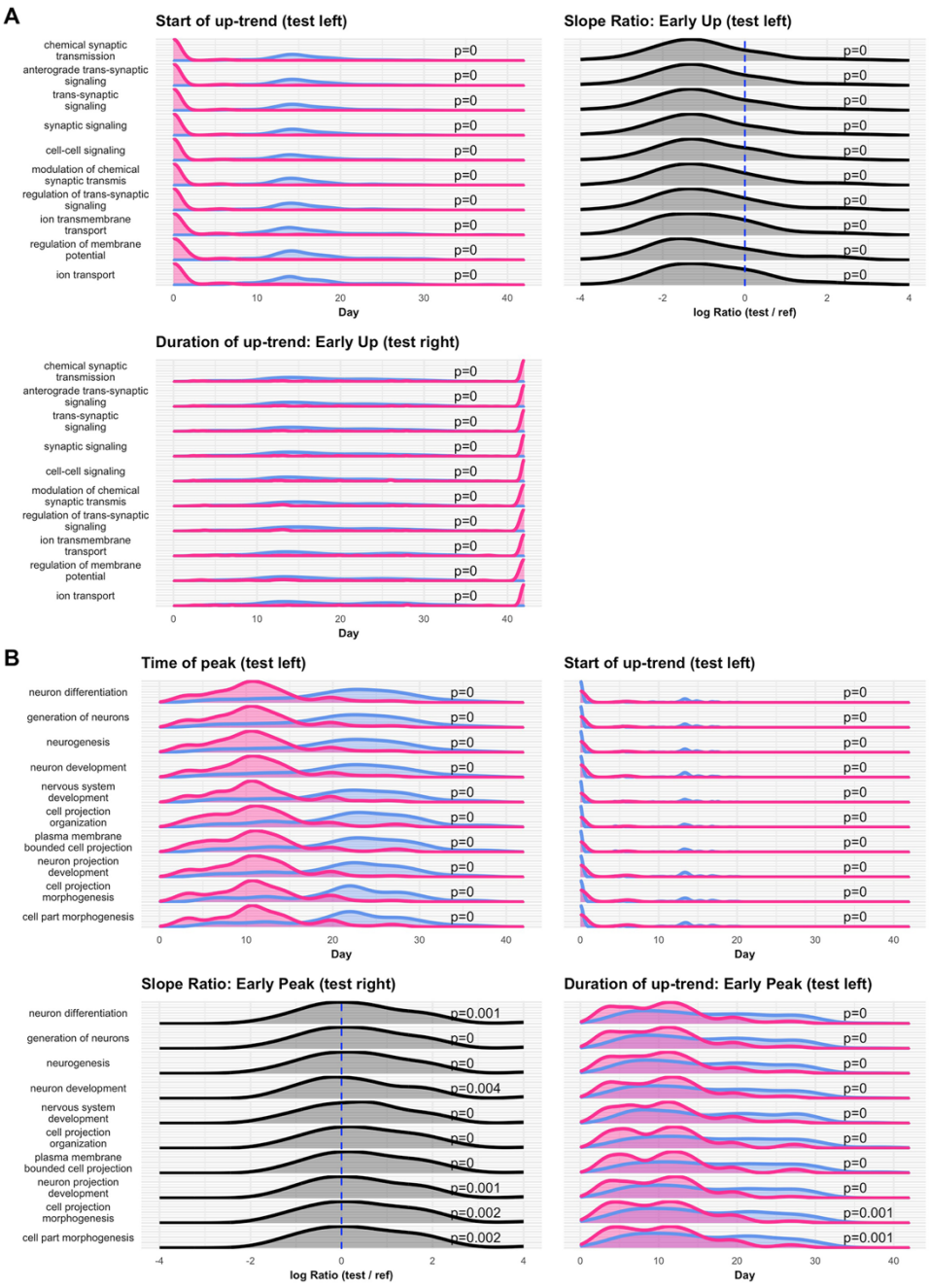
Supplemental Figure B.3: **Selected gene expression plots show characteristic differences between H100, H10, and M100.**

(A) Early-Up classified fitted trend lines (solid) are plotted for selected genes with overlaid normalized observed data (points). (B) Similar results are shown for selected Early-Peak classified genes (green = M100, pink = H10, blue = H100).



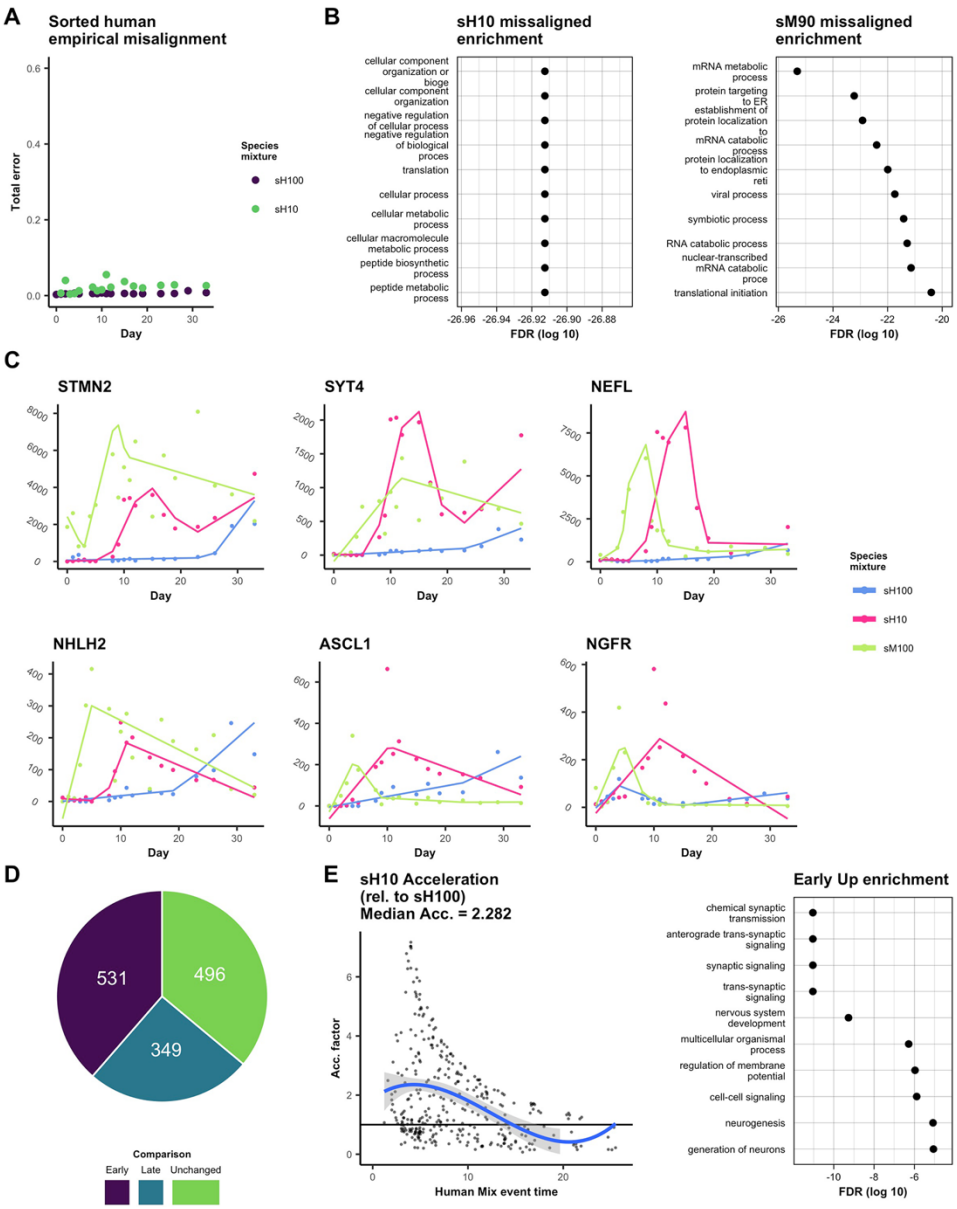
Supplemental Figure B.4: **Enrichment of late-up (LU) and late-peak (LP) genes fail to demonstrate a pattern of neuron development-related terms.**

(A) Top GO terms enriched for LU genes in H10 compared to H100 with corresponding FDR corrected p-values (log 10 scale). (B) Top GO terms enriched for LP genes in H10 compared to H100 with corresponding FDR corrected p-values (log 10 scale).



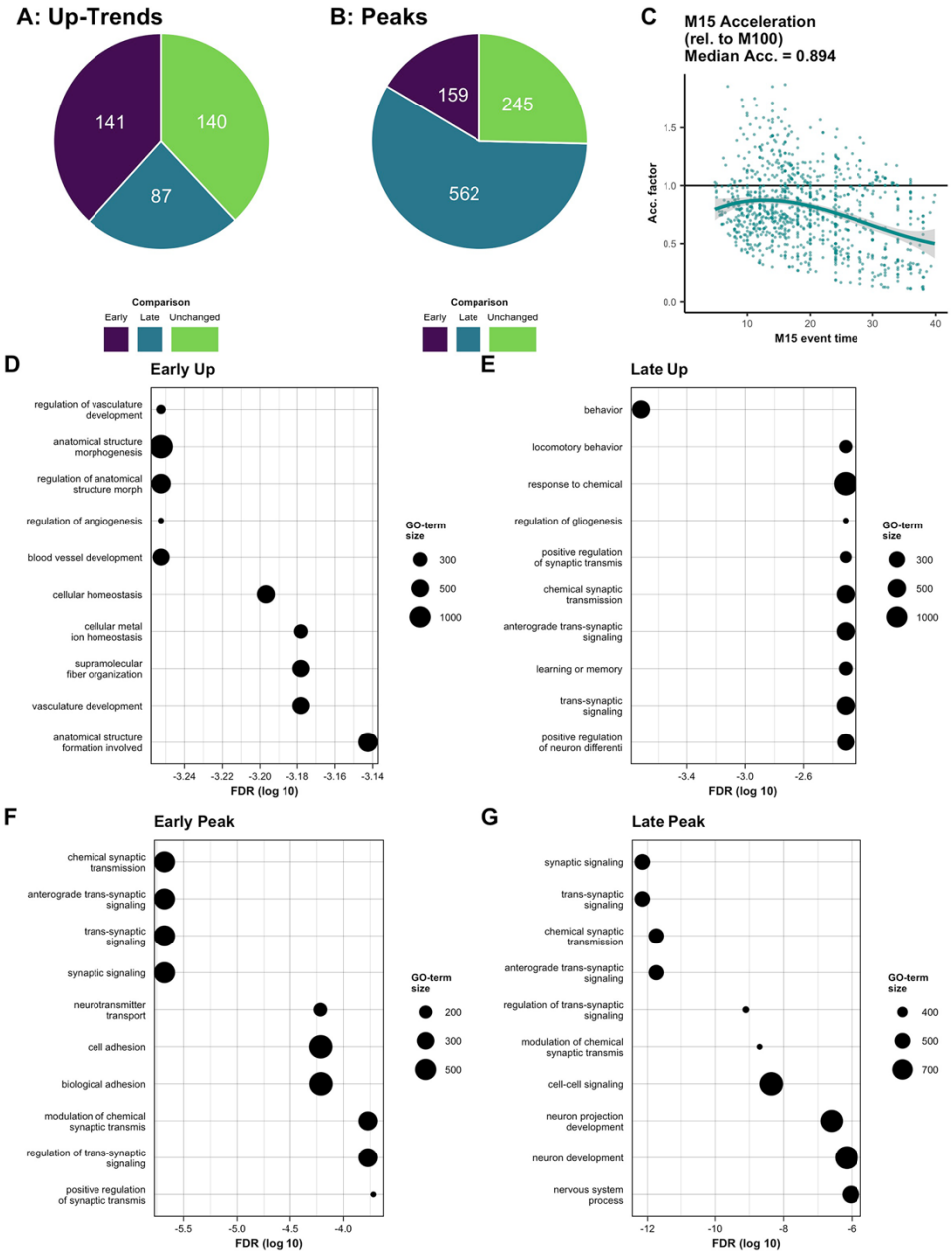
Supplemental Figure B.5: **Up-trends show defining shifts in H10 among EU and EP genes.**

(A) EU genes from each of the listed GO terms are plotted. The start of uptrends between H10 and H100 are plotted (top left) with KS testing sowing significant left shift corresponding to significantly earlier trend starts in H10. Slope ratio (ratio of H10 up-trend slope over H100 up-trend slope) densities are plotted (top right) on the log scale for top enriched GO terms with KS testing showing a significant left-shift corresponding to significantly reduced slopes in H10 among these genes. Densities of the duration of up-trends (bottom left) show significantly longer (KS test) trends for H10 (red) than H100 (blue). (B) EP genes from each of the listed GO terms are plotted. The timing of peaks are plotted (top left) with KS testing showing significant left shift corresponding to significantly earlier peaks in H10. Similar results for EP genes as the above EU genes show significantly earlier up-trend starts, significant increases in slope in H10, and reduced duration of up-trends (pink = H10, blue = H100).



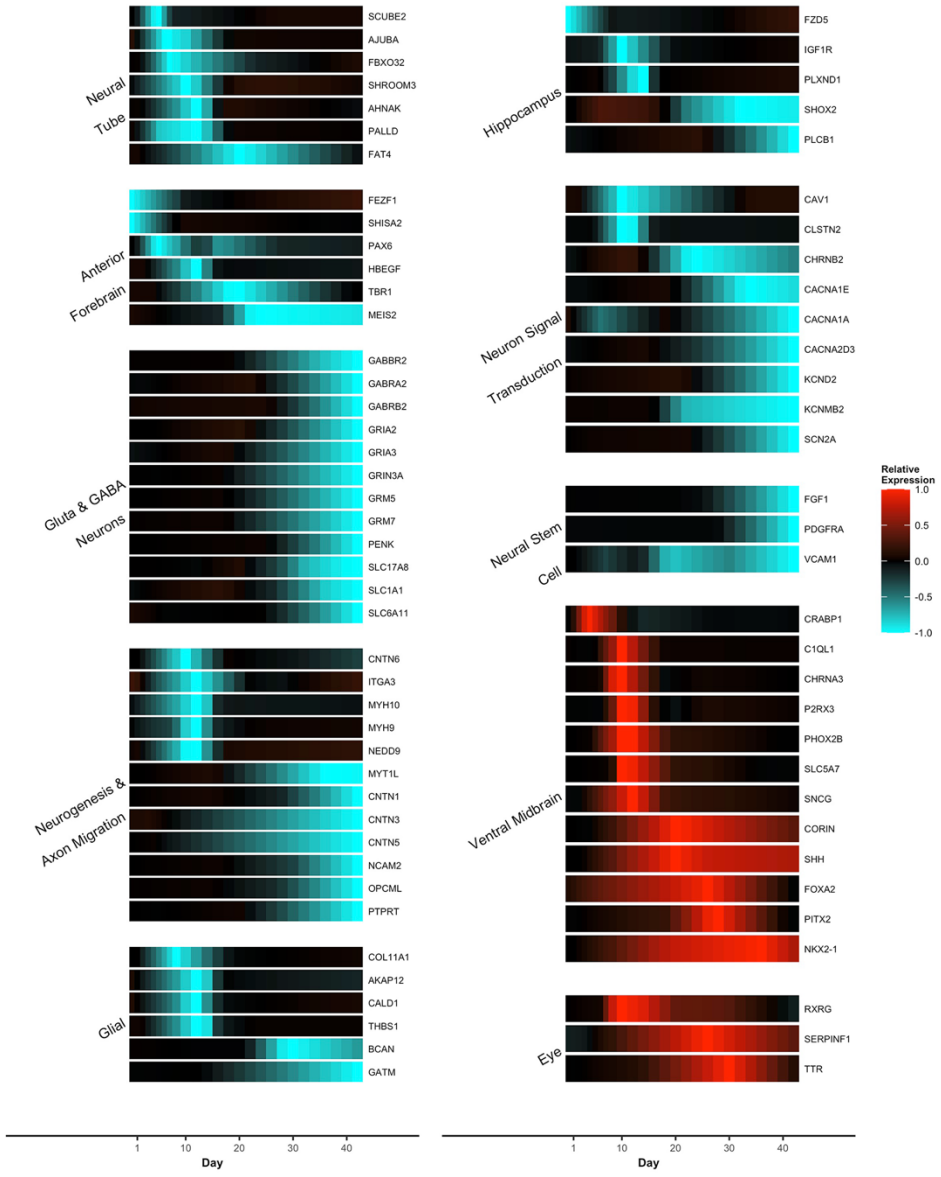
Supplemental Figure B.6: **Expression from sorted co-culture cells fails to show misalignment bias.**

(A) Empirical misalignment for sH100 and sH10 are plotted by day. (B) Misaligned genes for the sH10 and sM90 (mouse and human aligned reads respectively) are subset. Enrichment testing is performed on active genes, defined as those with 80% quantile of observed expression of at least 20 expected counts, and top terms are plotted against FDR corrected p-values (log 10 scale). (C) Expression from selected genes which are accelerated in the H10-H100 comparison are plotted for sH100, sH10, and sM100, and show similar acceleration effects in this sorted control dataset. (D) EU/LU genes are tabulated for sH10. (E) Continuous acceleration factors are calculated for sH10 and top EU enriched GO terms are plotted.



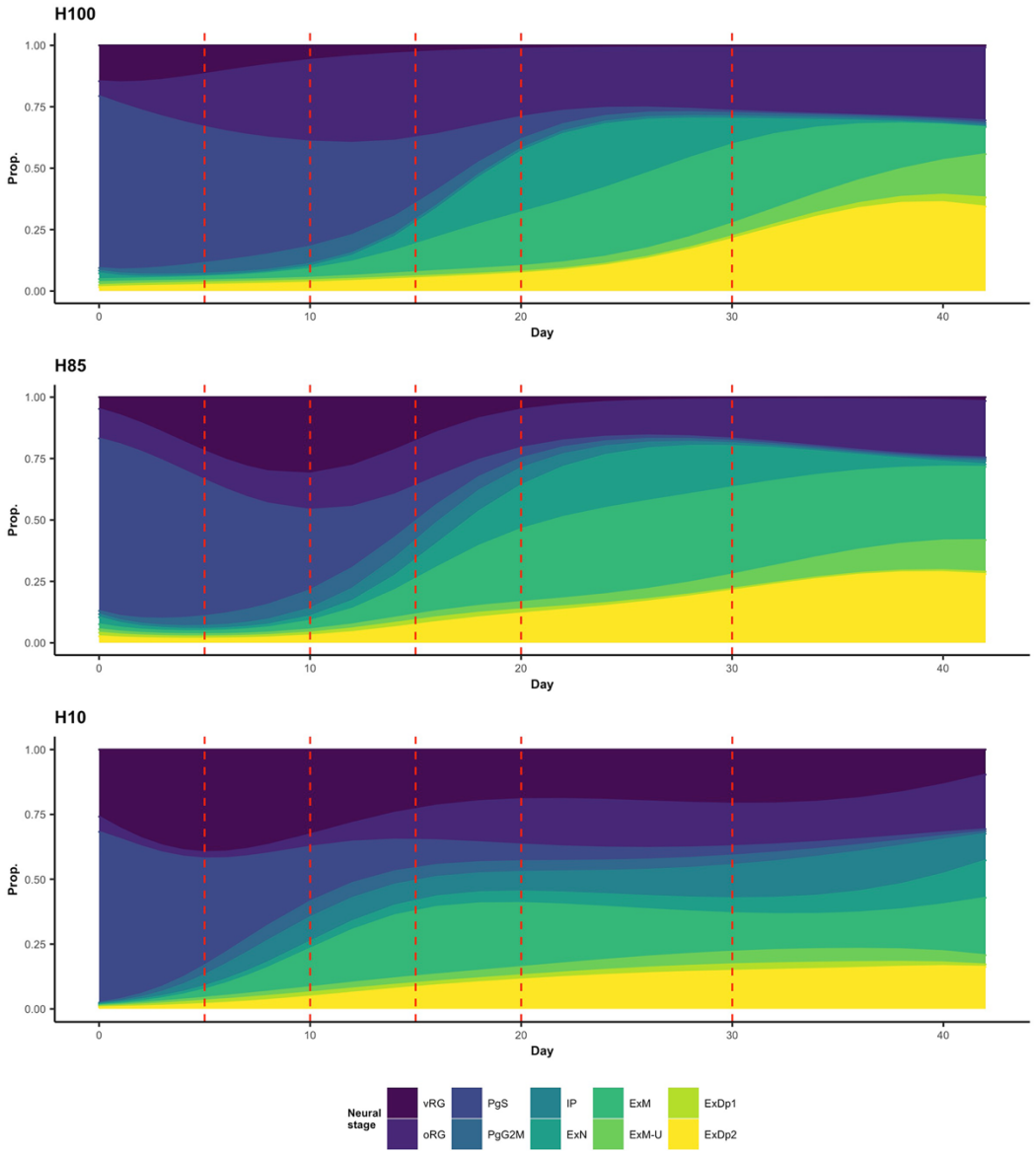
Supplemental Figure B.7: **Analysis of co-cultured mouse expression suggests deceleration of mouse gene expression patterns.**

(A-B) Genes identified as shared up-trends (excluding those which start to trend up on day 0 in both M100 and M15) or shared peaks between M15 and M100 are classified as either early, late, or unchanged, and then tabulated. (C) Shared up-trending and peaking genes are used to estimate a continuous acceleration factor for M15 relative to M100 in an identical manner to the human data. The median acceleration factor (over the first 16 days) of 0.894 indicates a deceleration in gene activity. (D-G) Top terms enriched for EU, LU, EP, and LP genes respectively are plotted against FDR corrected p-values. Neural associated terms are either unique to the late category or are more significant in that group, suggesting a deceleration effect specific to neural genes.



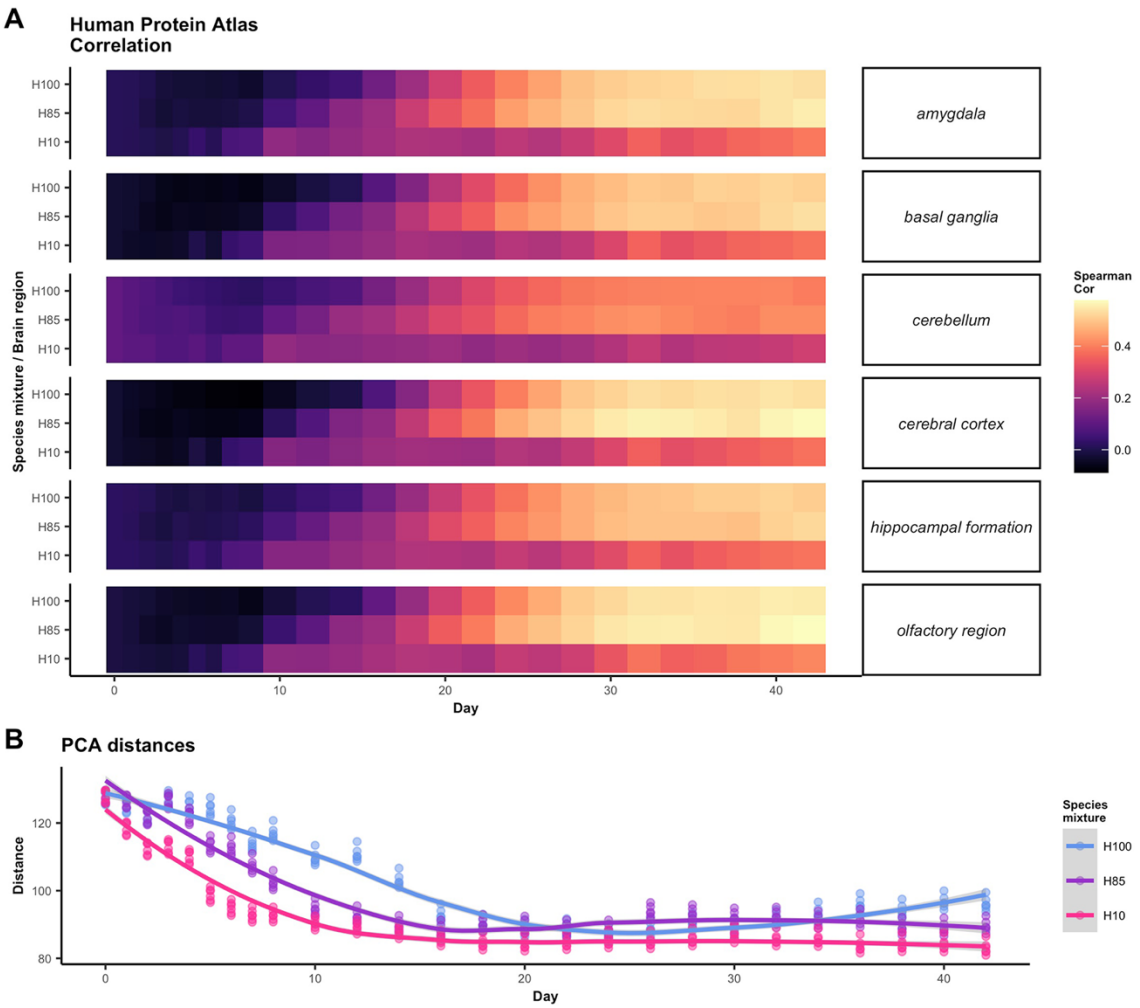
Supplemental Figure B.8: **Up/down regulation of genes in H10 show region specific patterns.**

Relative expressions of curated genes in regional/functional groups are plotted on a normalized -1 to 1 scale. Gene expression (within gene) is normalized such that the maximum difference in fitted expression (in H100 or H10) equals 1. Relative expressions are then calculated as the difference between H10 and H10 where higher H10 values tend towards 1 (red), lower H10 values tend towards -1 (blue), and equivalent values tend towards 0 (black).



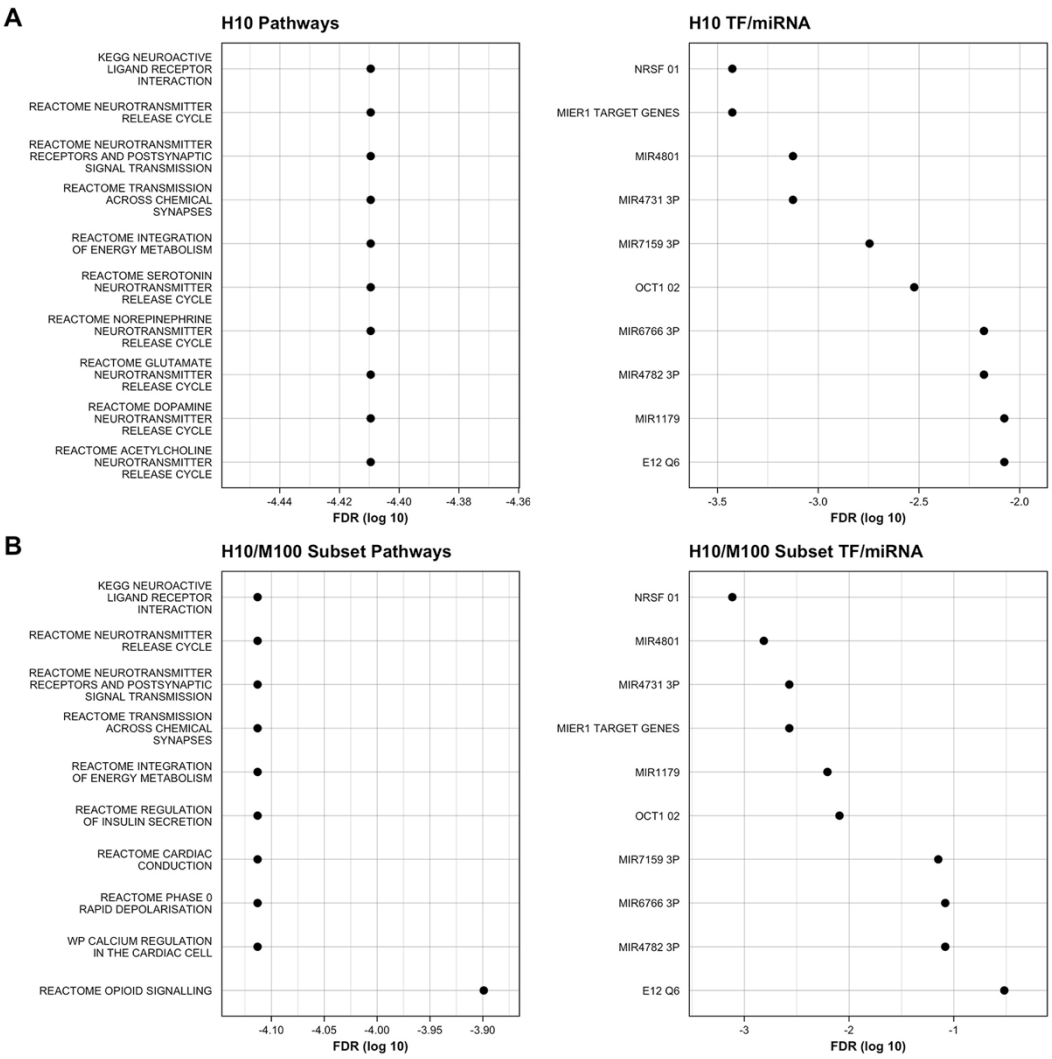
Supplemental Figure B.9: **Deconvolution analysis of mixed-species data supports dose-response effect.**

Expression data for H100, H85, and H10 respectively are deconvolved relative to the CoDEx reference dataset of annotated developing brain single cell expression. Deconvolution produces estimates of the relative proportions of reference cell-types present in the bulk data. Estimates are smoothed against time and plotted for each of H100 (top), H85 (middle), and H10 (bottom).



Supplemental Figure B.10: **Correlation with Human Protein Atlas (HPA) data further demonstrates dose response behaviors.**

Correlations (Spearman) between fitted trends HPA data are calculated across the thirteen HPA regions. Calculations are performed on a subset of highly dynamic genes (see Materials and Methods). Dissimilarity (PCA-based distance, see Materials and Methods) between species mixtures and each of 6 HPA cell-types are computed for each day and smoothed to estimate a continuous dissimilarity metric over time.



Supplemental Figure B.11: **Candidate pathways, transcription factors (TFs), and miRNAs mediate the observed acceleration.**

(A) Top pathways (left) and TFs/miRNAs (right) enriched for acceleration in H10 are plotted against their FDR corrected p-values. (B) Similar analysis is performed on M100 orthologs compared to H100 expression. Prior to plotting top pathways (left) and TFs/miRNAs (right), enriched terms are subset to include only those which are also significant (FDR corrected p-value ≤ 1e-2) in the above H10 comparison.

References

Alahverdi,A. *et al.* (2020) Involvement of EGFR, ERK-1,2 and AKT-1,2 Activity on Human Glioma Cell Growth. *Asian Pacific J. Cancer Prev.*, **21**, 3469–3475.

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Allan Human Brain Atlas: BrainSpan (Atlas of the Developing Brain).

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Andrew,G. and Gao,J. (2007) Scalable training of L1-regularized log-linear models. *ACM Int. Conf. Proceeding Ser.*, **227**, 33–40.

Árnason,Ú. *et al.* (2018) Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Sci. Adv.*, **4**.

Bacher,R. *et al.* (2017) SCnorm: Robust normalization of single-cell RNA-seq data. *Nat. Methods*, **14**, 584–586.

Bacher,R. *et al.* (2018) Trendy: Segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. *BMC Bioinformatics*, **19**, 1–10.

Bacher,R. and Kendziorski,C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 1–14.

Barry,C. *et al.* (2019) Automated minute scale RNA-seq of pluripotent stem cell differentiation reveals early divergence of human and mouse gene expression kinetics. *PLoS Comput. Biol.*, **15**, 1–24.

Barry,C. *et al.* (2017) Species-specific developmental timing is maintained by pluripotent stem cells ex utero. *Dev. Biol.*, **423**, 101–110.

Bauwens,C.L. *et al.* (2008) Control of Human Embryonic Stem Cell Colony and Aggregate Size Heterogeneity Influences Differentiation Trajectories. *Stem Cells*, **26**, 2300–2310.

Bayly,R.D. *et al.* (2012) A novel role for FOXA2 and SHH in organizing midbrain signaling centers. *Dev. Biol.*, **369**, 32–42.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Berahas,A.S. and Takáč,M. (2020) A robust multi-batch L-BFGS method for machine learning. *Optim. Methods Softw.*, **35**, 191–219.

Branham, R. L.,J. (1982) Alternatives to least squares. *Astron. J.*, **87**, 928.

Broccoli,V. *et al.* (2014) Modeling physiological and pathological human neurogenesis in the dish. *Front. Neurosci.*, **8**, 1–9.

Brons,I.G.M. *et al.* (2007) Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, **448**, 191–195.

Brown,J., Barry,C., *et al.* (2021) Interspecies chimeric conditions affect the developmental rate of human pluripotent stem cells. *PLOS Comput. Biol.*, **17**, e1008778.

Brown,J., Ni,Z., *et al.* (2021) Normalization by distributional resampling of high throughput single-cell RNA-sequencing data. *Bioinformatics*.

Brown,J.H. *et al.* (2004) TOWARD A METABOLIC THEORY OF ECOLOGY. *Ecology*, **85**, 1771–1789.

Brown,M.F. *et al.* (2007) Metabolic rate does not scale with body mass in cultured mammalian cells. *Am. J. Physiol. Integr. Comp. Physiol.*, **292**, R2115–R2121.

Bruce,A.W. *et al.* (2004) Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc. Natl. Acad. Sci.*, **101**, 10458–10463.

Burton,P.B.J. *et al.* (1999) An intrinsic timer that controls cell-cycle withdrawal in cultured cardiac myocytes. *Dev. Biol.*, **216**, 659–670.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Carlson,M. (2019) org.Hs.eg.db: Genome wide annotation for Human.

Chen,S.X. (2000) Probability Density Function Estimation Using Gamma Kernels. *Ann. Inst. Stat. Math.*, **52**, 471–480.

Chen,Y. and Wang,X. (2020) MiRDB: An online database for prediction of functional microRNA targets. *Nucleic Acids Res.*, **48**, D127–D131.

Chetty,S. *et al.* (2013) A simple tool to improve pluripotent stem cell differentiation. *Nat. Methods*, **10**, 553–556.

Chu,L.F. *et al.* (2019) An In Vitro Human Segmentation Clock Model Derived from Embryonic Stem Cells. *Cell Rep.*, **28**, 2247-2255.e5.

Chuang,J.H. *et al.* (2013) Differentiation of glutamatergic neurons from mouse embryonic stem cells requires raptor S6K signaling. *Stem Cell Res.*, **11**, 1117–1128.

Cordy,C.B. and Thomas,D.R. (1997) Deconvolution of a distribution function. *J. Am. Stat. Assoc.*,

92, 1459–1465.

D'Amour,K.A. *et al.* (2005) Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotechnol.*, **23**, 1534–1541.

Dale,J.K. *et al.* (1997) Cooperation of BMP7 and SHH in the induction of forebrain ventral midline cells by prechordal mesoderm. *Cell*, **90**, 257–269.

Das,S. *et al.* (2020) Generation of human endothelium in pig embryos deficient in ETV2. *Nat. Biotechnol.*, **38**, 297–302.

Dessaud,E. *et al.* (2008) Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network. *Development*, **135**, 2489–2503.

Dias,J.M. *et al.* (2020) A Shh/Gli-driven three-node timer motif controls temporal identity and fate of neural stem cells. *Sci. Adv.*, **6**, eaba8196.

Diaz-Cuadros,M. *et al.* (2020) In vitro characterization of the human segmentation clock. *Nature*, **580**, 113–118.

Ebisuya,M. and Briscoe,J. (2018) What does time mean in development? *Dev.*, **145**, 0–3.

Espuny-Camacho,I. *et al.* (2013) Pyramidal Neurons Derived from Human Pluripotent Stem Cells Integrate Efficiently into Mouse Brain Circuits In Vivo. *Neuron*, **77**, 440–456.

Finak,G. *et al.* (2015) MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 1–13.

Gao,F.B. *et al.* (1997) Oligodendrocyte precursor cells count time but not cell divisions before differentiation. *Curr. Biol.*, **7**, 152–155.

Gao,Y. *et al.* (2020) MicroRNA-1179 suppresses the proliferation and enhances vincristine sensitivity of oral cancer cells via induction of apoptosis and modulation of MEK/ERK and PI3K/AKT signalling pathways. *AMB Express*, **10**, 149.

Gaspard,N. *et al.* (2008) An intrinsic mechanism of corticogenesis from embryonic stem cells. *Nature*, **455**, 351–357.

Gholizadeh,N. *et al.* (2020) Association of MAPK and its regulatory miRNAs (603, 4301, 8485, and 4731) with the malignant transformation of oral lichen planus. *Mol. Biol. Rep.*, **47**, 1223–1232.

Greber,B. *et al.* (2010) Conserved and Divergent Roles of FGF Signaling in Mouse Epiblast Stem Cells and Human Embryonic Stem Cells. *Cell Stem Cell*, **6**, 215–226.

Grün,D. *et al.* (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

Gurok,U. *et al.* (2004) Gene expression changes in the course of neural progenitor cell differentiation. *J. Neurosci.*, **24**, 5982–6002.

Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296.

Hamilton,M.J. *et al.* (2011) Universal scaling of production rates across mammalian lineages. *Proc. R. Soc. B Biol. Sci.*, **278**, 560–566.

Haque,A. *et al.* (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, **9**, 1–12.

Harris,W.A. and Hartenstein,V. (1991) Neuronal determination without cell division in xenopus

embryos. *Neuron*, **6**, 499–515.

Hodge,R.D. *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature*, **573**, 61–68.

Hu,Z. *et al.* (2020) Transient inhibition of mTOR in human pluripotent stem cells enables robust formation of mouse-human chimeric embryos. *Sci. Adv.*, **6**, 1–17.

Huang,M. *et al.* (2018) SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.

Huang,W. *et al.* (2015) A broyden class of quasi-Newton methods for riemannian optimization. *SIAM J. Optim.*, **25**, 1660–1685.

Hwang,B. *et al.* (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, **50**.

Islam,S. *et al.* (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.

Jamshidian,M. and Jennrich,R.I. (1997) Acceleration of the EM Algorithm by using Quasi-Newton Methods. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, **59**, 569–587.

Jassal,B. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Kanehisa,M. *et al.* (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.

Kanehisa,M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kanehisa,M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein*

Sci., **28**, 1947–1951.

Kanton,S. *et al.* (2019) Organoid single-cell genomic atlas uncovers human-specific features of brain development.

Kasai,F. *et al.* (2013) Afrotheria genome; overestimation of genome size and distinct chromosome GC content revealed by flow karyotyping. *Genomics*, **102**, 468–471.

Kelava,I. and Lancaster,M.A. (2016) Stem Cell Models of Human Brain Development. *Cell Stem Cell*, **18**, 736–748.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Kolodziejczyk,A.A. *et al.* (2015) The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell*, **58**, 610–620.

Lancaster,M.A. *et al.* (2013) Cerebral organoids model human brain development and microcephaly. *Nature*, **501**, 373–379.

Law,C.W. *et al.* (2014) Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, 1–17.

van de Leemput,J. *et al.* (2014) CORTECON: A temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. *Neuron*, **83**, 51–68.

Liberzon,A. *et al.* (2015) The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.*, **1**, 417–425.

Liu,D.C. and Nocedal,J. (1989) On the limited memory BFGS method for large scale optimization. *Math. Program.*, **45**, 503–528.

De Los Angeles,A. *et al.* (2018) Generating human organs via interspecies chimera formation: Advances and barriers. *Yale J. Biol. Med.*, **91**, 333–342.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 1–21.

Lun,A.T.L. *et al.* (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, **20**, 63.

Lun,A.T.L. *et al.* (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 1–14.

Lupo,G. *et al.* (2006) Mechanisms of ventral patterning in the vertebrate nervous system. *Nat. Rev. Neurosci.*, **7**, 103–114.

Van Der Maaten,L. (2014) Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, **15**, 3221–3245.

van der Maaten,L. and Hinton,G. (2008) Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Maier,M.J. (2020) DirichletReg: Dirichlet Regression in R.

Marchetto,M.C. *et al.* (2019) Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *Elife*, **8**, e37527.

Maroof,A.M. *et al.* (2013) Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. *Cell Stem Cell*, **12**, 559–572.

Masaki,H. *et al.* (2015) Interspecific in vitro assay for the chimera-forming ability of human pluripotent stem cells. *Dev.*, **142**, 3222–3230.

Mascetti,V.L. and Pedersen,R.A. (2016) Human-Mouse Chimerism Validates Human Stem Cell Pluripotency. *Cell Stem Cell*, **18**, 67–72.

Matsuda,M., Yamanaka,Y., *et al.* (2020) Recapitulating the human segmentation clock with pluripotent stem cells. *Nature*, **580**, 124–129.

Matsuda,M., Hayashi,H., *et al.* (2020) Species-specific segmentation clock periods are due to differential biochemical reaction speeds. *Science (80-.).*, **369**, 1450–1455.

Matsumiya,M. *et al.* (2018) Es cell-derived presomitic mesoderm-like tissues for analysis of synchronized oscillations in the segmentation clock. *Dev.*, **145**.

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

Miller,J.A. *et al.* (2014) Transcriptional landscape of the prenatal human brain. *Nature*, **508**, 199–206.

Miyazawa,H. and Aulehla,A. (2018) Revisiting the role of metabolism during development. *Dev.*, **145**.

Nicholas,C.R. *et al.* (2013) Functional maturation of hPSC-derived forebrain interneurons requires an extended timeline and mimics human neural development. *Cell Stem Cell*, **12**, 573–586.

Nishimura,D. (2001) BioCarta. *Biotech Softw. Internet Rep.*, **2**, 117–120.

Pauklin,S. and Vallier,L. (2013) The cell-cycle state of stem cells determines cell fate propensity.

Cell, **155**, 135.

Perovanovic,J. *et al.* (2020) Oct1 recruits the histone lysine demethylase Utx to canalize lineage specification. *bioRxiv*, 2020.12.01.406488.

Placzek,M. and Furley,A. (1996) Neural development: Patterning cascades in the neural tube. *Curr. Biol.*, **6**, 526–529.

Polioudakis,D. *et al.* (2019) A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron*, **103**, 785-801.e8.

Pollen,A.A. *et al.* (2019) Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell*, **176**, 743-756.e17.

Powell,J.L. (1986) Censored regression quantiles. *J. Econom.*, **32**, 143–155.

Powell,J.L. (1984) Least absolute deviations estimation for the censored regression model. *J. Econom.*, **25**, 303–325.

Qi,C. *et al.* (2010) Riemannian BFGS algorithm with applications. *Recent Adv. Optim. its Appl. Eng.*, 183–192.

Qiu,X. *et al.* (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14**, 979–982.

Rayon,T. *et al.* (2020) Species-specific pace of development is associated with differences in protein stability. *Science (80-.).*, **369**, eaba7667.

Risso,D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284.

RNA FANTOM brain region gene data *Hum. Protein Atlas*.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

Roccio,M. *et al.* (2013) Predicting stem cell fate changes by differential cell cycle progression patterns. *Dev.*, **140**, 459–470.

Saha,K. and Jaenisch,R. (2009) Technical Challenges in Using Human Induced Pluripotent Stem Cells to Model Disease. *Cell Stem Cell*, **5**, 584–595.

Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.

Savage,V.M. *et al.* (2007) Scaling of number, size, and metabolic rate of cells with body size in mammals. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 4718–4723.

Schoenherr,C. and Anderson,D. (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science (80-.).*, **267**, 1360–1363.

Sebastiano,V. *et al.* (2010) Oct1 regulates trophoblast development during early mouse embryogenesis. *Development*, **137**, 3551–3560.

Shen,Q. *et al.* (2006) The timing of cortical neurogenesis is encoded within lineages of individual progenitor cells. *Nat. Neurosci.*, **9**, 743–751.

Shen,Z. *et al.* (2017) Enforcement of developmental lineage specificity by transcription factor Oct1. *Elife*, **6**.

Shi,Y. *et al.* (2012) Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat. Neurosci.*, **15**, 477–486.

Sjöstedt,E. *et al.* (2020) An atlas of the protein-coding genes in the human, pig, and mouse brain.

Science (80-.)., **367**, eaay5947.

Stephens,M. (2017) False discovery rates: A new deal. *Biostatistics*, **18**, 275–294.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Sun,N. *et al.* (2017) Inference of differentiation time for single cell transcriptomes using cell population reference data. *Nat. Commun.*, **8**, 1–12.

Sunkin,S.M. *et al.* (2013) Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.*, **41**.

Svensson,V. (2020) Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, **38**, 147–150.

Temple,S. and Raff,M.C. (1986) Clonal analysis of oligodendrocyte development in culture: Evidence for a developmental clock that counts cell divisions. *Cell*, **44**, 773–779.

Tesar,P.J. *et al.* (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*, **448**, 196–199.

Thompson,R. and Chan,C. (2018) NRSF and Its Epigenetic Effectors: New Treatments for Neurological Disease. *Brain Sci.*, **8**, 226.

Townes,F.W. *et al.* (2019) Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.*, **20**, 1–16.

Townes,F.W. (2019) Generalized Principal Component Analysis. 1–7.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

Tsoucas,D. and Yuan,G.-C. (2018) GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. *Genome Biol.*, **19**, 58.

Tung,P.Y. *et al.* (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, **7**, 1–15.

Väremo,L. *et al.* (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.*, **41**, 4378–4391.

Wang,X. *et al.* (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**.

Wang,X. *et al.* (2012) NRSF/REST is required for gastrulation and neurogenesis during zebrafish development. *Acta Biochim. Biophys. Sin. (Shanghai).*, **44**, 385–393.

Wheatley,D.N. and Clegg,J.S. (1994) What determines the basal metabolic rate of vertebrate cells in vivo? *Biosystems*, **32**, 83–92.

Wu,A.R. *et al.* (2017) Single-Cell Transcriptional Analysis. *Annu. Rev. Anal. Chem.*, **10**, 439–462.

Wu,J. *et al.* (2017) Interspecies Chimerism with Mammalian Pluripotent Stem Cells. *Cell*, **168**, 473-486.e15.

Yang,Y. *et al.* (2017) Derivation of Pluripotent Stem Cells with In Vivo Embryonic and Extraembryonic Potency. *Cell*, **169**, 243-257.e25.

Yevshin,I. *et al.* (2019) GTRD: A database on gene transcription regulation - 2019 update. *Nucleic Acids Res.*, **47**, D100–D105.

Ying,Q.L. *et al.* (2003) Conversion of embryonic stem cells into neuroectodermal precursors in

adherent monoculture. *Nat. Biotechnol.*, **21**, 183–186.

Yu,N.Y.L. *et al.* (2015) Complementing tissue characterization by integrating transcriptome profiling from the human protein atlas and from the FANTOM5 consortium. *Nucleic Acids Res.*, **43**, 6787–6798.

Zheng,G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 1–12.

Zhihong,Z. *et al.* (2020) MicroRNA-1179 regulates proliferation and chemosensitivity of human ovarian cancer cells by targeting the PTEN-mediated PI3K/AKT signaling pathway. *Arch. Med. Sci.*, **16**, 907–914.