STATISTICAL METHODS FOR RELIABLE INFERENCE IN RNA-SEQ EXPERIMENTS TO FACILITATE REGENERATIVE MEDICINE

by

Ning Leng

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

Doctor of Philosophy
(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 07/25/14

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

Christina Kendziorski, Professor, Department of Biostatistics and Medical Informatics

Ron Stewart, Associate Director of Bioinformatics in Regenerative Biology, Morgridge Institute for Research

Michael Newton, Professor, Department of Statistics and Department of Biostatistics and Medical Informatics

Colin Dewey, Associate Professor, Department of Biostatistics and Medical Informatics

Ming Yuan, Professor, Department of Statistics and Morgridge Institute for Research

Dedicated to my parents, my grandparents and my beloved husband Tan

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisor Professor Christina Kendziorski who has guided me through my Ph.D. studies with her great patience, immense knowledge and kind support. Without her guidance and persistent help this dissertation would not have been possible. Her support, understanding and kindness have been invaluable on both an academic and a personal level. I am also very grateful to my committee member Dr. Ron Stewart. His patient guidance helped me to develop my background and interests in computational biology during our four-year collaboration. I also give deep thanks to the other three committee members: Professor Michael Newton, Professor Colin Dewey and Professor Ming Yuan. Their insightful comments and valuable suggestions largely improved my dissertation work.

I thank to the members in our research group for generously sharing their ideas and suggestions during our group meetings and daily work. I also thank to current and former members in the bioinformatics team at Thomson lab for sharing their rich experiences in computational biology. I am also thankful to all my collaborators at Thomson lab, Gould lab, Attie lab, Pike lab and Sun lab for giving me opportunities to participate in their interesting projects. This cross-disciplinary experience largely inspired me to continue to pursue my career in statistical genomics.

I am so grateful to all my friends for all their kindness and helps. A special thanks goes to Bo Li for his efforts on the RSEM-EBSeq pipeline, to Yuan Li and Xiaomao Li for their hard work on the EBSeq-HMM project and the Oscope project, to Li-Fang Chu for the insightful discussions on the Oscope project and his valuable comments on my thesis writing, and to Qiuling He for her support and advice.

Last but not least, I would like to thank my parents, grandparents, other family members and my husband Tan for their love, support and encouragement.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Abstract

The last decade of genome research has led to major technological advances in sequencing, genotyping, and phenotyping. However, how best to derive useful information from them still remains to be explored by statistical scientists. In this dissertation, I develop, implement, evaluate and apply three statistical methods for high-dimensional data analysis to facilitate efforts in regenerative medicine. A unifying theme of this work is that all these methods utilize RNA-seq data. The first two methods focus on identifying changes of gene or isoform expression across two or more conditions, in experiments with ordered or unordered conditions. The third method focuses on detection of oscillatory gene sets using single-cell RNA-seq data from an unsynchronized cell population.

The first method is an empirical Bayes model called EBSeq for identifying differentially expressed (DE) genes and isoforms. Unlike microarrays, RNA-seq experiments allow for the identification of not only DE genes, but also their corresponding isoforms on a genome-wide scale. Several methods had been developed for identification of the DE genes, but we discovered that they are not applicable for isoform level analysis, because none of them accounts for varying quantification uncertainty among different isoform groups. Taking advantage of the merits of empirical Bayesian methods, we developed EBSeq which models the uncertainty groups via different priors. Our results demonstrate substantially improved power and performance of EBSeq for identifying DE isoforms compared to other competing methods. EBSeq also proves to be a robust approach for identifying DE genes.

The second method is an auto-regressive hidden Markov model called EBSeq-HMM for identifying expression changes across ordered conditions. With improvements in next-generation sequencing technologies and reductions in price, ordered RNA-seq experiments are becoming common. Of primary interest in these experiments is identifying genes that are changing over time or space, for example, and then characterizing the specific expression changes. Several RNA-seq DE methods

have been extended to accommodate experiments with more than two conditions. However, none of them accounts for dependence across conditions and thereby sacrifice power for identifying genes showing subtle, yet consistent, changes. In addition, these methods were not designed to classify DE genes into expression paths. In EBSeq-HMM, an autoregressive hidden Markov model is implemented to accommodate dependence in gene expression across ordered conditions. As demonstrated in simulation and case studies, the output proves useful in identifying DE genes, characterizing their changes over conditions, and classifying genes into particular expression paths. EBSeq-HMM may also be used for inference on isoform expression.

The third method is a statistical pipeline called Oscope for identifying oscillatory gene sets using unsynchronized single-cell RNA-seq data. Recent advance of single-cell RNA-seq enables precise quantification of gene expression among individual cells. This provides the potential to uncover oscillatory systems at single-cell level. However, methods to identify candidate oscillatory gene sets in an unsynchronized cell population are still lacking. Here we developed a statistical pipeline with 3 main modules - a paired-sine model to identify co-oscillating gene paires, a K-Medoid clustering module to group gene pairs into oscillatory gene sets, and an extended nearest insertion algorithm to recover base cycle profile of oscillatory genes. Simulation and case study results demonstrate that Oscope is powerful in identifying oscillatory gene sets, recovering base cycle expression profile of oscillators, and in estimating phase shifts among different oscillatory genes.

1 INTRODUCTION

Next generation sequencing (NGS) technologies are having a major impact on a broad range of biological endeavors. RNA-seq refers to the NGS of cDNA to quantify the mRNA transcripts present in a sample, and is now considered a revolutionary tool for transcriptomics (Wang et al., 2009b; Ozsolak and Milos, 2011). With the rapid decrease in the cost and the advancement of sequencing technologies, the debate is not about *whether* RNA-seq will replace microarrays, but *when* . This is due to the many compelling advantages of RNA-seq. Unlike microarrays, RNA-seq has a very wide dynamic range and low background noise; doesn't require probe design before the experiment; allows for discovery of novel splicing events, exons, genes and gene fusion events; and provides more precise expression estimations at the gene, isoform and exon level by its single-base resolution(Wang et al., 2009b).

Although many computational and statistical methods have been developed for analyzing RNA-seq data, the best practices are not yet clearly established. Some fields are still in a developmental stage requiring new methods. Here we propose three novel statistical approaches in identifying differential expressed genes and isoforms across two or more conditions, identifying genes and isoforms with expression changes in a time or spatial course experiment, and detecting oscillatory gene sets in an unsynchronized single-cell population using RNA-seq data.

The first part of my work was motivated largely by studies of pluripotent stem cells, as these cells have the potential to revolutionize medicine by facilitating new models for drug testing and discovery. They may also, someday, provide treatment options for a wide range of conditions and diseases that currently lack therapies. However, a number of challenges must be addressed before this potential becomes reality.

A seminal challenge concerns characterizing the extent to which different types of pluripotent stem cells vary. In 1998, The James Thomson Lab isolated the human embryonic stem cells (ESCs), which was a major breakthrough that led to all sorts of discoveries. In spite of the potential of human ESCs research, a number of efforts have hampered their widespread use due to ethical considerations and

potential immune rejection. Induced pluripotent stem cells (iPSCs) are similar to ESCs in many ways and do not carry the ethical concerns and immune rejection. Although similar, it has become clear that these two pluripotent cell types display significant differences. Identifying differences between ESCs and iPSCs is a first step in determining when, and how, the latter may be used in basic science and regenerative medicine.

Most previous studies in comparing ESCs with iPSCs were focused on identifying differentially expressed (DE) genes. However, a recent study (Wu et al., 2010) showed that splicing isoform diversity is much higher in ESCs compared to cells at other stages of differentiation, indicating that important information might be ignored if only considering DE genes, but not isoforms, in studies of pluripotent stem cells. To this end, we aimed to develop statistical methods that allows for a sensitive and specific characterization of the differences between ESCs and iPSCs using RNA-seq data.

Our method concerns the identification of isoforms that are differentially expressed across two or more groups, with a focous on differences between ESCs and iPSCs. A number of methods exist for identifying DE genes, but far fewer are available for identifying DE isoforms. When isoform DE is of interest, investigators often apply gene-level (count-based) methods directly to estimates of isoform counts. We discovered that doing so is problematic. In short, estimating isoform expression is relatively straightforward for some groups of isoforms, but more challenging for others. This results in estimation uncertainty that varies across isoform groups. Count-based methods were not designed to accommodate this varying uncertainty and consequently application of them for isoform inference results in reduced power for some classes of isoforms and increased false discoveries for others. Therefore, we developed an empirical Bayes method called EBSeq to identify DE isoforms as well as genes in an RNA-seq experiment. EBSeq accommodates estimation uncertainty by assigning different priors for isoforms from distinct uncertainty groups. More details of the approach, simulation results and case study results can be found at Chapter 2.

In addition to studies aimed at comparing ESCs vs. iPSCs, another important

problem in the development of regenerative medicine is understanding the connection between gene expression profiles and positional identity in adult mammal tissues. Throughout adulthood, humans and other mammals possess a very limited ability to regenerate body parts, like limb structures. In contrast, salamanders such as the axolotl can fully regenerate various body parts during their entire life. By studying the limb regenerative process in adult axolotl, researchers found that gene regulation plays an important role in positional identities. This critical relationship has also been identified in mammals, but only limited to the embryo stage. Inspiringly, a few recent studies (Rinn et al., 2006; Chang, 2009; Wang et al., 2009a) have demonstrated that embryonic spatial profiles in mammals are further retained into adulthood to maintain functioning tissues and organs. This very fact implies that regenerative potential exists in the adult mammal. However, the previous study was limited to inference of binary, rather than gradient, gene expression changes from proximal to distal ends due to the relatively insensitive measurements using microarrays. Taking advantage of the precise quantification of RNA-seq, we investigated experiments to study gradient expression changes over 7 positions along the limb of adult mouse. We are interested in identifying genes with expression changes along the spatial course, as well as clustering these genes based on their profiles.

As detailed in Chapter 2, several methods have been developed for identification of DE genes and/or isoforms, and may be used to detect DE genes in a multiple condition experiment (Anders and Huber, 2010; Robinson et al., 2010; Trapnell et al., 2012a). However, they are not directly applicable to our problem here. First, these methods identify a gene as differentially expressed if it shows significant change in at least one condition, but they do not distinguish among different types of changes (expression is increased in a single condition vs. monotonically increasing vs. increasing then decreasing, for example). Second, current methods do not accommodate dependence across conditions and thereby sacrifice power for identifying genes showing subtle, yet consistent, changes. Therefore, we developed EBSeq-HMM which implements an auto-regressive hidden Markov model to accommodate dependence in gene expression across ordered conditions. More

details of the approach, simulation results and case study results can be found at Chapter 3.

Another important issue of interest is studying expression profiles of oscillatory genes at the single-cell level. In most living organisms, gene oscillation systems such as cell cycle and circadian clock play important roles in regulating growth, metabolic activity, cell division and replication, etc. They are important during development as well as adulthood. Understanding the oscillatory systems is an essential, and very important step in differentiating ESCs or iPSCs towards target cell types. On studying mRNA expression profiles of oscillatory systems, previous work falls into three categories - transcriptome-wide studies on average signals of synchronized cells using mRNA microarrays or RNA-seq (Whitfield et al., 2002; Marcolino-Gomes et al., 2014), studies of selected genes on synchronized cell population using qt-PCR or live-cell imaging, and single-cell level studies on several pre-selected genes using live-cell imaging (Lionnet et al., 2011). Although many important findings have been obtained via these three approaches, each approach has its own limitations. The former two approaches take average gene expression from populations with millions of cells, which ignores differences between individual cells. In addition, the synchronization step is typically designed for one known oscillatory system and may disturb rhythms of any other. Thus it doesn't allow for de novo system findings. On the other side, the latter two approaches are limited to a few genes with prior knowledge, which is neither able to efficiently identify de novo oscillators nor to infer large gene networks.

Currently, no method is available to identify candidate oscillatory genes using single-cell RNA-seq data. Therefore, we developed a statistical pipeline called Oscope (oscilloscope) for identifying oscillatory gene sets in an unsynchronized cell population. The Oscope pipeline incorporates three modules. The three modules are designed to identify oscillatory genes, to group them into oscillatory gene sets, and to recover the base cycle expression profiles of the genes in each group by reordering the cells, respectively. More details of the approach, simulation results and case study results are included in Chapter 4.

2 EBSEQ: AN EMPIRICAL BAYES MODEL FOR IDENTIFYING DE GENES AND ISOFORMS

2.1 Background

Appropriate expression of a gene's isoforms via alternative splicing is fundamental to normal development and maintenance in eukaryotes; and aberrations in alternative splicing are common in disease (Wang et al., 2008; Stamm et al., 2005; Smith et al., 1989). Consequently, there is much interest in identifying isoforms with expression that varies on average across biological conditions. RNA-seq experiments provide the potential to identify such DE isoforms on a genome-wide scale, but statistical methods are required to ensure that accurate identifications are made.

The statistical methods available for identifying *differences* in isoforms in an RNA-seq experiment (e.g. MISO) have focused on changes in the proportion of gene-specific reads assigned to an isoform (Katz et al., 2010; Singh et al., 2011), so-called differential transcription or differential splicing. These methods do not consider changes in overall expression levels and are therefore not appropriate for identifying DE isoforms. At the same time, several methods already have been developed for identifying DE genes in an RNA-seq experiment (e.g. baySeq (Hardcastle and Kelly, 2010); DESeq (Anders and Huber, 2010); edgeR (Robinson and Smyth, 2008)), but very few accommodate isoforms.

When isoform DE is of interest, investigators often apply gene-level methods directly to estimates of isoform counts. However, most of the gene level methods are not appropriate for isoform-based inference. This is largely due to the fact that isoform expression is difficult to estimate, which results in isoform expression estimation uncertainty that varies systematically for different groups of isoforms. More specifically, estimating expression of an isoform unique to its parent gene is a relatively straightforward problem, but for genes with multiple isoforms, the problem is more challenging as reads mapping to overlapping exons must be allocated to isoforms in a way that is consistent with their expression (see Figure 2.1).

Consequently, there is increased uncertainty (on average) in expression estimates for isoforms with multiple overlapping exons, referred to hereinafter as complex isoforms. This uncertainty effects downstream analysis methods, such as methods for identifying DE genes.

Most approaches for identifying DE genes in an RNA-seq experiment focus exclusively on genes and consequently they require counts of reads that uniquely map to genes, as opposed to *estimates* of counts mapping to isoforms. When isoform DE is of interest, some count-based methods (e.g. edgeR) suggest choosing a single isoform (such as the isoform with the most counts within a gene (Robinson et al., 2014) or the longest isoform (Sandmann et al., 2011)) and estimating expression using reads mapping to the isoforms' constituent exons. In either case, information on other isoforms is lost, and reads mapped to multiple genes are ignored. A more serious consideration is that erroneous conclusions may be made due to differences in other isoforms (see Figure 2.1 (b) for an example). Other methods such as easyRNASeq (Delhomme and Padioleau, 2012) suggest that one assign all reads mapping to overlapping exons to each isoform separately (i.e. count reads mapping to exon 2 in Figure 2.1 (b) twice, once for each isoform), and then proceed with a count-based approach. As with the prior suggestion, this can lead to erroneous conclusions. Specifically, an isoform may appear to be EE (say isoform 1 in Figure 2.1 (b)), even if it is not.

A potentially more robust way to proceed is to estimate each isoform's mRNA counts using a method designed specifically to do so (such as in Li and Dewey (2011); Jiang and Wing (2009); Nicolae et al. (2010); Trapnell et al. (2012b)) and then apply a count-based approach directly to *expected* counts after rounding the expected counts to the nearest integer. What we discovered in the course of this work is that doing so is problematic. Count-based methods require gene-level counts and consequently do not account for uncertainty inherent in *estimated* counts. Furthermore, given that uncertainty varies systematically for different groups of isoforms, applications to isoform level inference results in reduced power for some classes of isoforms and increased false discoveries for others. In short, the test statistics used by most methods for DE gene identification calibrate a difference in

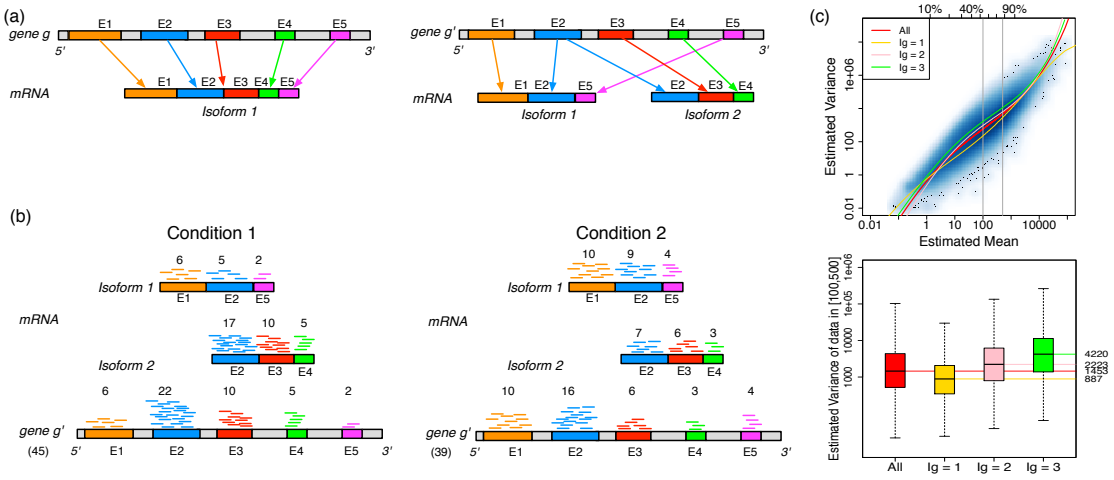


Figure 2.1: Panel (a) shows two hypothetical genes g and g′. Gene g has one isoform, denoted by $I_g = 1$; gene g′ has two ($I_{g'} = 2$). The problem of estimating expression for isoforms of g′ is complicated by the fact that reads mapping to exon 2 must be unambiguously assigned to each isoform. This results in increased uncertainty, on average, in expression estimates for isoforms sharing a parent. Panel (b) shows hypothetical expression of the isoforms from gene g′ in each of two conditions (assuming differences in library size have been accommodated). If one focuses on the longest isoform (isoform 1) and uses all reads mapping to its constituent isoforms to estimate its expression, the isoform is called equivalently expressed since there are 30 (6+22+2) reads mapped in condition 1 and 30 (10+16+4) mapped in condition 2. However, if the expression of other isoforms is considered, it becomes clear that isoform 1 contains almost twice as many reads in condition 2 as in condition 1 (23 vs. 13, respectively). Panel (c) demonstrates how estimation uncertainty changes as isoform complexity increases. We quantified isoform complexity here by I_g where the $I_g = k$ group represents isoforms from genes with k isoforms (here $I_g > 3$ isoforms are included in the $I_g = 3$ group; alternative definitions of complexity are discussed in the text). Shown top right are splines fit to the empirical variance as a function of the mean for all isoforms (red) as well as isoforms within groups defined by I_g for the two-group human embryonic stem cell RNA-seq experiment(details of the data set could be found in Appendix); bottom right considers isoforms with average expression (expected count) in [100, 500]. The range was chosen as it approximates the 50^{th} and 80^{th} percentiles of expression across all isoforms. Shown are box-plots of the variances of these isoforms collectively, and within I_g group. Median variance within each group is shown right.

expression levels between conditions by a variance, which is commonly estimated by fitting a spline to the mean-variance relationship observed in data. Figure 2.1 (c) shows that this relationship varies dramatically for different groups of isoforms, where groups are defined by the number of constituent isoforms of the parent gene (other definitions are possible as discussed below). Specifically, an isoform of gene g is assigned to the $I_g = k$ group, for example, where $k = 1, 2$ or 3, if the total number of isoforms from gene g is k (the $I_g = 3$ group contains all isoforms from genes having 3 or more isoforms).

As shown in Figure 2.1 (c), there is decreased variability in the $I_g = 1$ group, but increased variability in the others, due to the relative increase in uncertainty inherent in estimating isoform expression when multiple isoforms of a given gene are present. This observation is not specific to the data set and/or the method used for isoform expression estimation; it is also not specific to the particular method used for quantifying isoform complexity (see Appendix for additional examples). If isoforms are analyzed collectively, there is reduced power for identifying isoforms in the $I_g = 1$ group (since the true variances in that group are lower, on average, than that derived from the full collection of isoforms) and increased false discoveries in the $I_g = 2$ and $I_g = 3$ groups (since the true variances are higher, on average, than those derived from the full collection).

Cuffdiff (Trapnell et al., 2012b) and BitSeq (Glaus et al., 2012) are the only methods currently available that accommodates DE inference at the isoform level. However, the applicability of Cuffdiff and BitSeq are limited to cases where expression estimates are obtained via Cufflinks or via BitSeq itself. Therefore they are not suitable when investigators are using quantification methods other than Cufflinks or BitSeq. We also found that Cuffdiff is underpowered on data with biological replicates in both simulation and case study data sets. BitSeq is not evaluated here because it provides probability of positive log ratio instead of posterior probability of being DE, which is useful for ranking but with no FDR control.

Consequently, we developed an empirical Bayesian approach called EBSeq that accommodates varying levels of isoform expression estimation uncertainty and thereby facilitates improved inference in RNA-seq experiments at both the isoform

and gene level. Our method requires gene counts or estimates of isoform expression, but it is not specific to any particular quantification method (e.g. RSEM (Li and Dewey, 2011), Rseq (Jiang and Wing, 2009), Cufflinks (Trapnell et al., 2012b), or another method may be used).

2.2 Methods

EBSeq: An empirical Bayes model for identifying DE genes and isoforms

The general model is developed for isoform analysis. The gene-level model is a special case discussed at the end of this section. The model assumes the expected count for isoform i in gene g and sample s is distributed as Negative Binomial, $X_{g_i,s}$, where $g = 1, 2, ..., G$, $s = 1, 2, ..., S$, and $i = 1, 2, ..., N_g$; N_g denotes the number of isoforms of gene g.

We let $X_{g_i}^{C1} = X_{g_i,1}, X_{g_i,2}, ..., X_{g_i,S_1}$ denote data from Condition 1 and $X_{g_i}^{C2} = X_{g_i,(S_1+1)}, X_{g_i,(S_1+2)}, ..., X_{g_i,S}$ data from Condition 2. We assume that counts within condition C are distributed as Negative Binomial: $X_{g_i,s}^{C}|r_{g_i,s}, q_{g_i}^{C} \sim NB(r_{g_i,s}, q_{g_i}^{C})$ where

$$P(X_{g_i,s}|r_{g_i,s}, q_{g_i}^{C}) = \binom{X_{g_i,s} + r_{g_i,s} - 1}{X_{g_i,s}} (1 - q_{g_i}^{C})^{X_{g_i,s}} (q_{g_i}^{C})^{r_{g_i,s}} \tag{2.1}$$

and $\mu_{g_i,s}^{C} = r_{g_i,s}(1 - q_{g_i}^{C})/q_{g_i}^{C}$; $(\sigma_{g_i,s}^{C})^2 = r_{g_i,s}(1 - q_{g_i}^{C})/(q_{g_i}^{C})^2$.

We assume a prior distribution on $q_{g_i}^{C}$:$q_{g_i}^{C}|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$. The hyperparameter α is shared across isoforms while β depends on I_g, accommodating the systematic differences in variability among the I_g groups. I_g quantifies a measure of isoform complexity and may be defined by the user as the number of isoforms from a gene, as described in the previous section, or from an isoform's mappability score or credibility interval as provided by some isoform expression estimation approaches. We further assume that $r_{g_i,s} = r_{g_i,0} l_s$ where l_s represents the library size in sample s. $r_{g_i,0}$ is an isoform specific parameter common across conditions. Of interest is distinguishing between EE and DE (two expression patterns) where

H_0 (EE) : $q_{g_i}^{C1} = q_{g_i}^{C2}$ vs H_1 (DE) : $q_{g_i}^{C1} \neq q_{g_i}^{C2}$.

On the null hypothesis (EE), the data $X_{g_i}^{C1,C2} = X_{g_i}^{C1}, X_{g_i}^{C2}$ arises from the prior predictive distribution $f_0^{I_g}(X_{g_i}^{C1,C2})$:

$$f_0^{I_g}(X_{g_i}^{C1,C2}) = \left[\prod_{s=1}^{S} \binom{X_{g_i,s} + r_{g_i,s} - 1}{X_{g_i,s}}\right] \frac{Beta(\alpha + \sum_{s=1}^{S} r_{g_i,s}, \beta^{I_g} + \sum_{s=1}^{S} X_{g_i,s})}{Beta(\alpha, \beta^{I_g})} \tag{2.2}$$

Alternatively (DE), $X_{g_i}^{C1,C2}$ follows the prior predictive distribution $f_1^{I_g}(X_{g_i}^{C1,C2})$:

$$f_1^{I_g}(X_{g_i}^{C1,C2}) = f_0^{I_g}(X_{g_i}^{C1}) f_0^{I_g}(X_{g_i}^{C2}) \tag{2.3}$$

Denoting the latent variable Z_{g_i} where $Z_{g_i} = 1$ indicates that isoform g_i is DE and $Z_{g_i} = 0$ indicates isoform g_i is EE. $Z_{g_i} \sim Bernoulli(p)$. Thus, the marginal distribution of $X_{g_i}^{C1,C2}$ and Z_{g_i} is:

$$(1-p) f_0^{I_g}(X_{g_i}^{C1,C2}) + p f_1^{I_g}(X_{g_i}^{C1,C2}) \tag{2.4}$$

The posterior probability of being DE at isoform g_i is obtained by Bayes' rule:

$$\frac{p f_1^{I_g}(X_{g_i}^{C1,C2})}{(1-p) f_0^{I_g}(X_{g_i}^{C1,C2}) + p f_1^{I_g}(X_{g_i}^{C1,C2})} \tag{2.5}$$

Parameter estimation

With the assumption that $r_{g_i,s} = r_{g_i,0} l_s$, denote $\mu_{g_i,0}^C$ and $(\sigma_{g_i,0}^C)^2$ are the mean and variance of gene g isoform i under standard library size. Then $\mu_{g_i,0}^C = \frac{1}{l_s}\mu_{g_i,s}^C$ for any s within condition C, Assume there are S_C samples in condition C. We could obtain the unbiased estimator $\hat{\mu}_{g_i,0}^C = \frac{1}{S_C}\sum_{s \text{ in } C} \frac{1}{l_s}\hat{\mu}_{g_i,s}^C$. Where $\hat{\mu}_{g_i,s}^C = X_{g_i,s}^C$.

Since $(\sigma_{g_i,0}^C)^2 = \frac{1}{l_s}(\sigma_{g_i,s}^C)^2$ for any s within condition C, we could obtain the estimator $(\hat{\sigma}_{g_i,0}^C)^2 = \frac{1}{S_C}\sum_{s \text{ in } C} \frac{1}{l_s}(\hat{\sigma}_{g_i,s}^C)^2$, which is unbiased conditioning on $\mu_{g_i,0} = \hat{\mu}_{g_i,0}$ where $(\hat{\sigma}_{g_i,s}^C)^2 = (X_{g_i,s}^C - l_s\hat{\mu}_{g_i,0}^C)^2$.

Denote $\hat{\mu}_{g_i,0} = \frac{\hat{\mu}_{g_i,0}^{C1} + \hat{\mu}_{g_i,0}^{C2}}{2}$ and $\hat{\sigma}_{g_i,0}^2 = \frac{(\hat{\sigma}_{g_i,0}^{C1})^2 + (\hat{\sigma}_{g_i,0}^{C2})^2}{2}$ Then the estimator of $r_{g_i,0}$ is

obtained by $\hat{r}_{g_i,0} = \frac{\hat{\mu}^2_{g_i,0}}{\hat{\sigma}^2_{g_i,0} - \hat{\mu}_{g_i,0}}$.

$\hat{l_s}$ could be obtained by total number of reads, TMM (Robinson and Oshlack, 2010), Median Normalization (Anders and Huber, 2010), or Quantile Normalization (Bullard et al., 2010). Since total number of reads may be adversely affected by outliers from PCR or other artifacts, the latter 3 methods are more acceptable. We used Median Normalization.

The EM algorithm is used to estimate the α, β^{I_g} and p via the **optim** function in **R**.

Working with multiple conditions

EBSeq naturally accommodates multiple condition comparisons. For example, in a study with 3 conditions, there are 5 possible patterns in which latent levels of expression may vary across conditions:

$$
\begin{aligned}
&q^{C1}_{g_i} = q^{C2}_{g_i} = q^{C3}_{g_i};\\
&q^{C1}_{g_i} = q^{C2}_{g_i} \neq q^{C3}_{g_i};\\
&q^{C1}_{g_i} = q^{C3}_{g_i} \neq q^{C2}_{g_i};\\
&q^{C1}_{g_i} \neq q^{C2}_{g_i} = q^{C3}_{g_i};\\
&q^{C1}_{g_i} \neq q^{C2}_{g_i} \neq q^{C3}_{g_i}.
\end{aligned}
$$

The prior predictive distributions for these are given, respectively, by:

$$
\begin{aligned}
&g_1^{I_g}(X^{C1,C2,C3}_{g_i}) = f_0^{I_g}(X^{C1,C2,C3}_{g_i});\\
&g_2^{I_g}(X^{C1,C2,C3}_{g_i}) = f_0^{I_g}(X^{C1,C2}_{g_i}) f_0^{I_g}(X^{C3}_{g_i});\\
&g_3^{I_g}(X^{C1,C2,C3}_{g_i}) = f_0^{I_g}(X^{C1,C3}_{g_i}) f_0^{I_g}(X^{C2}_{g_i});\\
&g_4^{I_g}(X^{C1,C2,C3}_{g_i}) = f_0^{I_g}(X^{C1}_{g_i}) f_0^{I_g}(X^{C2,C3}_{g_i});\\
&g_5^{I_g}(X^{C1,C2,C3}_{g_i}) = f_0^{I_g}(X^{C1}_{g_i}) f_0^{I_g}(X^{C2}_{g_i}) f_0^{I_g}(X^{C3}_{g_i})
\end{aligned}
$$

in which $f_0^{I_g}$ is the same as in equation 2.2. Then the marginal distribution in

equation 2.4 becomes:

$$\sum_{j=1}^{5} p_j g_j^{I_g}(X_{g_i}^{C1,C2,C3}) \tag{2.6}$$

in which $\sum_{j=1}^{5} p_j = 1$.

Thus, the posterior probability that isoform g_i is in pattern P_J is readily obtained by:

$$\frac{p_J g_J^{I_g}(X_{g_i}^{C1,C2,C3})}{\sum_{j=1}^{5} p_j g_j^{I_g}(X_{g_j}^{C1,C2,C3})} \tag{2.7}$$

Estimates of posterior fold change

EBSeq also provides estimates of posterior fold change (posterior FC). Compared to empirical FC, the posterior FC tends to shrink the estimated ratio of low expressed genes. Denote $a_g^{C1} = \alpha + r_g^{C1} * S^{C1}$, and $b_g^{C1} = \beta + \mu_g^{C1} * S^{C1}$, in which S^{C1} denotes the number of samples within C1. Denote $\psi_g^{C1} = a_g^{C1}/(a_g^{C1} + b_g^{C1})$. Then the posterior FC can be written as (here we assume constant library size for simplicity):

$$\begin{aligned} \frac{(1-\psi_g^{C1})/\psi_g^{C1}}{(1-\psi_g^{C2})/\psi_g^{C2}} &= \frac{b_g^{C1}/a_g^{C1}}{b_g^{C2}/a_g^{C2}} \\ &= \frac{b_g^{C1}}{b_g^{C2}} * \frac{a_g^{C2}}{a_g^{C1}} \\ &= \frac{\beta + \mu_g^{C1} * S^{C1}}{\beta + \mu_g^{C2} * S^{C2}} * \frac{\alpha + r_g^{C2} * S^{C2}}{\alpha + r_g^{C1} * S^{C1}} \end{aligned} \tag{2.8}$$

When μ_g^{C1} and μ_g^{C2} are small, the posterior FC will be closer to 1 compared to the empirical FC μ_g^{C1}/μ_g^{C2}. Isoform level estimates can be derived similarly.

2.3 Results

Simulation set-up

To evaluate the model proposed in EBSeq, we followed the simulation setup of (Robinson and Smyth, 2007) by defining counts as Negative Binomial with isoform-specific mean in sample s and condition C given by $l_s \mu^C_{g_i}$ and variance $l_s \mu^C_{g_i}(1 + l_s \mu^C_{g_i} \phi_{g_i})$. The library size factors for both the isoform and gene-level simulations were randomly simulated from Uniform $(0.8, 1.3)$. One hundred simulated data sets were generated for each scenario considered.

Sim I: Isoform expression for each of 30,802 isoforms, four lanes in each of two conditions, is generated by sampling unknown parameters (μ_{g_i}, ϕ_{g_i}) from the case study comparing embryonic stem cells (ESCs) with induced pluripotent stem cells (iPSCs). The number of isoforms and sample sizes are taken to match those in the case study. The percentages of DE isoforms were set at 2%, 4% and 5% in the $I_g = 1, 2$ and 3 groups, also to match the case study data. Parameters for isoforms belonging to the same gene are sampled together to preserve dependence within isoforms common to a single gene. For DE isoforms, $\mu^{C1}_{g_i} = \mu^{C1}_{g_i} * \delta_{g_i}$ where δ_{g_i} is sampled from the 95%-97% quantile of fold changes in sample means across conditions; for EE isoforms $\mu^{C1}_{g_i} = \mu^{C2}_{g_i}$.

Sim II: Isoform expression for each of 30,802 isoforms is generated by sampling (μ_{g_i}) from case study data; ϕ_{g_i} is fixed for all g_i. Six sets of Sim II are considered to investigate the effects of systematic changes in variability, one set for each ϕ in $(5 * 10^{-4}, 1 * 10^{-3}, 5 * 10^{-3}, 1 * 10^{-2}, 5 * 10^{-2}, 1 * 10^{-1})$; DE and EE are as in Sim I. This setup is similar to that considered in (Robinson and Smyth, 2007). There, too, μ_{g_i} is sampled and ϕ is fixed, but here we simulate isoforms (not genes) and we consider more (and slightly different values) of ϕ.

Sim III: Gene expression for each of 20,000 genes is generated by sampling unknown parameters (μ_g, ϕ_g) from case study data. 2% DE genes are simulated to match the case study data.

Simulation results

Table 2.1 shows the power and false discovery rate (FDR) for EBSeq and Cuffdiff averaged across 100 Sim I simulations for a target FDR of 5%. Cuffdiff deems some isoforms unacceptable prior to analysis. Acceptable isoforms are called "OK" in Cuffdiff, and so results are reported for all genes as well as those deemed "OK" by Cuffdiff. As shown in Table 2.1, Cuffdiff has well controlled FDR, but reduced power compared with EBSeq (~44% vs. ~72%); the FDR of EBSeq is slightly elevated (~8%). Panel (a) of Figure 2.2 shows qualitatively similar results for lists of varying size, not determined by targeting a specific FDR. In particular, the ROC curves (true positive rate (TPR) vs. false positive rate (FPR) for lists of increasing size) show that the TPR is higher for lists provided by EBSeq for all FPRs considered.

A closer look into the DE calls from the Sim I simulations reveals that operating characteristics are sensitive to ϕ, which determines within-isoform variability. To demonstrate the effects, panel (b) of Figure 2.2 shows power and FDR for six other sets of simulations where ϕ is fixed at a specific value (detailed in Sim II). The solid lines show that the power of both methods decreases as variability (ϕ) increases, with a greater loss in power for Cuffdiff; the dashed lines show that FDR increases slightly but remains well-controlled for both methods. Panel (c) shows the cumulative distribution functions (CDFs) of ϕ in four empirical data sets as well as the average CDF from 100 Sim I simulations to demonstrate that the values of ϕ considered in panel (b) are typical of those observed in data. Panel (c) also demonstrates systematic differences between the three data sets with biological replicates and the MAQC data. Given that the MAQC data is comprised of technical replicates, it is not surprising to observe relatively smaller values of ϕ compared with data sets having biological reps. However, since the operating characteristics of some DE identification methods vary with ϕ (as shown in panel (b)), an evaluation of methods based on MAQC data alone is cautioned.

We also applied count-based DE methods on Sim I simulations. As demonstrated in Figure 2.1, count-based DE methods expected to be underpowered in the $I_g = 1$ group when applied directly to estimates of isoform expression (Shown in

Table 2.1: DE analysis: isoform simulation results

	Power	FDR
Cuffdiff	33.6%	0.2%
Cuffdiff(OK)	44.4%	0.2%
EBSeq	72.2%	8.2%

The power and false discovery rate (FDR) for EBSeq and Cuffdiff averaged across 100 Sim I simulations where target FDR was set at 5%. Cuffdiff deems some isoforms unacceptable prior to analysis. Operating characteristics are reported overall, as well as within those deemed acceptable ("OK") by Cuffdiff. Standard errors on average power (FDR) were less than 2% (0.2%) and are not shown.

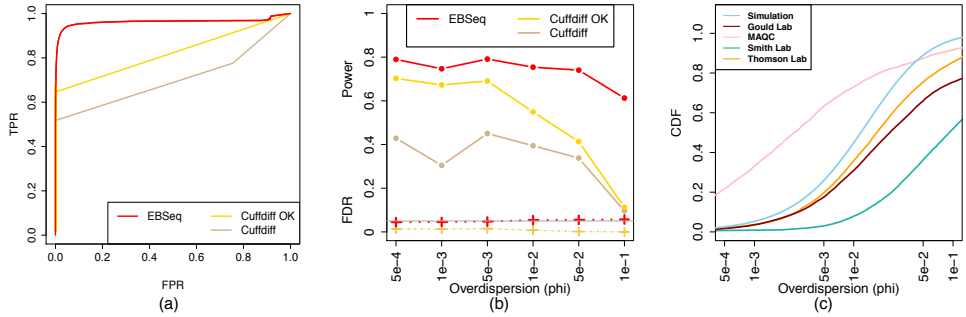


Figure 2.2: Panel (a) shows ROC curves (true positive rate (TPR) vs. false positive rate (FPR)). The curves are obtained from averaging over 100 Sim I simulations. Cuffdiff deems some isoforms unacceptable prior to analysis; isoforms deemed acceptable by Cuffdiff are denoted "OK"; and results are reported here for both. Panel (b) shows the operating characteristics of EBSeq and Cuffdiff as a function of ϕ, described in Sim II. The solid and dashed lines indicate power and FDR, respectively, at 5% target FDR. Panel (c) shows the cumulative distribution function (CDF) of ϕ in four empirical data sets, as well as the CDF averaged across 100 Sim I simulations.

Table 2.2). Each of the count-based methods was also applied to 100 Sim III data sets. Table 2.3 shows the power and FDR of each method at 5% target FDR averaged over the 100 simulated data sets. In short, all methods have well controlled FDR. EBSeq shows the highest power (~79%) with DESeq and edgeR showing comparable performance (~73%); and although baySeq seems to outperform DESeq and edgeR

Table 2.2: DE analysis: applying gene level (count-based) methods on simulated isoform data

		baySeq	DESeq	edgeR	EBSeq
All Isoforms	Power	56.9%	72.3%	81.2%	81.4%
	FDR	0%	0.5%	13.1%	5.0%
$I_g = 1$ Isoforms	Power	**53.3%**	**56.8%**	**60.9%**	**78.2%**
	FDR	0%	0%	0%	0.5%
$I_g = 2$ Isoforms	Power	56.3%	75.2%	84.3%	83.1%
	FDR	0%	1.2%	4.6%	5.7%
$I_g = 3$ Isoforms	Power	59.4%	78.1%	90.3%	81.5%
	FDR	0.5%	0.9%	17.3%	7.0%

Power and FDR averaged across 100 isoform simulations. Thresholds were chosen to control FDR at 5% for each approach. Count-based DE methods are significantly underpowered in the $I_g = 1$ group when applied directly to estimates of isoform expression.

with respect to ranking genes, it has lower power (~61%) than both DESeq and edgeR when FDR is controlled at 5%. Although the FDR of edgeR is well-controlled overall, simulation and case study results suggest that the false calls that are made by edgeR are almost always in genes with outliers (See Appendix Figure A.6 for more details).

Table 2.3: DE analysis: gene simulation results

	Power	FDR
baySeq	60.8%	0.4%
DESeq	73.4%	0%
edgeR	73.1%	4.6%
EBSeq	78.8%	2.7%

The power and false discovery rate (FDR) of baySeq, DESeq, edgeR and EBSeq averaged across 100 Sim III simulations where target FDR was set at 5%. Standard errors on average power (FDR) were less than 2.5% (1.4%) and are not shown.

Case study results

To further evaluate and compare methods, we analyze data from an experiment comparing human ESCs with iPSCs using DESeq, edgeR, and baySeq (with expression counts obtained from HTSeq) as well as Cuffdiff (with expression estimated via Cufflinks). EBSeq is evaluated on both HTSeq and Cufflinks processed data. DESeq, edgeR, baySeq, and Cuffdiff identify 127, 377, 34, and 54 DE genes at 5% FDR, respectively. EBSeq identifies 334 from HTSeq counts and 351 from Cufflinks estimated counts. These results are largely consistent with those observed in the simulation studies.

In particular, panels (a) and (b) of Figure 2.3 show the number of genes found by each approach. There are 114 genes found by DESeq, edgeR and EBSeq via HTSeq processed data; 161 genes found exclusively by EBSeq (neither DESeq nor edgeR find these); and 197 found exclusively by edgeR (neither DESeq nor EBSeq find these). Figure 2.3 shows the boxplots of each gene's 75^{th} percentile of expression across all samples in the groups of 114, 161 and 197. The genes identified exclusively by edgeR have lower expression on average than the other groups while the genes identified exclusively by EBSeq tend to be highly expressed. Figure 2.3 (d) shows the CDFs of each gene's 75^{th} percentile of expression in the groups identified by each of the five methods. Approximately 20% of the genes identified by edgeR are with 75^{th} percentile of expression less than 20.

Results from the gene and isoform comparisons between EBSeq and Cuffdiff via Cufflinks processed data are also consistent with the simulation study, with Cuffdiff identifying far fewer genes and isoforms than EBSeq. Specifically, Cuffdiff identifies 7 isoforms, each of which is identified by EBSeq; but EBSeq also finds additional isoforms to be DE (935 in total). Furthermore, in this case study, isoform-level results obtained from EBSeq are more consistent with gene-level results than those obtained from Cuffdiff. Specifically, there are 12,404 single-isoform genes. For these, we expect isoform and gene-level inference to match (i.e. if the isoform is DE, the gene should also be DE given there is only a single isoform in that gene). Out of the 54 genes identified as DE by Cuffdiff, 39 have single isoforms; only 5

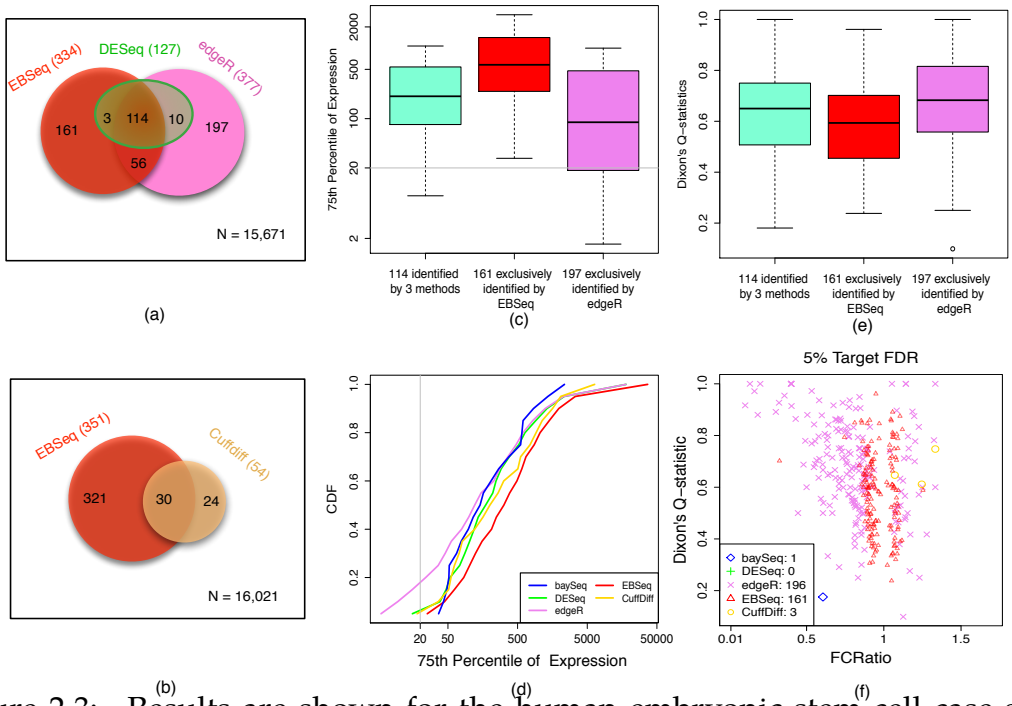


Figure 2.3: Results are shown for the human embryonic stem-cell case study, which compares ESCs with iPSCs.Panel (a) shows a Venn diagram of the genes identified as DE by DESeq, edgeR or EBSeq using HTSeq for quantification. Panel (b) shows a Venn diagram of the genes identified by Cuffdiff and EBSeq using Cufflinks processed data. Panel (c) shows boxplots of each gene's 75^{th} percentile of expression for the three groups of genes - the 114 identified by DESeq, edgeR and EBSeq; the 161 identified by EBSeq but not DESeq or edgeR; and the 196 identified by edgeR but not DESeq or EBSeq. Panel (d) shows the cumulative distribution function (CDF) of the 75^{th} percentile of expression among the 34, 54, 127, 377, and 334 DE genes identified by each method. Panel (e) shows boxplots of Dixon's Q-statistics in three groups of genes defined in panel (c). Panel (f) shows the FCRatios and Dixon's Q-statistics of the genes identified exclusively by each method, but not the other four methods (in this panel, five methods are compared). Note that baySeq, Cuffdiff, DESeq, edgeR and EBseq (via HTSeq) identify 34, 54, 127, 377, and 334 DE genes, respectively, at 5% FDR.

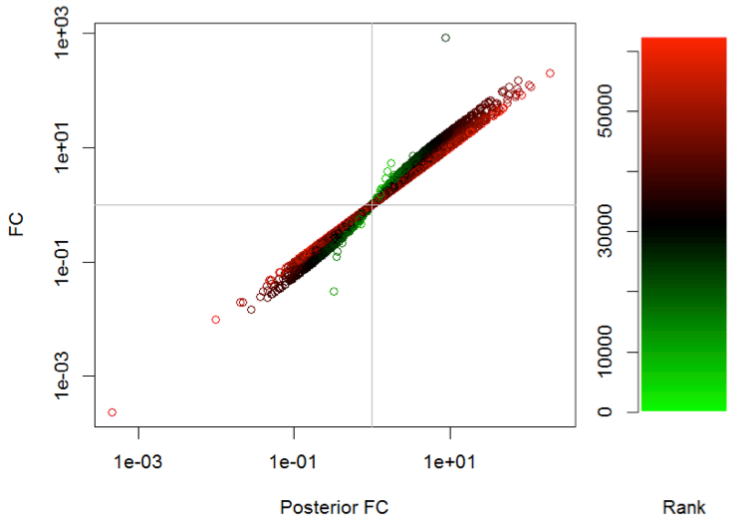


Figure 2.4: Shown are FCs vs. posterior FCs on case study data. Genes are ranked by their cross-condition mean (adjusted by the library size factors). High/low expressors are colored with red/green, respectively.

out of the 39 are also called DE on isoform level. Out of the 351 genes identified as DE by EBSeq, 226 have single isoforms and 225 out of the 226 are also called DE on the isoform level. Furthermore, many important genes confirmed to be DE between ESCs and iPSCs in previous studies (Phanstiel et al., 2011; Ohi et al., 2011; Bock et al., 2011) are missed by Cuffdiff but not EBSeq, including DPP6, FAM19A5, SOX17 and DNAJC15.

Figure 2.4 shows the FC vs. posterior FC on gene expression estimates. The genes are ranked by their cross-condition mean. Recall that the posterior FC estimated by EBSeq tends to shrink the low expressers (genes with small rank). In Figure 2.4, genes with small cross-condition mean are colored as green. For most of genes, the posterior FC shows concordance to the empirical FC. Looking into more details of the low expressed genes with extreme empirical FC, their posterior FCs are shrunk by > 10 fold compared to the empirical FC.

Outlier analysis using EBSeq multiple condition model

We found that many of the genes identified by edgeR but not EBSeq are due to outliers. To identify putative outliers in the case studies, for each gene we evaluated Dixon's Q-statistic (Dixon, 1950) as well as the fold change ratio (FCRatio). A Dixon's Q-statistic for a collection of values is defined as the gap over range, where gap is the absolute difference between an outlier in question and the number closest to it; the range is the max minus min. For each gene in each condition, we calculated the Dixon's Q-statistics for the smallest and the largest value. The sample with the largest Dixon's Q-statistic was defined as the potential outlier for that gene; and the largest Dixon's Q-statistic (over the two conditions) was taken as the Dixon's Q-statistic for the gene. The FCRatio is the ratio of the fold change without the outlier over the fold change with the outlier. A gene containing an outlier will have a Dixon's Q-statistic near 1 and FCRatio far from 1.

Figure 2.3 (e) shows boxplots of Dixon's Q-statistics for these 114, 161 and 197 genes. As shown, the genes exclusively identified by edgeR tend to have higher Dixon's Q-statistics, and are therefore more likely to contain outliers. Of course a gene may contain an outlier and still be DE. To assess this possibility, Figure 2.3 (f) considers how a gene's fold change changes when its most extreme value is removed, as quantified by the FCRatio. If a gene's most extreme value is not largely responsible for the DE call, fold changes with and without the value will remain largely unchanged, and FCRatio will be near one. As shown, edgeR tends to favor genes with FCRatios far from 1, suggesting that the genes identified may be due to a single outlier in an otherwise EE gene.

The experiment comparing ESCs vs iPSCs has been further repeated twice. We used the combined data of 3 experiments (triplicates for each of 8 cell lines) to identify genes that are aberrant in each cell line. We applied the multiple condition EBSeq model on the combined data set with 24 samples to evaluate the posterior probability of each of 18 patterns of expression. The patterns included EE and DE, as in the two condition comparison (panel (a) in Figure 2.5), as well as EE with an outlier cell line (EEO; 8 possibilities since the outlier could be in any one of the 8

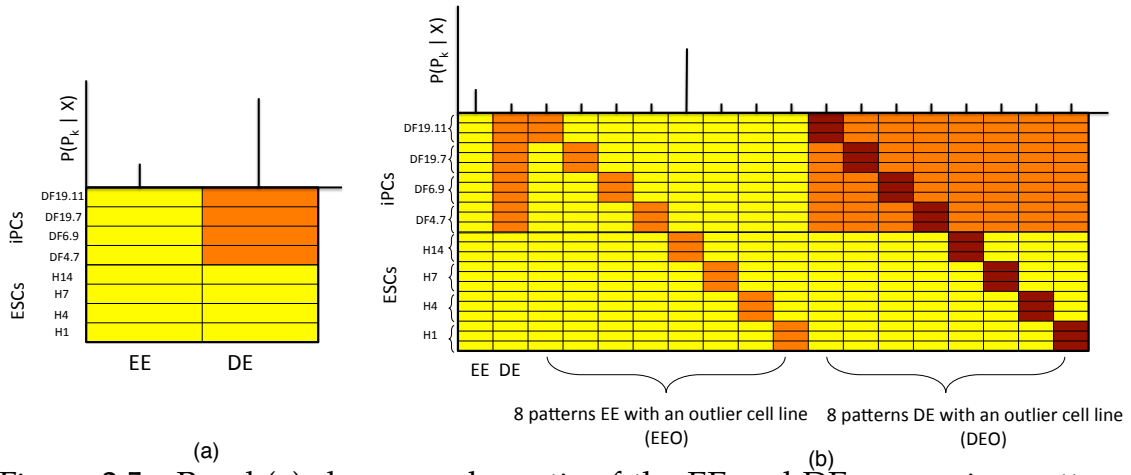


Figure 2.5: Panel (a) shows a schematic of the EE and DE expression patterns considered in a two condition EBSeq model for the experiment comparing ESCs with iPSCs. The upper part shows the posterior probability (PP) of each pattern for a hypothetical gene; these probabilities classify the gene into one of the patterns. Panel (b) shows the patterns used in the multiple condition EBSeq model on the combined data set with 24 samples. The patterns included EE and DE as well as 8 patterns for EE with an outlier cell line (EEO) and another 8 for DE with an outlier cell line (DEO).

Table 2.4: DE analysis: outlier test using multiple condition model

Outlier Cell Line	H1	H7	H9	H14	DF4.7	DF6.9	DF19.7	DF19.11
EEO	98	65	296	135	539	63	38	59
DEO	142	182	203	105	308	51	29	16

The table shows the number of genes identified as DE or EE with an outlier cell line by EBSeq multiple condition test in the experiment comparing ESCs with iPSCs.

cell lines) and DE with an outlier cell line (DEO; another 8 possibilities); see panel (b) in Figure 2.5.

Table 2.4 shows the number of genes falling into each of 16 EEO and DEO patterns with posterior probability greater than 0.95. Results indicate that H9 (DF4.7) is the most different one among ES (iPS) cell lines according to the number of outlier genes.

2.4 Implementation

We've implemented EBSeq as an R package. For those not familiar with R, a graphical user interface (GUI) and a Galaxy-based version of EBSeq are also available. R/EBSeq is freely available on Bioconductor; its source file and vignette may be found at

`www.bioconductor.org/packages/devel/bioc/html/EBSeq.html`

The GUI and galaxy wrapper are available at the EBSeq website:

`http://www.biostat.wisc.edu/~kendzior/EBSEQ/`

Expression estimates: EBSeq requires a (# genes)-by-(# samples) expression matrix for gene analysis or a (# isoforms)-by-(# samples) expression matrix for isoform analysis. It is not specific to any particular quantification method. It accepts gene counts, estimates of gene expression, or estimates of isoform expression. EBSeq takes unnormalized expressions and provides implementations of normalization methods developed in Anders and Huber (2010) and Bullard et al. (2010). Normalization factor estimated by other softwares is also acceptable.

Uncertainty group definitions: EBSeq accommodates varying levels of expression estimation uncertainty in isoform level data. Consequently, for an isoform-level analysis, EBSeq requires that isoform uncertainty groups be specified. A simple but effective grouping is determined by the number of isoforms transcribed from a parent gene, since an isoform transcribed from a gene having multiple isoforms tends to have higher expression estimation uncertainty than an isoform that does not share a parent gene. Specifically, an isoform of gene g may be assigned to the uncertainty group $I_g = k$, for example, where $k = 1$, 2 or 3, if the total number of isoforms from gene g is k (and uncertainty group $k = 3$ contains all isoforms from genes having 3 or more isoforms). This works well for well-annotated genomes, where isoforms and their corresponding parent genes are well defined. However, when the transcriptome of a species of interest is not well annotated, a user may use another measure of isoform complexity to define the uncertainty groups. For example, expression estimation methods often provide isoform-specific measures of uncertainty. RSEM provides an unmappability score; RSEM, Cufflinks and RSeq

provide confidence intervals. Applying clustering algorithm to these statistics, would be a quick way to define uncertainty groups. The RSEM-EBSeq pipeline provides functions and an example using the unmappability scores of RSEM; further details are below.

Output: Unknown parameters in the EBSeq model are estimated using the EM algorithm; Once model parameters are well estimated, the main output of interest consists of a (# genes)-by-(# patterns) or (# isoforms)-by-(# patterns) matrix of PP's which may be used to identify genes or isoforms that are in a particular pattern at a target FDR level. In a two-condition experiment, for example, genes (isoforms) having PP of DE exceeding 0.95 comprise the list with FDR controlled at 5%. Similarly, in a multiple-condition experiment, to obtain a list of genes (isoforms) in a specific DE pattern with target FDR α, a user may identify genes (isoforms) with PP of being in that pattern greater than 1 - α.

RSEM-EBSeq pipeline: An RSEM-EBSeq pipeline is also available for studying known as well as *de novo* assembled transcriptomes. With this pipeline, a user may easily apply RSEM to quantify expression and then EBSeq to identify DE genes and isoforms (or contigs in *de novo* assembled transcriptomes) from the shell command line. To define the I_g vector, the RSEM-EBSeq pipeline takes unmappability scores and applies a K-means algorithm to cluster the isoforms into 3 uncertainty groups. The unmappability scores are also provided as output and, consequently, a user can easily apply a K-means algorithm with different K's or apply another clustering algorithm.

EBSeq GUI: For users not familiar with R, the EBSeq GUI may prove useful. Figure 2.6 (a) shows the interface for a DE analysis across two conditions. The expression matrix (and I_g vector for isoform analysis) can be stored in a .csv, .xls, or .xlsx format and uploaded via the interface. The full utility of EBSeq is available; output is stored in a .csv file which could be easily opened by many applications (e.g. Microsoft Excel, Open Office Calc). The EBSeq GUI requires R packages EBSeq and RGtk2. Once they are installed, a single command `EBSeqInterface()` from the R command line brings up the GUI shown in Figure 2.6. Information on downloading and installing these packages is available at the EBSeq website, along with a manual

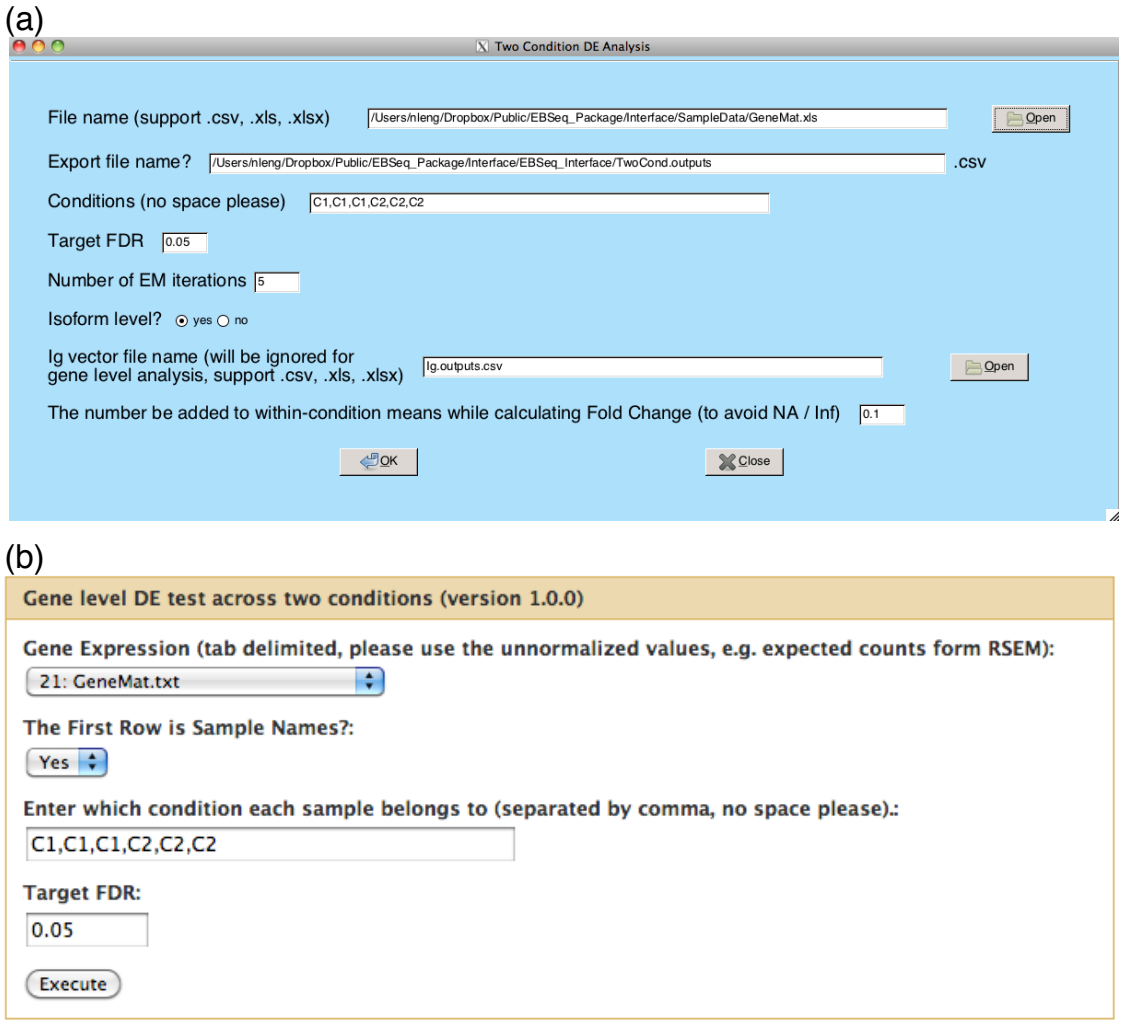


Figure 2.6: Panel (a) shows the EBSeq GUI for the gene (isoform) DE analysis across two conditions. A user may provide expression estimates (and the I_g vector for isoform analysis) in .csv, .xls, or .xlsx format, and can also specify a target FDR level and the number of iterations for the EM algorithm. (b) shows the Galaxy interface for gene DE analysis across two conditions.

that provides examples.

EBSeq in Galaxy: EBSeq is also available in Galaxy, an open, web-based platform that aims to make computational biology accessible to a wide-range of research scientists (Goecks et al., 2010). Figure 2.6 (b) shows the Galaxy interface for a gene DE analysis across two conditions. To use EBSeq in Galaxy, both EBSeq and Galaxy must be installed along with the EBSeq Galaxy wrappers. These wrappers and additional information are available in the Galaxy toolshed.

2.5 Discussion and future work

The main difference between EBSeq and the other approaches considered here is that EBSeq models isoform expression directly, as opposed to gene expression, and in so doing accommodates isoform expression estimation uncertainty. In particular, estimation uncertainty is partitioned into three groups defined by isoform complexity ($I_g = 1, 2$, or 3), following our empirical observation that uncertainty is increased on average in isoforms that share a parent gene. EBSeq is not restricted to three groups and for some genomes, additional I_g groups may be warranted. As detailed in section 2.4, EBSeq is also not restricted to this definition of complexity.

EBSeq shows increased power over Cuffdiff2 for identifying DE isoforms. Although developed to facilitate isoform inference, like Cuffdiff2, EBSeq may also be used for identifying DE genes. It shows slightly increased power over most count-based methods in both simulation and case studies, without major losses in efficiency when outliers are present.

A second difference is that, unlike most approaches which classify non-DE genes as EE, EBSeq is based on a parametric mixture model which facilitates evaluation of the posterior probabilities associated with DE, as well as EE. The particular parameterization provides closed form predictive distributions that facilitate efficient computation. (However, diagnostics should always be checked to ensure model fit (for example, Appendix Figure A.8)). Once posterior probabilities are obtained from a well-fit model, a user may identify an FDR controlled list of EE genes. This may be of particular interest for genes with more than one isoform,

since compensatory mechanisms may give rise to DE isoforms in EE genes; and consequently subtle, yet important, differences may be missed if focus is placed exclusively on DE genes alone. EBSeq also provides estimation of posterior FC across conditions in a two or more condition experiment. Using the posterior FC, instead of the empirical FC, is more robust for low expressed genes.

EBSeq may be improved in a number of ways. One interesting application would be accommodating continuous covariates (e.g. age, drug dosage). Suppose a continuous covariate V_s is measured for each sample s. To accommodate V_s, instead of assuming $r_{gi,s} = r_{gi,0} * l_s$, we may assume $r_{gi,s} = r_{gi,0} * l_s * (1 + \alpha * V_s)$. Under this assumption, the expected mean of isoform g_i in sample s and condition C is $\frac{(1+\alpha*V_s)*r_{gi,0}*l_s*(1-q^C_{gi})}{q^C_{gi}}$; $r_{gi,0}$ and l_s may be estimated as described before. α is a parameter shared by all the samples and isoforms, and might be estimated via EM algorithm. Multiple continuous covariates could be modeled similarly. Additionally, we may introduce a random effect factor to accommodate experiments with paired samples.

Also, we might improve EBSeq by modeling the variance-covariance structure among multiple isoforms within the same gene. We may model expression of isoforms in the same gene as a multi-variate Poisson distribution with inflated variance. The inflated variance may be modeled by a Gamma distribution. This will give us more precise estimates of isoform mean and variance, and also provide information of co-regulation structures among these isoforms. By doing so we may also be able to test differential transcription events (changes in the proportion of reads assigned to an isoform) and differeitial co-expression events among multiple isoforms, in addition to DE inference of each individual isoform. More details may be found at Appendix Section A.6. In addition, another interesting extension that would improve power in detecting changes in an time course or spatial course experiment is accounting for the dependence over the ordered conditions. We focus on this problem in Chapter 3.

3 EBSeq-HMM: An auto-regressive HMM model for identifying gene/isoform expression changes in ordered RNA-seq experiments

3.1 Background

Of primary interest in the mouse limb experiment and other experiments with ordered conditions is characterizing how genes are changing over time, space, gradient, etc. For example, an investigator may be interested in genes with expression profiles that are monotonically increasing (or decreasing), that increase initially then decrease, that increase initially then remain unchanged, and so on. We refer to these types of changes in expression hereinafter as *expression paths*.

We consider three types of expression paths: (i) constant paths: expression remains unchanged, or equally expressed, over all conditions; (ii) sporadic paths: expression shows some change between at least one pair of adjacent conditions, but remains unchanged between at least one other pair; and (iii) dynamic paths: expression changes continuously. Among the three categories, of primary interest in most ordered RNA-seq experiments are dynamic paths since genes from these paths directly reflect the continuous response to treatments or environmental changes along the time or spatial course of the experiment.

A number of robust statistical methods have been developed for identifying DE genes (EBSeq (Leng et al., 2013), DESeq (Anders and Huber, 2010), edgeR (Robinson et al., 2010), baySeq (Hardcastle and Kelly, 2010), Cuffdiff2 (Trapnell et al., 2012a)) as well as isoforms (EBSeq, rSeqDiff (Shi and Jiang, 2013), Cuffdiff2, BitSeq (Glaus et al., 2012)) in a two-condition RNA-seq experiment. The majority of these methods may also be used to identify changes across multiple conditions. In most cases, the statistical tests employed are designed to identify a gene as DE if it shows a change in at least one condition; and, consequently, non-constant genes are detected collectively. Subsequent analysis must be done to distinguish sporadic paths from dynamic ones, to classify genes into distinct paths, and to

assess the associated classification uncertainty. In addition, as these approaches were not designed specifically for ordered experiments, they sacrifice power by not accommodating dependence across conditions.

To directly address the main questions of interest in an ordered RNA-seq experiment, we have developed an empirical Bayes autoregressive hidden Markov model (HMM) based approach called EBSeq-HMM. The model extends our previous work, EBSeq, for identifying DE genes and isoforms across two or more conditions. As detailed in Methods, an autoregressive process describes changes in expression over conditions, and a hidden Markov component is used to accommodate dependence.

3.2 Methods

EBSeq-HMM: An auto-regressive HMM model for identifying gene/isoform expression changes in ordered RNA-seq experiments

EBSeq-HMM requires estimates of gene or isoform expression collected over three or more ordered conditions. The general model is presented for gene-level analysis; the isoform-level model is discussed in the last part of this Methods section. To simplify the presentation, we refer to ordered conditions as time points denoted by $t = 1, 2, \ldots, T$, noting that the method directly accommodates other ordered data structures (e.g. space, gradient, etc.).

Let $\mathbf{X}_t$ be a $G \times N_t$ matrix of expression values for G genes in N_t samples at time t. The full set of observed expression values is then denoted by $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T)$. With a slight abuse of notation, let $\mathbf{X}_g$ denote one row of this matrix containing data for gene g over time; X_{gtn} denotes expression values for gene g at time t in sample n. Of interest are changes in the latent mean expression levels for gene g: $\mu_{g1}, \mu_{g2}, \ldots, \mu_{gT}$. We allow for three possibilities, or states, to describe such changes: Up, Down, EE. If $\mu_{t-1} < \mu_t$, we define state $S^{\Delta t}$ as Up; if $\mu_{t-1} > \mu_t$, $S^{\Delta t}$ is Down; and $\mu_{t-1} = \mu_t$ defines $S^{\Delta t}$ as EE. The main goals in an ordered RNA-seq

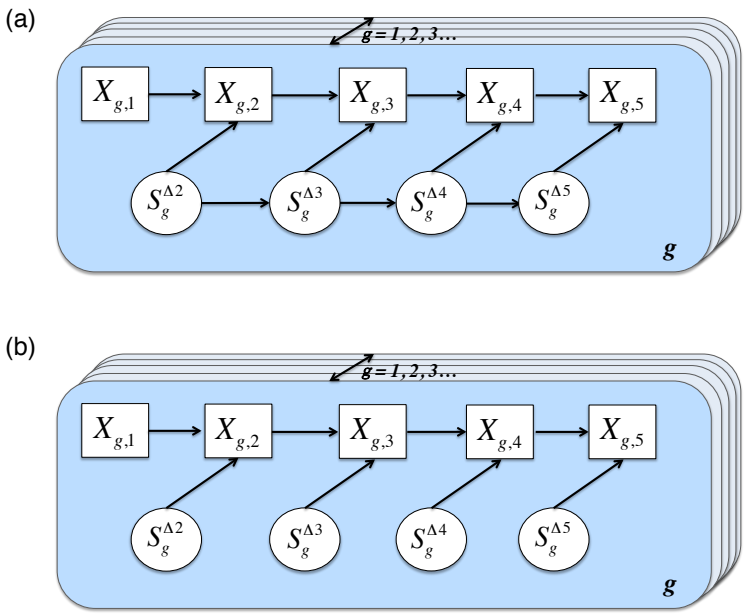


Figure 3.1: (a)An auto-regressive hidden Markov component models dynamic paths. (b)A non-markov component models constant and sporadic paths.

experiment - identifying genes that change over time, and specifying each genes' expression path - can be restated as questions about these underlying states. In short, for each gene g and each transition between $t-1$ and t, we would like to estimate the probability of each state. A gene is said to follow a non-constant path if at least one state is not EE. We would also like to estimate the most likely expression path, which is given by the configuration of expression states over time ($S_g^{\Delta 2}$, $S_g^{\Delta 3}$, ..., $S_g^{\Delta T}$), noting that the most likely configuration of states need not equal the collection of states that define $S_g^{\Delta t}$ marginally at each t.

To make inference regarding these states, we propose a model for the set of expression measurements taken on a gene g. We make the common and well-supported assumption that gene expression in an RNA-seq experiment is well described by a Negative Binomial (NB) distribution. Were we to consider time t in isolation, this implies $X_{gtn}|r_{gt}, q_{gt} \sim NB(r_{gt}, q_{gt})$ where the NB distribution may

be parameterized such that $\mu_{gt} = r_{gt}(1 - q_{gt})/q_{gt}$. For simplicity of notation, we assume equal library sizes. Details on adjustments for unequal library sizes are given in the part below.

Since our interest here is in quantifying changes in $\mathbf{X}_{gt}$ over time, we assume expression at time t depends on that at $t-1$ through parameters r and q. Specifically, $\left(X_{gtn}|r_{g,t-1}, q_{g,t-1}, S_g^{\Delta t} = s\right) \sim \text{NB}\left(r_{g,t-1}\xi_g^s, q_{g,t-1}\right)$ where $\xi_g^s = c$ if s is Up; $\xi_g^s = 1/c$ if s is Down; and $\xi_g^s = 1$ if s is EE. The data dependent parameter c specifies the expected change associated with each state. For example, if $c = 2$, then $S_g^{\Delta t} = \text{Up}$ refers to a two-fold increase in expression between $t-1$ and t. Although c may be defined by a user, we suggest estimation by maximum likelihood. We further model fluctuations in μ_{gt} by defining a prior distribution for $q_{gt} : q_{gt}|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$ for all g and t. Given this set-up, the marginal predictive conditional distribution describing expression (or emissions) for each state is Beta-Negative Binomial: $\left(X_{gtn}|X_{g,t-1} = x_{g,t-1}, S_g^{\Delta t}, \Theta\right) \sim$ Beta-NB$(\alpha + N_{t-1}r_{g,t-1}, \beta + \sum_j x_{g,t-1,j}, \xi_g^s r_{g,t-1})$ where $\Theta = \left[\alpha, \beta, r_{g,t-1}, \xi_g^s\right]$. The expected mean is then defined as $\xi_g^s r_{g,t-1}(\beta + \sum_j x_{g,t-1,j})/(\alpha + N_{t-1}r_{g,t-1})$.

To model dependence among states, we separately consider genes in dynamic, sporadic, and constant paths. For genes with dynamic paths, each state $S_g^{\Delta t}$ is dependent on the prior state $S_g^{\Delta t-1}$ since these genes represent continuous changes over time. To accommodate this dependence, we assume that the state process is described by a Markov chain. The sporadic and constant genes do not show continuous changes over time, and consequently we assume states are independent. Taken together, since we do not know the expression path type *a priori*, the model for the full set of expression measurements is a mixture over these two possibilities.

In summary, the time course X_g for a dynamic gene is governed by two interrelated probabilistic mechanisms: the conditional distribution (emissions model) at each time and the process describing the evolution of states over time. Initially, we assume that the observed expression vector can be characterized by the Beta-Negative Binomial model described above and that the state process can be described by a Markov chain. Were it the case that dependence among measurements is fully captured by the state process, the proposed model would be a standard

hidden Markov model. However, this last assumption does not hold, given that X_t depends not only on the state $S^{\Delta t}$ but also on X_{t-1} through r_{t-1}. Consequently, the model for dynamic genes is given by a Markov-switching autoregressive model, as in Hamilton (1989) and Ailliot and Monbet (2012) (see Figure 3.1). For sporadic and constant genes, we assume the same emissions model, but do not assume the state process is Markov. The marginal distribution of the data is then given by a two-component mixture over dynamic and sporadic/constant genes.

Model specification taking accounts of library size factors

Denote the library size factor of sample n at time t as l_{tn}. Taking account of library size factors, we assume $r_{g,tn} = r_{g,t} l_{tn}$. Then $X_{g,tn}|r_{g,t}, q_{g,t}, l_{tn} \sim NB(l_{tn} * r_{g,t}, q_{g,t})$ where the mean is defined as $\mu_{g,tn} = l_{tn} * r_{g,t}(1 - q_{g,t})/q_{g,t}$.

Denote $\mu_{g,t}$ and $(\sigma_{g,t})^2$ as the mean and variance of gene g at time t under the standard library size. Then $\mu_{g,t} = \frac{1}{l_{tn}}\mu_{g,tn}$ for any n at time point t. Assuming there are N_t samples at time point t, we obtain the unbiased estimator $\hat{\mu}_{g,t} = \frac{1}{N_t}\sum_{n \text{ at } t} \frac{1}{l_{tn}}\hat{\mu}_{g,tn}$ where $\hat{\mu}_{g,tn} = X_{g,tn}$. Since $(\sigma_{g,t})^2 = \frac{1}{l_{tn}}(\sigma_{g,tn})^2$ for any n at t, we obtain the estimator $(\hat{\sigma}_{g,t})^2 = \frac{1}{N_t}\sum_{n \text{ at } t} \frac{1}{l_{tn}}(\hat{\sigma}_{g,tn})^2$, which is unbiased conditioning on $\mu_{g,tn} = \hat{\mu}_{g,tn}$ where $(\hat{\sigma}_{g,tn})^2 = (X_{g,tn} - l_{tn}\hat{\mu}_{g,t})^2$. Then the estimator of $r_{g,t}$ is obtained by $\hat{r}_{g,t} = \frac{\hat{\mu}^2_{g,t}}{\hat{\sigma}^2_{g,t} - \hat{\mu}_{g,t}}$.

Subsequently, the conditional probability becomes $\left(X_{g,tn}|X_{g,t-1} = x_{g,t-1}, S_g^{\Delta t}, l_{tn}\right) \sim$ Beta-NB$(\alpha + N_{t-1}r_{g,t-1}, \beta + \sum_j x_{g,t-1,j}, l_{tn}\xi_g^s r_{g,t-1})$. The expected mean is then defined as $l_{tn}\xi_g^s r_{g,t-1}(\beta + \sum_j x_{g,t-1,j})/(\alpha + N_{t-1}r_{g,t-1})$.

Parameter estimation

In the emissions distributions, the unknown parameters (r's, α, and β) are estimated using the method of moments; c is estimated via maximum likelihood. r's are estimated within time point and α and β are estimated using all samples. Recall that EBSeq-HMM assumes a mixture model with a Markov component $m1$ and a non-Markov component $m2$. We assume equal prior probabilities of being in each

mixture component, and assume equal probabilities of being in each state in the non-Markov component $m2$.

In Markov chain $m1$, the Baum-Welch algorithm is used to estimate initial $\pi_j = P(S_g^{\Delta_2} = j|m1)$ and state transition probabilities $a_{dj}^{t,t+1} = P(S_g^{\Delta_{t+1}} = j|S_g^{\Delta_t} = d, m1)$ for $t \geqslant 2$. Here we assume a non-homogeneous Markov chain for the hidden states so $a_{dj}^{t,t+1}$'s are different for different t's. Denote the vector of initial probabilities and the state transition matrices estimated from the last step as $\tilde{\pi}, \tilde{A}$. Given parameter estimates $\tilde{\pi}, \tilde{A}$, define $z_g^{m1} = P(M_g = m1|X_g, \tilde{\pi}, \tilde{A})$ and $b_j(X_{gt}) = P(X_{gt}|S_g^{\Delta_t} = j, X_{g,t-1} = x_{g,t-1})$. The forward and backward steps of the Baum-Welch algorithm are then defined as follows:

$$\begin{aligned}\alpha_{g,j}(t) &= [\sum_d \alpha_{g,d}(t-1)\tilde{a}_{dj}^{t-1,t}]b_j(X_{g,t}) \\ &\propto P(X_{g1}, ..., X_{gt}, S_g^{\Delta_t} = j|m1)\end{aligned}$$

$$\begin{aligned}\beta_{g,j}(t) &= \sum_d [\beta_{g,d}(t+1)b_d(X_{g,t+1})\tilde{a}_{jd}^{t,t+1}] \\ &\propto P(X_{g,t+1}, ..., X_{gT}|X_{gt}, S_g^{\Delta_t} = j, m1)\end{aligned}$$

The initial and transition probabilities are updated by:

$$
\begin{aligned}
a_{D,J}^{t,t+1} &= \frac{\sum_g P(S_g^{\Delta_t} = D, S_g^{\Delta_{t+1}} = J, M_g = m1 | X_g, \tilde{\pi}, \tilde{A})}{\sum_g \sum_j P(S_g^{\Delta_t} = D, S_g^{\Delta_{t+1}} = j, M_g = m1 | X_g, \tilde{\pi}, \tilde{A})} \\
&= \frac{\sum_g \alpha_{g,D}(t) \tilde{a}_{D,J}^{t,t+1} b_{g,J}(X_{g,t+1}) \beta_{g,J}(t+1) z_g^{m1}}{\sum_g \sum_j \alpha_{g,D}(t) \tilde{a}_{D,j}^{t,t+1} b_{g,j}(X_{g,t+1}) \beta_{g,j}(t+1) z_g^{m1}} \\
\pi_J &= \frac{\sum_g P(S_g^{\Delta_2} = J, M_g = m1 | X_g, \tilde{\pi})}{\sum_g \sum_j P(S_g^{\Delta_2} = j, M_g = m1 | X_g, \tilde{\pi})} \\
&= \frac{\sum_g \alpha_{g,J}(2) \beta_{g,J}(2) z_g^{m1}}{\sum_g \sum_j \alpha_{g,j}(2) \beta_{g,j}(2) z_g^{m1}}
\end{aligned}
$$

Parameters are estimated by fixing expected fold-change c at 1.2. The process is then repeated for c in (1.4, 1.6, ...,3); and the parameter set with maximum likelihood is used in the final model.

Inference at the isoform level

The model detailed in the previous parts applies to gene counts. To apply the approach to isoforms, the uncertainty inherent in isoform expression estimation should be accommodated. In short, estimating expression at the gene-level is a relatively easy task in RNA-seq as all reads mapping to a gene's constituent exons may be used. The same holds true for estimating expression for an isoform unique to its parent gene. However, for genes with multiple isoforms, the problem is more challenging as reads mapping to overlapping exons must be allocated to isoforms in a way that is consistent with their expression. Consequently, there is increased uncertainty (on average) in expression estimates for isoforms with multiple overlapping exons, referred to as complex isoforms; and the uncertainty has been shown to have a substantial effect on downstream analysis methods.

Specifically, define an isoform of gene g as belonging to the $I_g = k$ group, for example, where $k = 1, 2$ or 3, if the total number of isoforms from gene g is k (the $I_g = 3$ group contains all isoforms from genes having 3 or more isoforms). In Chapter 2 we demonstrated that there is decreased variability in the $I_g = 1$ group, but increased variability in the others, due to the relative increase in uncertainty inherent in estimating isoform expression when multiple isoforms of a given gene are present. This observation is not specific to the dataset and/or the method used for isoform expression estimation; it is also not specific to the particular method used for quantifying isoform complexity.

To adjust for the increased uncertainty inherent in complex isoform expression estimates, we allow the Beta prior to depend on isoform group: $q^C_{g_i}|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$. As above, the hyperparameter α is shared across isoforms, but here β depends on I_g, accommodating the systematic differences in variability among the I_g groups. I_g quantifies a measure of isoform complexity and may be defined by the user as the number of isoforms from a gene, as described above. It could also be defined by an isoform's mappability score or credibility interval as provided by Koehler et al. (2011), Li and Dewey (2011), or Derrien et al. (2012) .

3.3 Results

Simulated data

We followed the simulation setup of Robinson and Smyth (2007) by defining counts as Negative Binomial with gene-specific mean in sample n and time point t given by μ_{gt} and variance $\mu_{gt}(1 + \mu_{gt}\phi_{gt})$. The (μ_{gt}, ϕ_{gt})'s were sampled as pairs from the mouse limb case study data described in the next part of this Results Section. Paired sampling was done to preserve the mean-variance relationship observed in most RNA-seq datasets. Each simulated dataset contains 10000 genes and 15 samples which represent three biological replicates at each of five time points. One hundred datasets were considered for each simulation scenario.

Sim I considers dynamic changes over time for 60% of the genes. For these

genes, paths were generated from an HMM.

With five conditions, there are four states in the hidden chain, so three state transition matrices were used. We defined equal starting probabilities, and defined the state transition matrices as $\left(\begin{smallmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{smallmatrix}\right)$, and $\left(\begin{smallmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{smallmatrix}\right)$. The main expression paths were defined as Up-Up-Down-Down and Down-Down-Up-Up under this set-up. Other paths were simulated as well, although they were less common. Once a gene's particular path (collection of states) was generated, $\mu_{g,t+1}$ was simulated as μ_{gt} multiplied (divided) by δ if $S_g^{\Delta_{t+1}}$ was Up (Down). For $1/2$ of the dynamic genes, we simulated strong effects, with δ sampled from empirical fold-changes (FCs) between 1.4 and 1.5 calculated using case study data. The other $1/2$ represent weak effects with δ sampled from empirical FCs between 1.2 and 1.4. The remaining 40% of genes were simulated as constant meaning the latent level of expression remains unchanged across conditions. To simulate genes following constant paths, we only took the genes whose empirical FC of medians between any two adjacent time points is within $(1/1.2, 1.2)$.

Sim II For this simulation scenario, 40% of the 10000 genes were simulated as dynamic as in Sim I and another 20% were simulated as sporadic. For dynamic genes, paths were generated from an HMM as described in Sim I; half were simulated as strong effects and the other half were with weak effects. For the sporadic genes, a time point t was chosen at random and μ_{gt} was defined to be $\mu_{gt} * \delta$, where δ was sampled from empirical FCs between 1.4 and 1.5. The remaining 40% of genes were simulated as constant, again as described in Sim I.

Case Study Data

Of interest in our case study, detailed below, is RNA-seq data from the James Thomson Lab at the Morgridge Institute for Research. We evaluated gene expression from 7 positions along the mouse limb: proximal stylopod, distal stylopod, elbow, proximal zeugopod, distal zeugopod, autopod, digit. Three 12-week old C57BL/6J female mice were euthanized by cervical dislocation, followed by the extraction of the right forelimb. The tissues were treated with RNAlater (Sigma), per manufac-

turerâŁ™s instructions, dissected using a SteREO Discovery.V8 microscope (Zeiss), and stored at -20°C. The tissues were homogenized and lysed using a variable speed rotor stator homogenizer and Qiazol (Qiagen). Message RNA was extracted from the homogenized tissue samples using Qiagen's RNeasy Lipid Tissue Mini (digits) and Midi (all other) Kits. A total of 21 samples were sequenced using Illumina's Directional mRNA-Seq protocol (Part # 15018460 Rev. A). The reads are single-end with read length 42-bp. Each sample was run on one lane of an Illumina GAII. Alignment was done using Bowtie (Langmead et al. (2010)) with the hg19 RefSeq annotation. Expression estimates were obtained from RSEM (Li and Dewey (2011)) and library size factors were obtained using median normalization Anders and Huber (2010).

Identification of DE genes and classification

EBSeq-HMM is compared with a naive method based on fold-change (FC), DESeq2 (version 1.4.1), and edgeR (version 3.6.0). Two tasks are of interest: identifying DE genes, defined as those showing any change across conditions; and assigning DE genes into their most likely expression path.

Identification of DE genes: Recall that EBSeq-HMM provides gene-specific posterior probabilities associated with each expression path. To identify a list of DE genes with FDR α via EBSeq-HMM, we take those genes for which the posterior probability of being constant is less than or equal to α. With FDR controlled at α, the most likely path of a DE gene is then defined as the path with highest posterior probability (PP).

For the naive FC method, denote med_g^t as the median expression of gene g at time point t. A gene g is called Up (Down) between t and $t+1$ if $\frac{med_g^{t+1}}{med_g^t}$ is greater than (less than) K; otherwise, it is EE. We evaluate five values of K: 1.2, 1.3, 1.5, 2, and 2.5. A gene is defined as DE if it is non-EE at any transition.

Both DESeq2 and edgeR implement a generalized-linear model to test data ~ intercept vs. data ~ intercept + condition with derived p-values adjusted for multiplicities using Benjamini-Hochberg. To construct a list of DE genes with target

FDR α, we consider those genes with adjusted p-values less than or equal to α.

Classification of genes into expression paths: For EBSeq-HMM, a DE gene is classified into a specific expression path if its posterior probability (PP) of being in that path exceeds 0.5. Selecting genes with PP>0.5 ensures that the posterior maximizing class always minimizes the Bayes risk regardless of choice of the metric loss function (Schlüter et al. (2005)), although we note that there may be reasons to consider different thresholds (see Discussion). Since no uncertainty measure of assignment is available using FC, for the FC analysis a gene is classified into the path defined by the Up/Down/EE calls across transitions. For DESeq2 and edgeR, classifying DE genes into expression paths is not of interest, and no clear guidelines on how to do so is provided. Consequently, these methods are not evaluated for expression path classification.

Simulation Results

Simulation studies were conducted to investigate the operating characteristics of EBSeq-HMM and to assess how it compares with FC, DESeq2, and edgeR. As detailed in Methods, each simulated dataset derives counts from a Negative Binomial model. Like EBSeq-HMM, DESeq2 and edgeR also assume that counts are distributed as Negative Binomial, and consequently this assumption should not provide advantage, or lack thereof, to any one method in particular. As the form of the variance is that assumed in edgeR, there may be a slight advantage given to that method. Parameter estimates were derived from case study data to help ensure that many features of real data are preserved in the simulation (e.g. mean/variance relationship and magnitude of FCs; see Appendix Figure B.1 for more details).

Table 3.1 shows the power and FDR for identifying dynamic genes in Sim I where the target FDR is controlled at 5%. In addition to showing power overall, it is also shown separately for strong and weak effects (FDR is not shown for each subgroup because false discoveries are discoveries of EE genes and therefore cannot be classified as strong or weak). EBSeq-HMM has higher power overall, which is largely due to its ability to identify genes showing subtle, yet consistent, changes

Table 3.1: Operating characteristics of identification of expression changes in Sim I

	Power	FDR	Power (strong)	Power (weak)
EBSeqHMM	99.3%	3.3%	99.9%	98.7%
DESeq2	94.8%	0%	97.6%	92.0%
edgeR	94.8%	0.1%	97.8%	91.8%
FC (2.5)	0.9%	0.2%	1.2%	0.6%
FC (2)	5.4%	0.8%	7.5%	3.2%
FC (1.5)	62.0%	2.4%	84.5%	39.3%
FC (1.3)	94.7%	8.1%	99.7%	89.8%
FC (1.2)	99.4%	19.6%	100%	98.9%

The first two columns show the average power and FDR for detecting DE genes in Sim I. Power within the strong and weak groups is further evaluated in columns 3 and 4. Averages are calculated over 100 Sim I simulations. The standard deviations (not shown) for EBSeq-HMM, DESeq2, and edgeR (and in most cases FC) were $\leqslant 0.005$.

over time. Specifically, the power of the three methods is comparable for genes with strong effects, but EBSeq-HMM has a substantial advantage in identifying genes where changes between any two points are relatively small. An example of two genes identified by EBSeq-HMM but neither edgeR nor DESeq2 is shown in Figure 3.2 (a) and (b). It is clear from the figure that the change between any two points is small (FC$<$ 1.4) and in some cases changes would not be identified by a marginal analysis between adjacent time points (e.g. time points 1 and 2 in Figure 3.2(b)), but EBSeq-HMM identifies the genes as dynamic given the consistent changes over time. The FC analysis works best at threshold 1.3, but is still inferior to the other methods.

Table 3.2 shows the power and FDR for identifying DE genes (either dynamic or sporadic) in Sim II where, again, the target FDR is controlled at 5%. The increased power of EBSeq-HMM in identifying dynamic genes with weak effects that was demonstrated in Sim I results persists when sporadic genes are present. Note all methods have reduced power in identifying sporadic genes. This is because

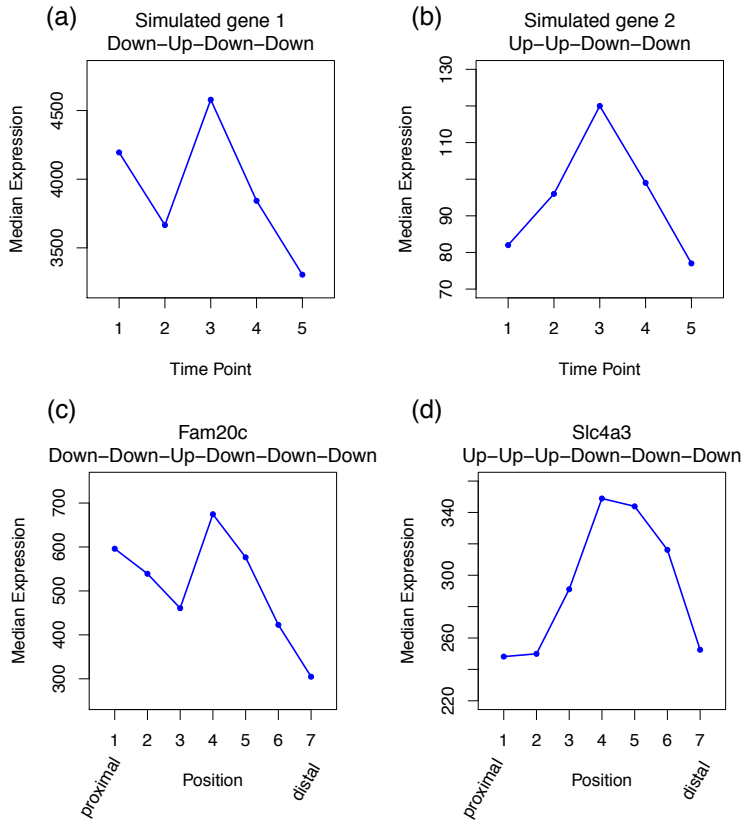


Figure 3.2: Shown are two genes identified by EBSeq-HMM but neither DESeq2 nor edgeR in Sim I data (upper) and in case study data (lower). The x-axis shows time points (upper) and positions on mouse limb (lower) and the y-axis shows median gene expression adjusted for library sizes.

the differences in sporadic genes are weaker than dynamic genes. For example, assuming a constant fold change 1.5 for any DE transitions over 5 time points. In a monotone increasing case, the max ratio between any two μ_t's would be $1.5^4 = 5.06$ and the max ratio between μ_t and estimated intercept would be around $\sqrt{5.06} = 2.25$. However, in a sporadic gene, the max ratio between any two μ_t's would be 1.5 and the max ratio between μ_t and estimated intercept would be less than 1.5. Table 3.2 also suggests that EBSeq-HMM has higher power than DESeq2 or edgeR for identifying sporadic genes. While fitting the alternative hypothesis H1

Table 3.2: Operating characteristics of identification of expression changes in Sim II

	Power	FDR	Power (strong)	Power (weak)	Power (sporadic)
EBSeqHMM	97.8%	3.4%	99.9%	98.6%	94.7%
DESeq2	90.3%	0%	97.6%	92.2%	81.2%
edgeR	90.6%	0%	97.8%	92.4%	81.5%
FC (2.5)	0.7%	0.4%	1.3%	0.6%	0.1%
FC (2)	4.2%	1.3%	7.7%	3.3%	1.6%
FC (1.5)	56.4%	2.7%	84.5%	39.4%	45.4%
FC (1.3)	91.9%	8.3%	99.7%	89.8%	86.1%
FC (1.2)	98.5%	19.8%	100%	99.0%	96.4%

The first two columns show the average power and FDR for detecting DE genes in Sim II. For dynamic genes, the power within the strong and weak groups is further evaluated in columns 3 and 4. Power within the sporadic group is evaluated in column 5. Averages are calculated over 100 Sim II simulations. The standard deviations (not shown) for EBSeq-HMM, DESeq2, and edgeR (and in most cases FC) were $\leqslant 0.005$.

as defined in Method section, DESeq2 and edgeR estimate coefficients of the second to T^{th} condition. To detect DE genes, they apply Wald test and Likelihood Ratio test respectively to compare the fitted model under H1 to H0. Both Wald test and Likelihood Ratio test are test assuming approximated null distribution. To infer the significance in difference, each compares its test statistics to a χ^2 distribution whose degree of freedom (df) is defined to be df difference between the two models ($\mathtt{df}(\mathrm{H1}) - \mathtt{df}(\mathrm{H0})$). The df difference would be large when an experiment has large number of conditions. As a result, a large df difference will reduce statistical power of the test, especially in this case when signals in sporadic genes are weak. Instead of adapting the tests with asymptotic assumptions, EBSeq-HMM builds on Bayes theory which directly models the probabilities of expression paths. Therefore it provides higher power in cases with weaker signals.

In addition to identification of DE genes, we also evaluated the ability of EBSeq-HMM and FC to classify genes into distinct expression paths (DESeq2 and edgeR

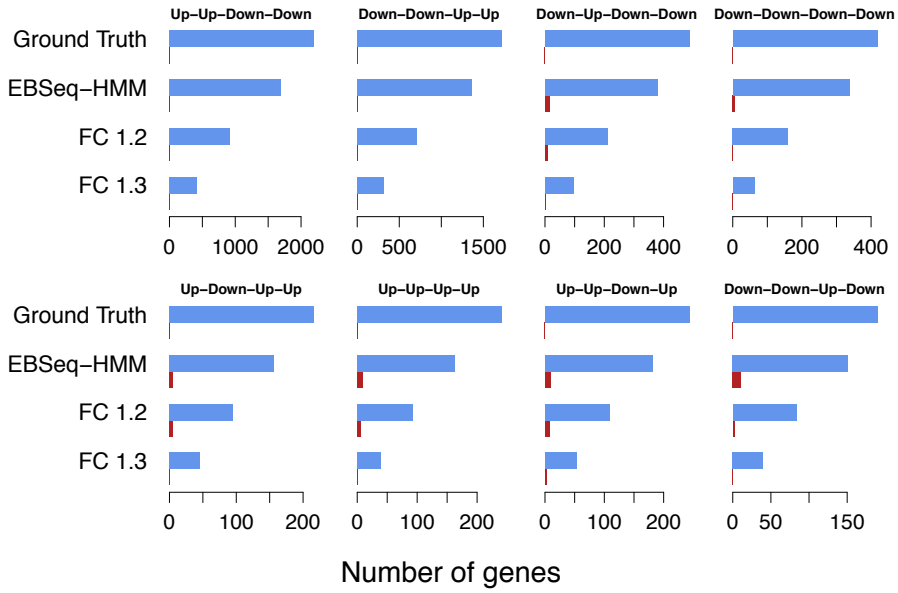


Figure 3.3: Shown are the number of genes (ground truth) simulated in Sim I as being in each of 8 dynamic paths (these 8 are shown as they represent the most genes among all simulated paths). Also shown are the average number classified into each path by EBSeq-HMM and by FC analysis at thresholds 1.2 and 1.3 (averages are calculated over 100 Sim I datasets). Correct classifications are shown in blue; incorrect are shown in red.

were not evaluated as they were not developed for this purpose). Figure 3.3 shows results for 8 dynamic paths simulated in Sim I; these 8 were chosen as they represent the most genes among all simulated paths. The ground truth shows the number of genes simulated in each expression path. Also shown are the average number classified into each path by EBSeq-HMM and by FC analysis at FC threshold $K = 1.2$ and 1.3 (averages are calculated over 100 Sim I datasets). Correct classifications are shown in blue; incorrect are shown in red. For FC analysis, we chose 1.2 and 1.3 as they performed best under all thresholds considered. As shown, EBSeq-HMM identified more True Positives than FC, while the FDR is well below 5%. Similar results were observed in Sim II data (see Appendix Figure B.2).

Case Study Results

We applied each method on the mouse limb case study data. EBSeq-HMM, DESeq2 and edgeR identified 14817, 11517 and 9520 DE genes at a 5% target FDR. Of the 11517 (9520) genes identified by DESeq2 (edgeR), 93% (96%) are also found by EBSeq-HMM. FC analyses identified 4225, 6500, 10881, 14016 and 15877 genes for $K = 2.5$, 2, 1.5, 1.3 and 1.2, respectively; and the identifications showed the lowest overlap with the other three methods. This, coupled with the poor performance of FC in the simulation study, leads us to exclude FC from further evaluation.

Given that the majority of genes identified by edgeR and DESeq2 are also identified by EBSeq-HMM, we focus initially on genes that are identified exclusively by EBSeq-HMM. Figure 3.2 (c) and (d) show two examples. As in the simulated data (shown in (a) and (b)), these genes have subtle but consistent changes over the 7 limb positions, again demonstrating that by accommodating dependence, EBSeq-HMM has increased power to identify genes showing relatively weak, but consistent, changes over ordered conditions. Similar results for other genes identified exclusively by EBSeq-HMM may be found at Appendix Figure B.3.

In addition to increased power for identifying DE genes, EBSeq-HMM may prove useful in classifying genes into particular expression paths. To illustrate, we consider Hox genes, a set of genes that are of primary interest here as they are well known to play an important role in maintaining positional identity in adult cells (Wang et al., 2009a; Rinn et al., 2006). In our case study data, 33 out of 39 Hox genes were identified as DE by EBSeq-HMM. Figure 3.4 shows expression levels of the 33 genes along with their most likely expression paths. Although the positional changes for most Hox genes are not well known, it is known that Hoxb4 and Hoxb8 have up-regulated expression in proximal sites (Wang et al., 2009a; Rinn et al., 2006). The EBSeq-HMM paths for these genes are consistent with these prior studies and provide further information as they characterize changes across the seven positions. In addition, the overall pattern of Hox gene expression with, in general, higher numbered Hox genes being unregulated distally and lower numbered Hox genes being unregulated proximally is in agreement with existing data and models of

proximal distal patterning of the limb (Zakany and Duboule, 2007).

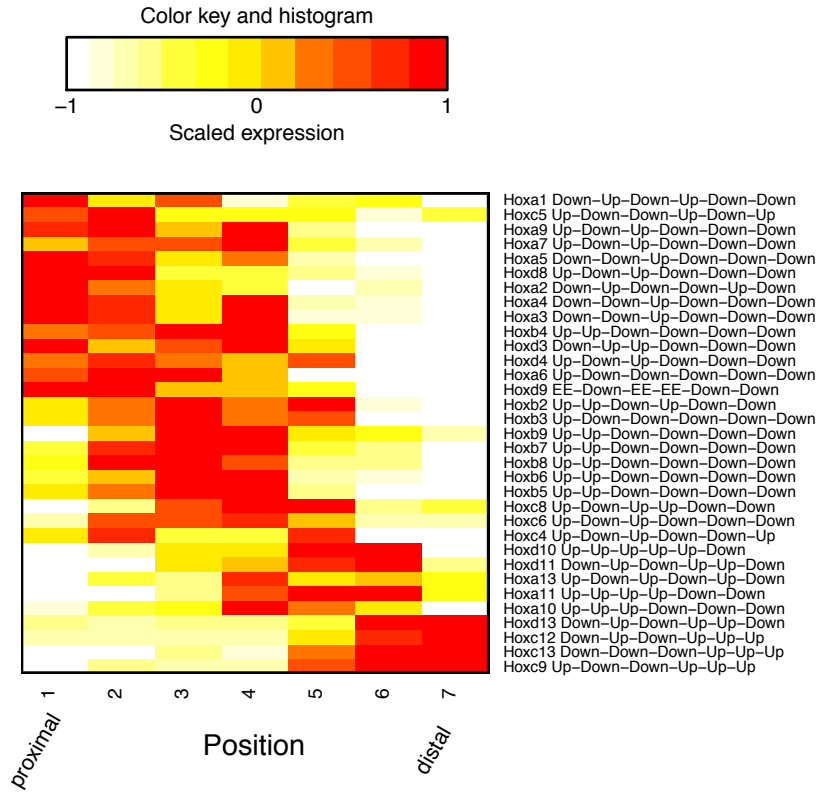


Figure 3.4: Shown are scaled expression of 33 Hox genes identified as DE by EBSeq-HMM. The expression values were adjusted for library size and further scaled to mean 0 and standard deviation 1 for each gene. The x-axis shows 7 positions over the mouse limb, median expression of each position is shown. Genes were clustered via hierarchical clustering using Euclidean distance and complete linkage.

To explore other genes beyond the Hox family that may be involved in positional identity, we consider 2347 genes that are classified by EBSeq-HMM into one of 64 possible dynamic paths. Among the 64 clusters formed by these dynamic genes, the two largest are Up-Down-Up-Down-Down-Down (827) and Down-Up-Down-Up-Up-Up (218). Figure 3.5 (a) and (b) show median expression of each position for each of these genes. As these groups each contain Hox genes but also previously

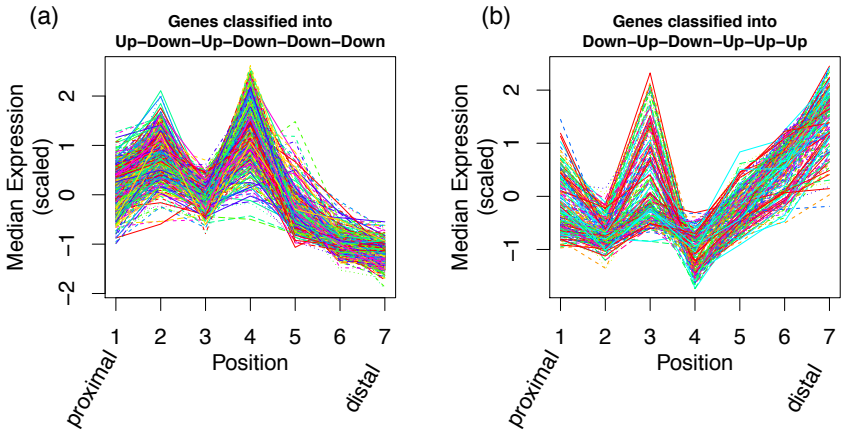


Figure 3.5: (a), (b) Shown are genes classified as following an Up-Down-Up-Down-Down-Down (left panel, 827 genes) or Down-Up-Down-Up-Up-Up (right panel, 218 genes) expression path in the case study data. Each line indicates one gene. The x-axis shows 7 positions over the mouse limb; the y-axis shows median scaled expression within each position.

unknown genes showing similar dynamics across position, the novel identifications define candidates for further study.

3.4 Implementation

EBSeq-HMM is implemented as an R package (EBSeq-HMM), currently available at `http://www.biostat.wisc.edu/~kendzior/EBSEQHMM/`. EBSeq-HMM requires estimates of gene or isoform expression, but is not specific to any particular estimation method. To estimate library sizes, EBSeq-HMM defaults to median normalization (Anders and Huber, 2010); TMM (Robinson and Oshlack, 2010) and Quantile Normalization (Bullard et al., 2010) are also available in the package.

Like most methods, EBSeq-HMM makes assumptions regarding the distribution governing expression measurements. Consequently, poor performance may result if there are strong departures from assumptions. Model diagnostics are implemented in EBSeq-HMM to ensure that assumptions can be easily checked. They should be considered with each application and results should not be used if serious departures from model assumptions are observed. A typical diagnostic summary for the case study data is shown in Appendix Figure B.4.

3.5 Discussion and future work

We have developed an approach called EBSeq-HMM for analysis of ordered RNA-seq experiments. EBSeq-HMM may be used to identify genes that are differentially expressed across a set of ordered conditions and to classify genes into their most likely expression paths. There are a number of methods available for identifying DE genes that may be used when data from multiple conditions is available. EBSeq-HMM has two main advantages over these approaches. First, it accommodates dependence across ordered conditions and consequently has increased power to identify genes showing subtle, yet consistent, changes. Second, for every gene, EBSeq-HMM calculates the gene-specific posterior probability associated with each possible expression path and in doing so allows for genes to be classified into distinct expression paths within a pre-specified level of uncertainty.

Simulations demonstrated the power of EBSeq-HMM over other approaches to identify DE genes. Results showed that most approaches perform well when

changes are relatively strong, but that EBSeq-HMM has increased power to identify genes showing weaker changes. EBSeq-HMM also worked well for identifying genes showing sporadic changes (where there is no dependence across ordered conditions as for some genes in Sim II).

In addition to DE gene identification, EBSeq-HMM performed well for classifying genes into expression paths. We defined a gene as being in a particular path if the gene was classified as DE at FDR 5% (posterior probability of EE was less than 0.05) and the posterior probability of being in that path exceeded 0.5. Given the two step process, observed mis-classification rates were well controlled. Note that in some cases, a DE gene may not be classified to any particular path. For example, if the last time point of a four-condition experiment is known to be noisy, a gene that is initially increasing may have equal posterior probability, say 1/3, of being Up-Up-Up, Up-Up-EE, and Up-Up-Down. This gene would be called DE with 5% FDR since PP(EE-EE-EE)$<$0.05, but it would not be assigned into a particular expression path if threshold 0.5 was used. In some cases, a user may want to modify these thresholds. If a false negative classification was considered more serious than a false positive, this threshold could be adjusted. Motivation for doing so under varying loss functions is discussed in Berger (1985).

If no dependence is present in the data, EBSeq-HMM will reduce to a mixture model with equal initial probabilities and equal transition probabilities. It will still provide posterior probabilities of being in each path, a sorted list of genes for each path, and the global profiles of expression paths. If the user is interested in one particular path and wants to maximize the power in detecting genes following this path, EBSeq-HMM may be applied with fixed initial probabilities and transition probabilities. For example, if monotone increasing and decreasing are of primary interest and the user wants to maximize the power in identifying genes following these 2 paths, EBSeq-HMM may be applied with transition probabilities $\left(\begin{smallmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{smallmatrix}\right)$ for all the state transitions.

As in EBSeq, there are a number of ways to extend the EBSeq-HMM model. For example, accommodating continuous covariates, accounting for dependence between isoforms from the same gene, etc, as discussed in Chapter 2. Of particu-

lar interest would be comparing two time course (or spatial course) experiments. For example, we might be interested in a comparison between two time course experiments with different treatments. Under this type of experimental design, in addition to identifying genes following dynamic or sporadic expression paths within each experiment, we may also want to infer the DE/EE status at each time point across two experiments. We may adapt the model framework in Yuan and Kendziorski (2006), in which a hidden Markov model was implemented on the DE/EE status at each time point to accommodate dependence over the time course.

4 OSCOPE: A STATISTICAL PIPELINE FOR IDENTIFYING OSCILLATORY GENES USING UNSYNCHRONIZED SINGLE-CELL RNA-SEQ DATA

4.1 Background

Recent advances in single-cell RNA-seq technology enable investigators to conduct transcriptome-wide gene expression studies at the single-cell level. Several publications utilized single-cell RNA-seq to identify heterogeneity and subpopulations among cells (Wu et al., 2014; Brennecke et al., 2013), to detect differentially expressed genes (Kharchenko et al., 2014), and to improve temporal resolution by reconstructing a pseudotemproal order using selected cells sharing similar transcriptome profiles (Trapnell et al., 2014). Single-cell technology also provides the potential in identifying de novo oscillators or oscillatory systems, but no method has been developed for such intention yet. Here we propose a statistical pipeline called Oscope (oscilloscope) for identifying oscillatory gene sets in an unsynchronized cell population. Oscope is able to identify potential oscillatory genes along the whole transcriptome, group them into oscillatory groups, and then reconstruct base cycle profiles of genes in each group by reordering cells. Furthermore, Oscope is an unsupervised algorithm that does not require prior knowledge of key marker genes.

The most challenge part in discovering oscillators in a single-cell RNA-seq experiment is that we are not able to trace expression of a certain cell over time. Here we utilize co-regulation information among genes in an unsynchronized cell population to scan for potential oscillating genes and to infer gene expression paths based on grouped oscillators. Figure 4.1 (a)-(c) shows an illustration of a typical single-cell experiment on an unsynchronized population. As an example, we assume 3 genes from the same oscillatory group oscillate together with the same frequency. In Figure 4.1 (a), each panel shows expression profiles of these 3 genes in one cell. Due to the unsynchronization, cell $1, 2, ..., S$ started oscillating at

different time points $t_{0,1}, t_{0,2}, ..., t_{0,S}$. For example, $t_{0,s}$ may be cell division time for the daughter cell s.

In a single-cell experiment, cells are collected by taking a "snapshot" at time T. We define the lifetime of a cell s as the duration between the time that the cell started oscillating and the collecting time T. The lifetime can be written as $t_s = T - t_{0,s}$. We assume that for a certain gene, it oscillates following the same expression path along the lifetime in all cells (for example, same starting phase and same frequency in all cells). We also assume different genes from the same oscillatory group share the same frequency but may have different starting phases. As shown in Figure 4.1 (b), if we were able to sort cells by their lifetime, we may reconstruct gene expression paths along the lifetime since we collected cells with different lifetime. In that case, existing methods studying dynamic expression profiles may be used to detect oscillators and to infer their features. Unfortunately, lifetime information is unobserved in single-cell RNA-seq experiments. With the collecting order in the snapshot, gene expression of the oscillators looks as noisy as random noise (shown in Figure 4.1(c)). Therefore existing methods are not applicable here.

Under this circumstance, the only information preserved is co-regulation information among genes in the same oscillatory group. And we may use this information to recover the "lifetime" by reordering the cells. For example, in Figure 4.1 (b), if we observe a cell has low expression in gene 1, moderate expression in gene 2 and high expression in gene 3, then its lifetime is likely to be close to cell 2 or 4. Then we may put this cell to a position closed to cell 2 or 4 during the reordering. However, we are not able to tell if the lifetime of this cell is more close to t_2 or t_4. That is because cell 2 and 4 have the same observable profile since they "traveled" to the same relative position on the base cycle at time T. Here we define the base cycle as the minimal unit that is repeated in a periodic change (an example is shown in Figure 4.1 (b)). With the limited observable information, in Oscope we only focus on recovering a cell's relative position on a base cycle (which may be viewed as t_s mod period duration τ). Figure 4.1 (d) shows such reconstruction of example cells.

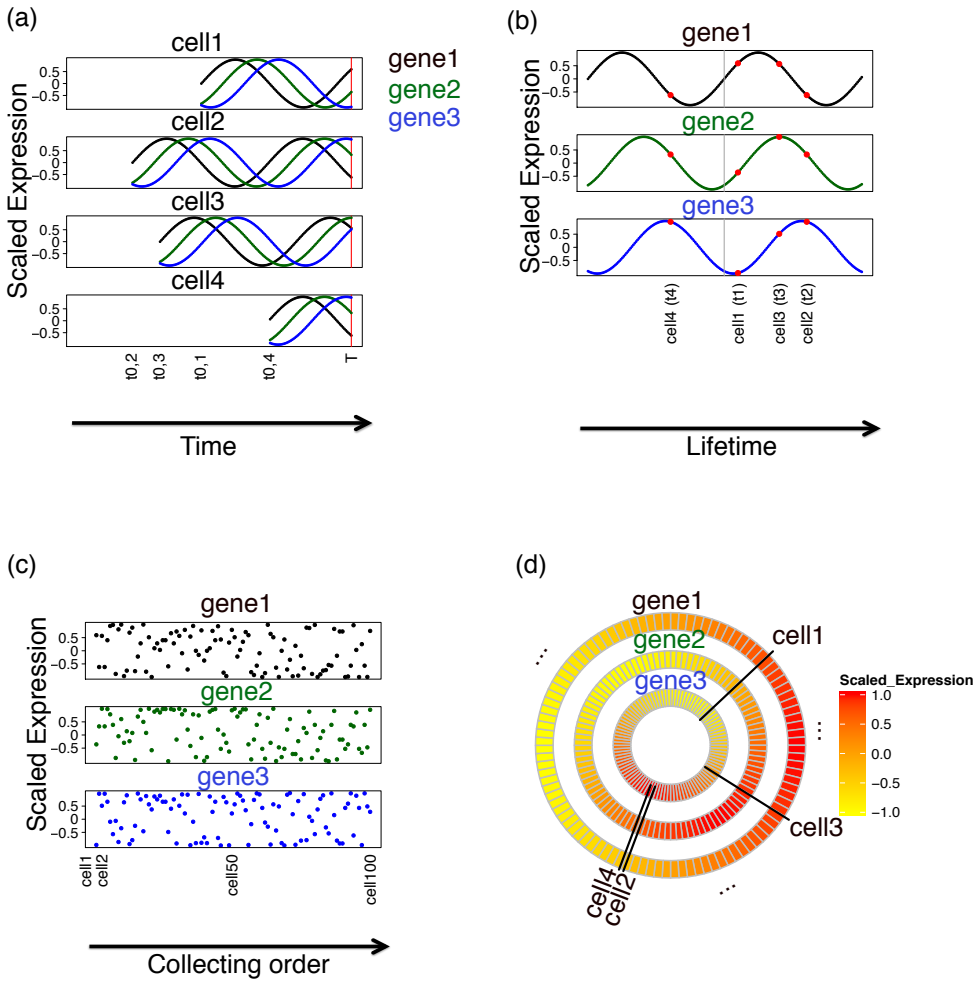


Figure 4.1: Panel (a) illustrates of a single-cell experiment on an unsynchronized cell population. The x axis shows the physical time and the y axis shows scaled gene expression (with mean 0 and standard deviation 1). Genes 1-3 are in the same oscillatory group. They oscillate with the same speed but with different starting phase. Cells 1-4 are the first 4 cells collected in the experiment; they start oscillating at different time point $t_{0,1}$, ..., $t_{0,4}$. In the single-cell experiment, cells are collected at the same time point T. Panel (b) shows the same genes and cells as in (a). Instead of calendar time, here the x axis shows the lifetime of each cell. For a given cell s, the lifetime is defined as $t_s = T - t_{0,s}$. Two base cycles are separated by the vertical gray line. Panel (c) shows the gene expression profiles of genes 1-3 based on the collecting order of an unsynchronized population of 100 cells. Panel (d) shows results of the base cycle reconstruction of the 100 cells shown in (c). Scaled expression is shown in gradient colors. Three loops here represent three genes. Cells are reordered and cells 1-4 are marked specifically.

4.2 Methods

Oscope pipeline

Figure 4.2 shows a flowchart of the Oscope pipeline. Here we developed a paired-sine model to calculate the likelihood of gene pairs to be oscillating together. If two genes both follow sinusoid process over time and with some phase shift, the paired expression profile from these two genes can be characterized by a certain form even if the lifetime is unobserved (see the next part of this Methods section). This paired-sine model picks gene pairs of best fit to this form. To infer oscillatory groups, we further cluster selected gene pairs into groups using a K-Medoid Algorithm.

After getting the oscillating groups, Oscope will reorder the cells based on expression of genes from the same group. We view the recovery of the base cycle profiles as a traveling salesman problem (TSP). The objective of TSP is searching for an optimal order of cities that constructs a path with shortest distance between adjacent cities, in which the longitude and latitude are unknown, instead, distance between each pair of cities is provided. In our problem, the objective is to search for an optimal order of cells that reconstructs smooth cyclic profiles for genes in one oscillatory group. Here the lifetime of each cell is unknown, instead, expression difference between each pair of cells is observed. Maximizing the smoothness of recovered cyclic profiles would also be viewed as minimizing expression differences between adjacent cells. To search for the optimal order that simultaneously recovers smooth base cycle profiles for all genes in a group, we developed an Extended Nearest Insertion (ENI) algorithm. The ENI algorithm extends the Nearest Insertion heuristics as proposed in Rosenkrantz et al. (1977). The ENI starts with a loop consisting of an arbitrary group of three cells, then chooses a cell not yet included in each step. This cell is inserted into the existing loop between two consecutive cells, such that the smoothness of expression profiles is maximized. The algorithms stop when all cells are included. Once the recovered order for each group is obtained, other algorithms developed for time course analysis may be applied. For example, identification of more (weaker) oscillatory genes, or inference of speed differences

across gene groups.

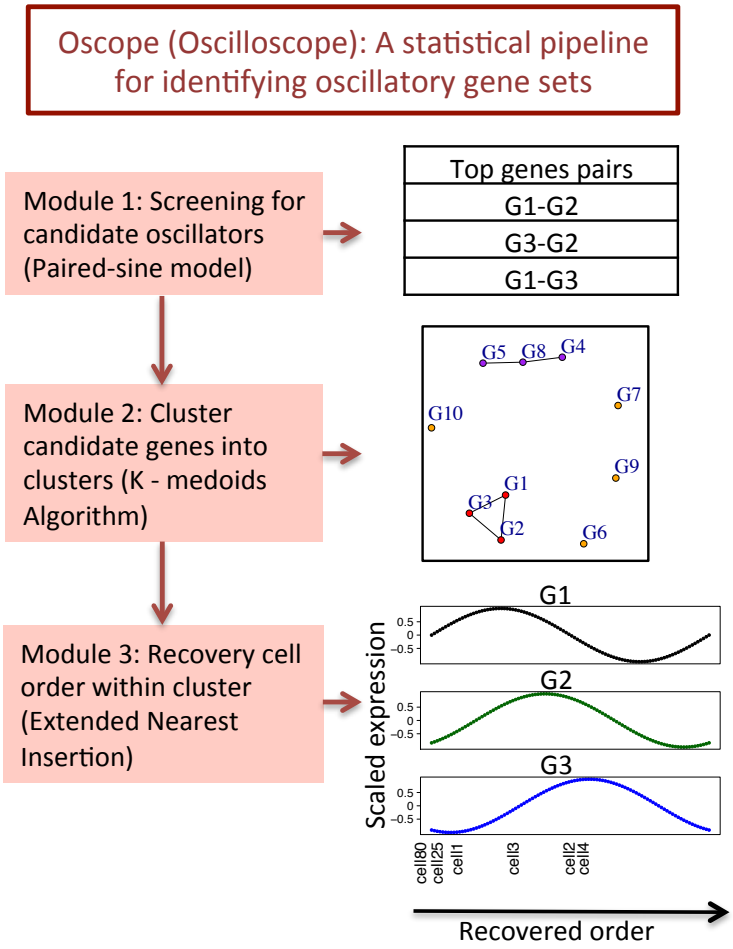


Figure 4.2: Shown is the workflow of the Oscope pipeline. Firstly, the paired-sine module selects candidate gene pairs with high likelihood of being oscillating together. Then a K-Medoid clustering module clusters candidate genes into oscillatory groups. Finally, the ENI module recovers the base cycle of each group by reordering the cells.

Oscope modules

Oscope: paired-sine model: Assume gene expressions are rescaled to range [-1, 1] for each gene. For each pair of genes X and Y, denote the matched gene expressions as (x_1, y_1), (x_2, y_2), ... , (x_S, y_S). For cell s in 1,...,S, if two genes follow the same sinusoid process and with some phase shifts (for example, genes in Figure 4.1), then the following equations hold:

$$x_s = sin(\tau_s); y_s = sin(\tau_s + \phi_{xy})$$

in which $\tau_s = t_s \bmod \tau$ indicates position of cell s on the base cycle, and ϕ_{xy} indicates amount of shift between two genes. ϕ_{xy} is shared by all cells and τ_s is cell specific. By trigonometric identities, the following equation holds:

$$y_s^2 + x_s^2 - 2y_s x_s cos(\phi_{xy}) - sin^2(\phi_{xy}) = 0 \tag{4.1}$$

which is obtained by $y_s = sin(\tau_s)cos(\phi_{xy}) + cos(\tau_s)sin(\phi_{xy}) = x_s cos(\phi_{xy}) \pm \sqrt{1 - x_s^2} sin(\phi_{xy})$. Define:

$$\epsilon_{xy} = y_s^2 + x_s^2 - 2y_s x_s cos(\phi_{xy}) - sin^2(\phi_{xy}) \tag{4.2}$$

If two genes follow the same sinusoid process, there exists an optimal ϕ_{xy} for which ϵ_{xy}^2 is close to 0. To search for gene pairs with associated dynamic changes, we estimate optimal ϕ_{xy} and calculate the associated ϵ_{xy}^2 for all gene pairs. The candidate oscillatory gene pairs are these with an optimal ϕ_{xy} that leads to small ϵ_{xy}^2.

Oscope: K-Medoid clustering: To identify oscillatory groups, we cluster candidate oscillators detected from the paired-sine model by a K-Medoid algorithm. ϵ_{xy}^2 is used as the dissimilarity metric. Therefore gene pairs with small ϵ_{xy}^2's are more likely to be clustered into the same cluster. We applied the K-Medriod algorithm with varying number of clusters k. The optimal k is picked by maximizing the Silouette distance.

Oscope: Extended Nearest Insertion: The ENI algorithm is developed to re-

cover optimal cell order for each oscillatory gene group defined in the K-Medoid clustering step. The optimal cell order recovered by ENI aimed at reconstructing smooth base cycle profiles for all genes in an oscillatory group.

We start with 3 randomly selected cells and form them as a loop. Then we randomly pick the 4th cell and insert it into 3 cell-cell gaps on the loop, this forms 3 candidate orders. We evaluate each order using aggregated mean squared error (MSE) of a sliding polynomial regression (SPR). For a given order, SPR is fitted on expressions within each gene. To capture the cyclic feature of the base cycle, SPR is defined as fitting m polynomial regression models starting with m evenly distributed points on the loop. The largest MSE among the m models is defined as the MSE of the SPR on this gene. For each order, the aggregated MSE of an oscillatory gene set is defined as the summation of the SPR's MSEs among all genes. The optimal order of first 4 cells is then selected as the one that minimizes the aggregated MSE. We repeat this process to insert the 5th cell and so on, until all cells are in the loop. Afterwards, the 2-opt algorithm is applied to avoid the local maxima.

4.3 Results

Simulation Setup

We conducted two simulation scenarios to evaluate Oscope. In **Sim I**, for a given oscillatory gene g in a cell with lifetime t, we simulated expression $X_{g,t}$ as a sinusoid signal following $\text{Sine}(\omega_g t + \varphi_g)+\epsilon_g$, in which ω_g is the angular frequency and φ_g is the starting phase of gene g; ϵ_g is Gaussian noise with mean 0 and standard deviation (SD) σ_g. To make the simulated data set more realistic, we conducted another simulation scenario **Sim II**. In **Sim II**, instead of a sinusoid signal, we simulated oscillatory genes based on profiles of cyclic genes identified in Whitfield data (Whitfield et al., 2002). Here $X_{g,t}$ is simulated as $\mu_{g,t} + \epsilon_g$. In which $\mu_{g,t}$ is obtained from imputed Whitfield data and ϵ_g is defined similarly as in **Sim I**.

In both simulation studies, we simulated 1000 genes and 100 cells. 90 out of

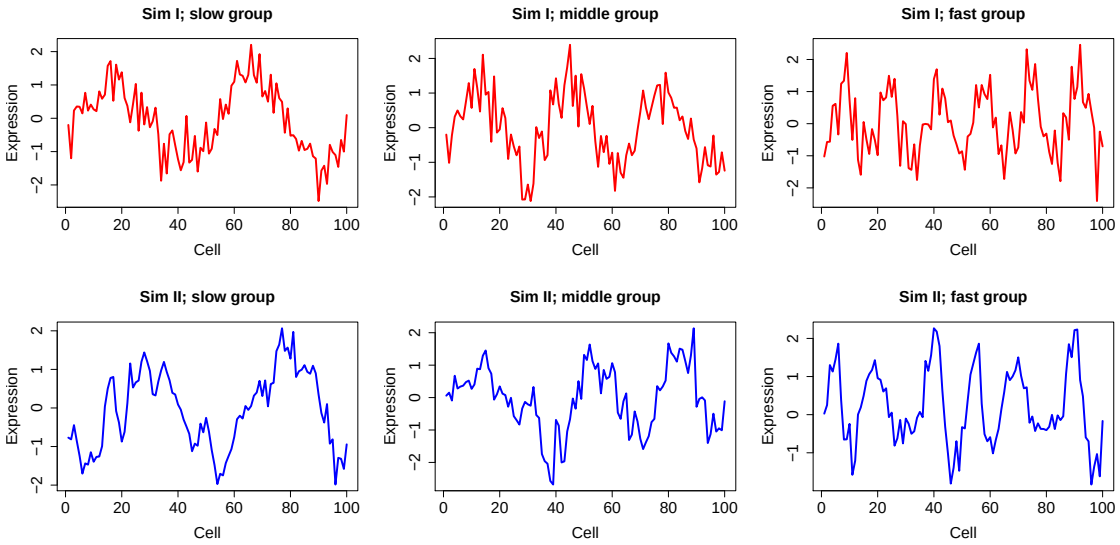


Figure 4.3: Shown are example oscillatory genes in 3 frequency groups in **Sim I** (upper panels) and **Sim II** (lower panels). The x axis shows original order along the simulated time course; the y axis shows simulated expression. (σ_g is defined as 0.2 for these genes).

the 1000 genes were simulated as oscillators. The 90 oscillators were simulated as 3 frequency groups, each group contains 30 genes. In **Sim I**, the relative speeds of the three groups are proportional to 6:3:2. Example genes can be found in the upper panels of Figure 4.3.

In **Sim II**, we implemented an imputation algorithm to extend 48 samples in Whitfield et al. (2002) to time course data with 100 cells. We took the expression profiles of the first 100 cyclic genes defined in the original paper as kernel signal in our imputation. Recall in Whitfield data, 48 samples were measured and oscillators have roughly 3 cycles over the time course. To simulate a gene in the fastest group, we randomly selected one of the top 100 genes, repeated the 48 measures twice to obtain $\mu_{g,1}, ..., \mu_{g,48}$ and $\mu_{g,51}, ..., \mu_{g,98}$. We then imputed $\mu_{g,49}$, $\mu_{g,50}$, $\mu_{g,99}$ and $\mu_{g,100}$ by taking average expression of $\mu_{g,48}$ and $\mu_{g,1}$ plus random noise. To simulate a gene in the group of median speed, we also randomly selected one gene from the top 100, then extended the 48 measures to 100 cells. We first obtained $\mu_{g,1}, \mu_{g,3}, ..., \mu_{g,99}$

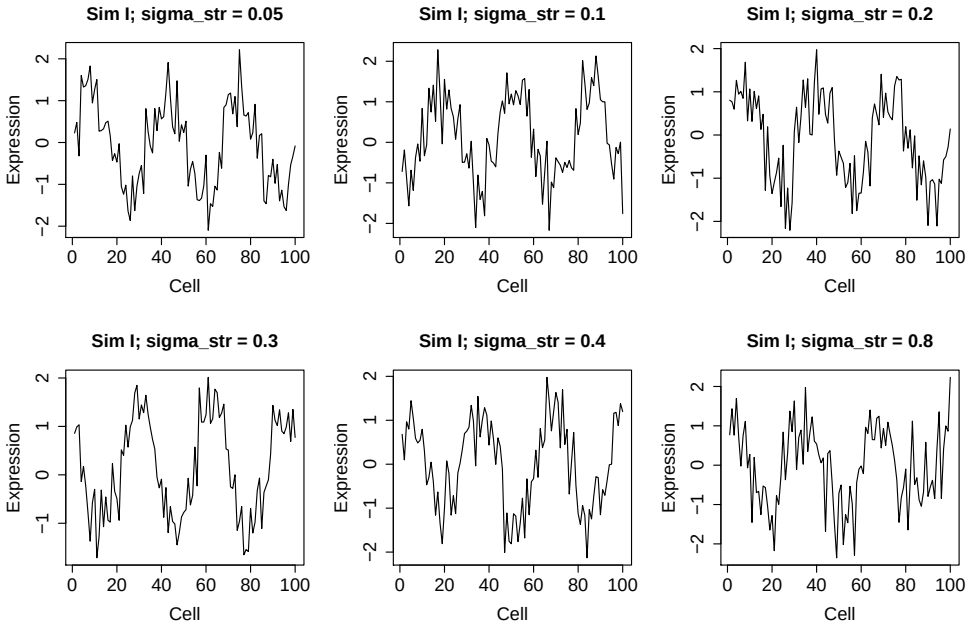


Figure 4.4: Shown are example oscillatory genes with varying noise levels under scenario **Sim I**.

by taking the 48 measures from Whitfield data (the first 2 measures were used twice to get $\mu_{g,97}$ and $\mu_{g,99}$). Then we generated imputed measures $\mu_{g,2}, \mu_{g,4}, ...$ The imputed measure was generated as averaged expression of its two adjacent cells plus random noise. For the slowest group, the first 33 samples from the original data of a selected gene were used and then extended to 100 cells in a similar way. As a result, three groups of genes have approximately 6, 3 and 2 cycles respectively. Example genes can be found at lower panels of Figure 4.3.

Within each frequency group, genes were further simulated with strong and weak signals. Half of the oscillatory genes were simulated as strong oscillators with $\sigma_g = \sigma_{str}$. The other half were simulated as weak oscillators with $\sigma_g = \sigma_{wk} = 2 * \sigma_{str}$. Starting phase φ_g varies in different genes within a frequency group. The remaining genes except the oscillators are called noise genes. Noise genes were simulated as random Gaussian noise. The noise level was adjusted to be comparable to the average noise signal among all oscillators.

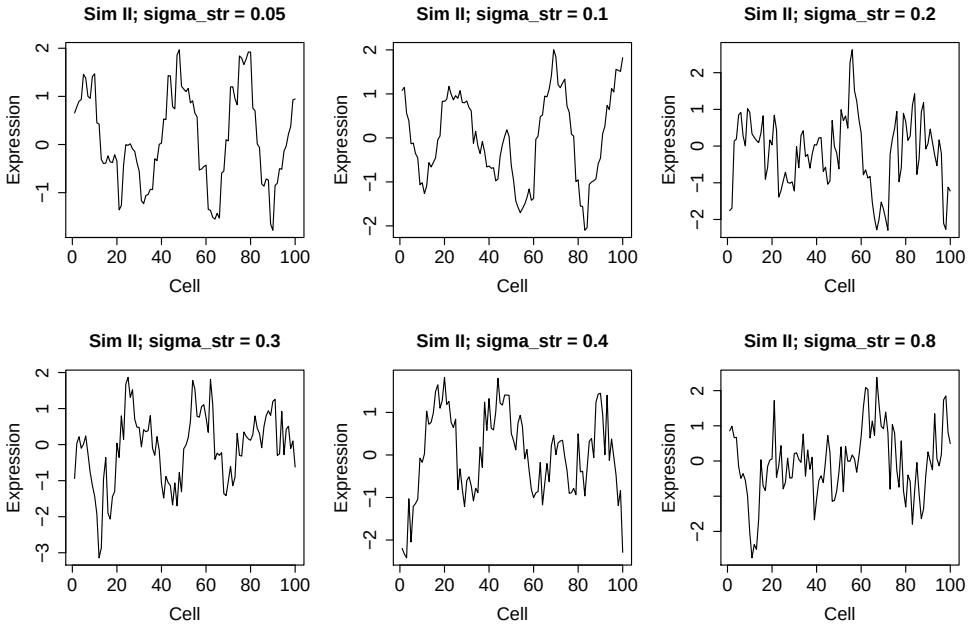


Figure 4.5: Shown are example oscillatory genes with varying noise levels under scenario **Sim II**.

Evaluation

To evaluate Oscope, we ran 6 simulation studies with varying noise levels for each simulation scenario. The σ_{str} varies from 0.05 to 0.8 in 6 simulation studies. Note that the noise level in the weak oscillatory group is always defined as $\sigma_{wk} = 2 * \sigma_{str}$. For each simulation study, 10 simulations were ran. Figure 4.4 shows example oscillators with varying noise levels under scenario **Sim I**. Similar trace plots for **Sim II** are shown in Figure 4.5.

We applied Oscope on each of the simulated data sets, the top 10% of the genes were selected from the paired-sine model and further clustered into oscillatory groups by the K-Medoid algorithm. To evaluate the paired-sine model, we consider two summary statistics:

TPR: $\frac{\text{Num True Positive genes detected by Oscope}}{\text{Num genes simulated as oscillator}}$

FDR: $\frac{\text{Num False Positive genes detected by Oscope}}{\text{Num genes detected by Oscope}}$

To evaluate the K-Medoid algorithm, we call a group a True Positive if more than half of the genes in this group were simulated as oscillating. Then we consider:

GP (Group-wise precision): $\frac{\text{Num True Positive groups detected by Oscope}}{\text{Num groups detected by Oscope}}$

The ENI algorithm was further applied to gene groups defined by the K-Medoid clustering algorithm. To evaluate the ENI recovered order, we consider statistics defined as follows. Recall in **Sim I**, gene expression $X_{g,t}$ is simulated following Sine($\omega_g t + \varphi_g$)+ϵ_g. Therefore, for a gene with angular frequency ω_g, the base cycle position of a sample at time t can be calculated from ($\omega_g t \bmod 2\pi$). For each frequency group, we split samples into 5 bins based on their base cycle positions. Figure 4.6 (a) shows bin specification on a simulated data set with $\sigma_{str} = 0.3$. Shown are 6 genes from three frequency groups. For each frequency group, samples in 5 bins are shown in different colors. Note the bin specifications are different across different frequency groups. If the base cycle profile was reconstructed successfully based on the ENI recovered order, the five bins should be separated clearly.

To assess this, we evaluated the recovered order by calculating the number of jumps. For a given gene in a recovered cell order, a jump is called if the bin classification of a sample is different from the previous sample. Therefore, a perfectly recovered order would give very few jumps, whereas a completely incorrect recovery will give ~ 99 jumps in our simulation with 100 samples (for example, based on our input order from random permutation, the median number of jumps across all simulated data sets is 79). Using the same data set as in Figure 4.6 (a), Figure 4.6 (b) shows recovered base cycle profiles of the 6 genes. The K-Medoid algorithm grouped the potential oscillatory genes into two clusters. Genes 1 and 2 were classified into the first cluster and the other 4 were classified into the second cluster (other genes in these clusters are not shown here). We ran ENI on two clusters separately. Figure 4.6 (b) shows these 6 genes in recovered order. Note when genes with multiple speeds are present in one cluster, ENI will recover one base cycle for the slowest oscillatory gene(s). So for a given cluster, its reference bin specification is defined as the bin specification based on the slowest gene in this cluster. In Figure 4.6 (b), each sample is marked with the color representing its bin assignment from the ground truth of its reference bin specification. Recovered

orders of clusters 1 and 2 have 22 and 37 jumps respectively. Note in cluster 2, the recovered base cycle profile successfully reconstructed one cycle for frequency group 2 and two cycles for frequency group 3. Figure 4.6 (c) shows step plots of the sample's ground truth bin assignment. Samples are shown in recovered order. It indicates that most jumps happened at transition points.

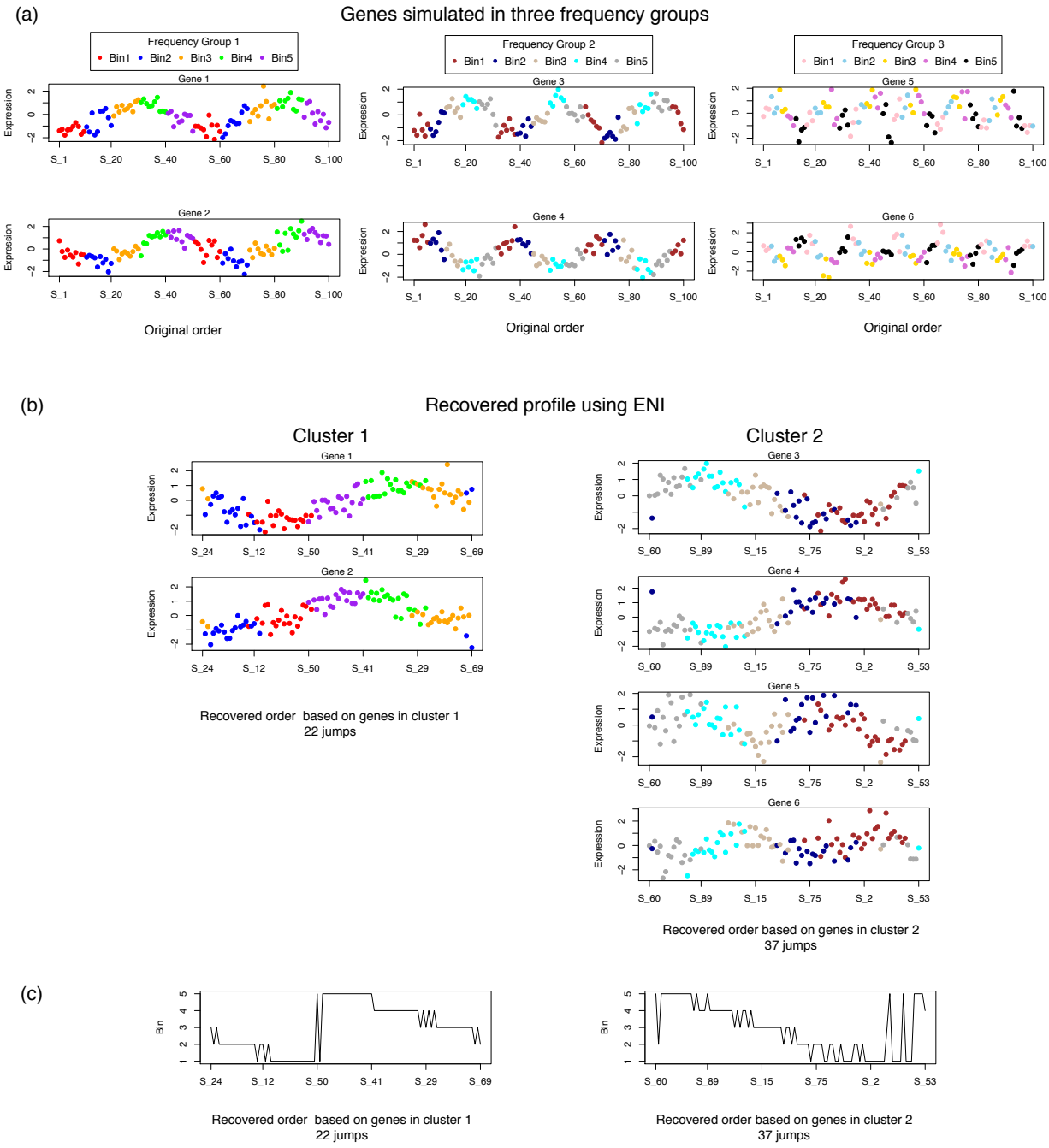


Figure 4.6: (a) Shown are 6 genes from one data set simulated following **Sim I** with $\sigma_{str} = 0.3$. Samples from different bins are shown in different color. (b) Shown are ENI recovered base cycle profiles of 6 genes in (a). Samples are colored by their bin assignments in their reference bin specification. (c) Shown are ground truth bin assignments of samples. Samples are shown in the ENI recovered order.

Simulation Results

Figures 4.7 and 4.8 show operating characteristics evaluating Oscope on simulation studies under scenario **Sim I** and **Sim II**, respectively.

Under scenario **Sim I** in which oscillators were simulated following sinusoid signal, Oscope has high TPR/GP and well controlled FDR unless the noise level is very high ($\sigma_{\text{str}} = 0.8$). Under scenario **Sim II**, Oscope has high TPR/GP and well controlled FDR when σ_{str} is lower than 0.2.

We also evaluated the ENI results on all **Sim I** data sets. Figure 4.9 shows the number of jumps based on the recovered order. Based on the input order of the simulated data set, the median number of jumps is 79. Using the ENI recovered orders, the number of jumps are controlled below 30 unless the noise level is very high ($\sigma_{\text{str}} = 0.8$).

In **Sim II**, expression is imputed from empirical data. It is very challenging to define ground truth of base cycle position for the samples. As a result, the ENI evaluation is not applied on **Sim II** data sets.

Case study results

In case study analyses, we first consider a microarray data set of a time course experiment studying cell cycle genes (Whitfield et al., 2002) in the HeLa cell line. In this data set, gene expression of a synchronized cell population were measured over 48 hours. To mimic the collecting order in a single-cell experiment (as shown in Figure 4.1 (c)), we randomly permuted the sample order in the time course data prior to applying Oscope. Paired-sine module in Oscope identified an oscillatory gene group with 69 genes, in which 65 out of 69 genes are also claimed as oscillatory genes in Whitfield et al. (2002). 44 out of 69 are also in the top 69 identified in the original paper. Figure 4.10 (a) and (b) show 4 example genes identified by Oscope. In Figure 4.10 (a), the samples are shown using the ENI module recovered order. While in Figure 4.10 (b) the samples are shown following the original time course order. Gray lines show peak of each gene in the Oscope recovered base cycle profile or in the first cycle along the original time course. Comparing Figure 4.10 (a) with

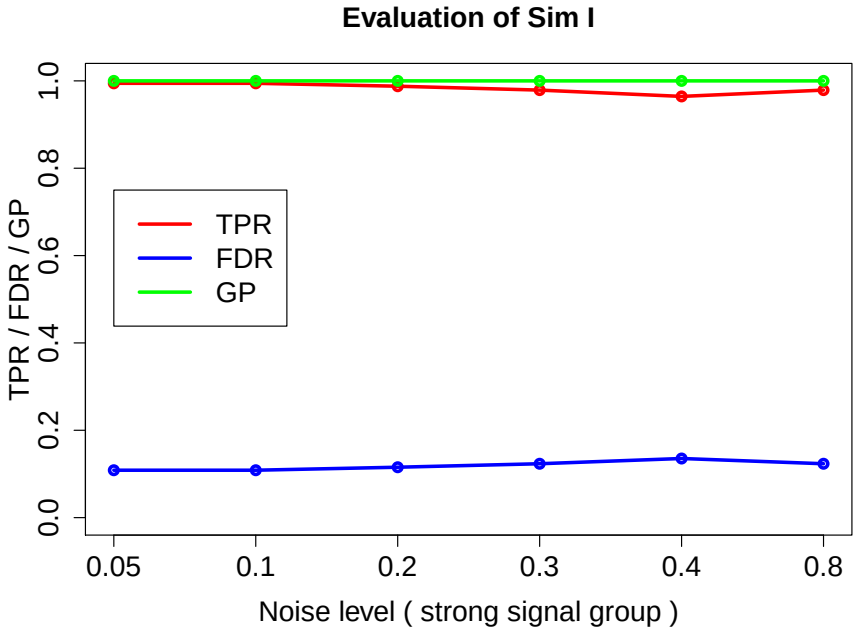


Figure 4.7: Shown are evaluation results from 6 simulation studies with varying noise levels under **Sim I** scenario. We ran 10 repeated simulations for each study. The y axis shows TPR, FDR and GP using Oscope averaged across 10 repeats. The x axis shows σ_{str} defined in each study.

(b), the optimal order reconstructed by Oscope successfully recovered the base cycle of each gene. In addition, peak shifts across different genes are correctly inferred using Oscope recovered order. Comparison between recovered order and original order for all 69 genes may be found at Appendix Figure C.3.

We then examined Oscope on a single-cell RNA-seq data set. 73 cells were collected from H1 ESCs. Among the top 3 clusters identified by Oscope, one cluster with 32 genes is found to be enriched by cell cycle (CC) pathway (22 out of 32 are in GO CC pathway). Figure 4.10 (c) shows expression of 4 example genes in the CC cluster, cells are sorted by ENI using all 32 genes. Peak shift was observed across two groups of genes, CENPE and BUB3 vs. DLGAP5 and CCNB1. To further validate the recovered base cycle profile and observed peak shift, we looked at these

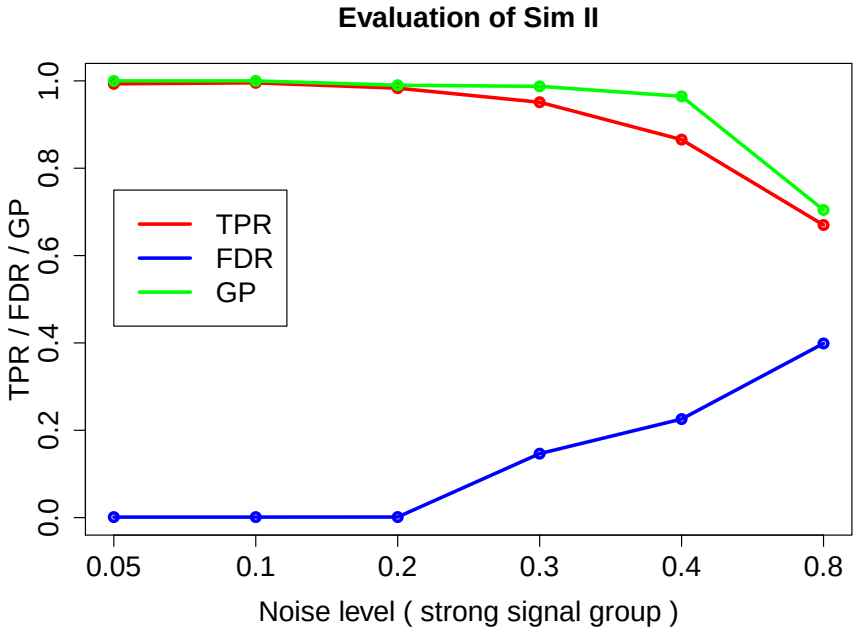


Figure 4.8: Shown are evaluation results from 6 simulation studies with varying noise levels under **Sim II** scenario. We ran 10 repeated simulations for each study. The y axis shows TPR, FDR and GP using Oscope averaged across 10 repeats. The x axis shows σ_{str} defined in each study.

32 genes in H1-FUCCI data set generated by the Fluorescence Ubiquitination Cell-cycle Indicator system. In H1-FUCCI data set, cells from G1, S or G2/M phase are sorted apart and performed single-cell RNA-seq separately. For testing purpose, we ignored cells' CC phase assignments and randomly permuted the cell order prior to applying Oscope. We applied ENI using the 32 genes on permuted H1-FUCCI data. Figure 4.10 (d) shows the 4 example genes on H1-FUCCI data, cells are sorted by ENI. To evaluate the sorting results, in Figure 4.10 (d), we color cells with three different colors based on their experimental CC phase assignments. The recovered order from ENI successfully seperated three CC phases with very few misclassified cells. And the peak shift across two groups of genes is agreed with those suggested in Figure 4.10 (c). Results of all 32 genes may be found at Appendix Figure C.5

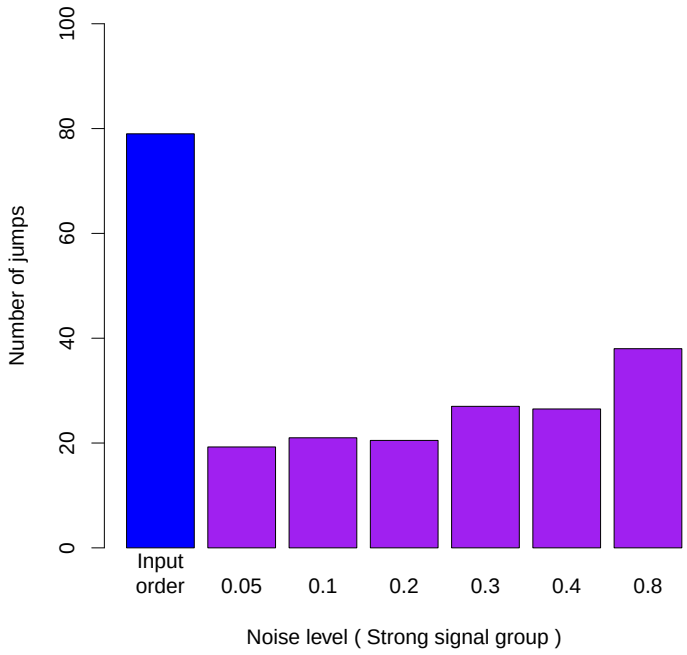


Figure 4.9: Shown are ENI evaluation results from 6 simulation studies with varying noise levels under **Sim I** scenario. We ran 10 repeated simulations for each study. The y axis shows median number of jumps based on the recovered order of each cluster averaged across 10 runs. The x axis shows σ_{str} defined in each study. The first bar shows median number of jumps across all simulated data sets using input order from random permutation.

We further explored other gene groups identified by Oscope. When doing so, we found that genes from one of the top clusters has an artificial trend that may be related to an environmental effect on the Fluidigm C1 chip. These genes all have high expression in cells collected from spots with small or large IDs on the chip. We call this "ordering effect". We further repeated the experiment twice to examine the ordering effect. Figure 4.10 (e) shows 4 example genes with such effect across 3 experiments. The trend is very reproducible in repeated experiments. To further justify the artifact, we conduct qPCR on 2 genes, PFN1 and MIF, using

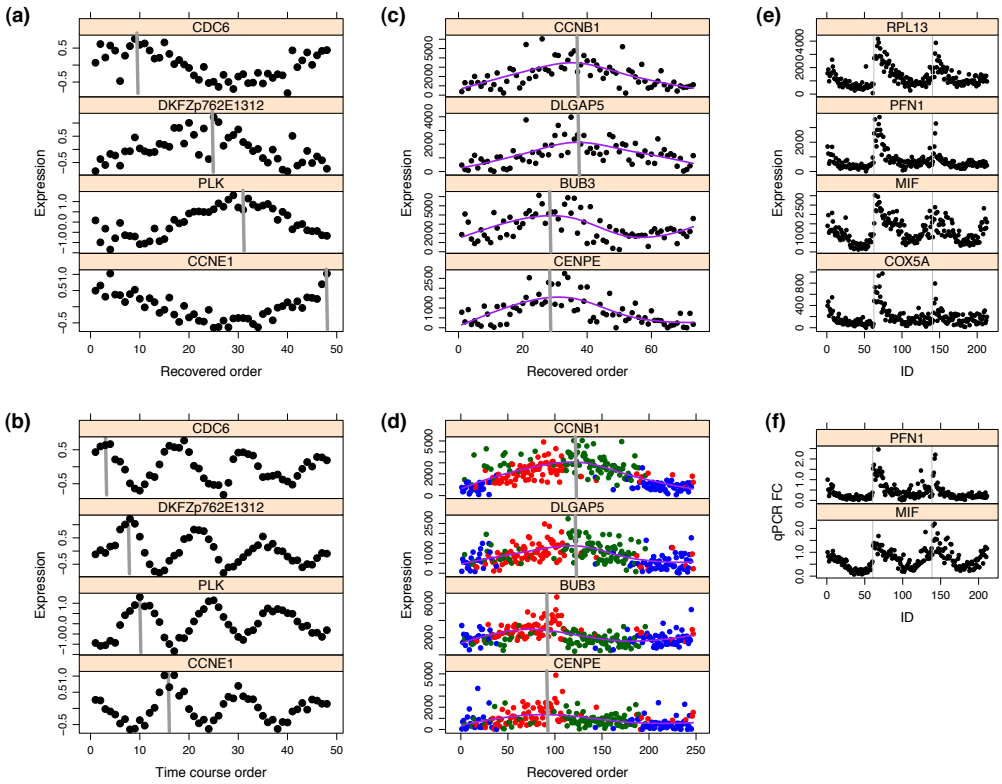


Figure 4.10: (a) Shown are 4 genes identified by Oscope on Whitfield data. The y axis shows expression from microarrays data. On the x axis, samples are shown following the order recovered by ENI using 69 genes identified by paired-sine model. Peak of base cycle is marked with a gray line. (b) Shown are the same set of genes as in (a). Here samples are shown following the original order over time. Peak of the first cycle is marked with a gray line. (c) Shown are 4 genes identified by Oscope on single-cell RNA-seq data of H1 cells. The y axis shows expression after adjusting for sequence depth. On the x axis, samples are shown following the order recovered by ENI using 32 genes identified by the paired-sine model. The peak of the base cycle is marked with a gray line. (d) Shown are the same set of genes as in (c). Sample order was recovered by ENI using the 32 genes on FUCCI data. The phase assignments were ignored while applying ENI. Samples from S, G2 or G1 are marked as blue, red or green on the figure after applying ENI. (e) Shown are 4 genes with potential ordering effects that were identified by Oscope. The y axis shows expression after adjusting for sequence depth. Three experiments of H1 data are separated by gray lines. Samples are shown following the collecting order on the x axis. (f) Shown are qPCR results on genes PFN1 and MIF. The y axis shows fold changes obtained from qPCR measures. Three experiments of H1 data are separated by gray lines. Samples are shown following the collecting order on the x axis.

the original single-cell cDNA libraries prepared for sequencing. Figure 4.10 (f) shows qPCR results of these 2 genes across 3 experiments. The artificial trend is present in qPCR data as well. We suspect that the ordering effect is possibly due to environmental variation in different spots on the chip (More details may be found at Appendix Figure C.2). We also checked the ordering effect in two public-available data sets (Trapnell et al., 2014; Wu et al., 2014) and another in-house data set. The trend occurs as well and results can be found at Appendix Figure C.1. This finding further proved that Oscope is able to identify gene groups with associated dynamic profiles.

We showed that Oscope is able to uncover oscillatory gene networks in unsynchronized single-cell RNA-seq experiment. We anticipate it would be useful in identifying undocumented oscillators and de novo oscillatory systems.

4.4 Implementation

Oscope is implemented as an R package (R/Oscope). R/Oscope takes expression estimates, TPM, RPM, FPKM or RPKM. To estimate library sizes, R/Oscope defaults to median normalization (Anders and Huber, 2010). R/Oscope also provides a function that can robustly rescale estimates to values with the range of -1 and 1 for each gene. For a certain gene, the function will push the top and bottom α% values to -1 and 1 respectively. The threshold is defined to be 5% and can be changed by a user. This step will be able to reduce the impact of outliers.

The output of R/Oscope will be (1) a sorted list of gene pairs detected by the paired-sine model; (2) a sorted list of gene clusters defined by the K-Medoid clustering step; and (3) the recovered cell order based on each gene cluster. R/Oscope also provides a fast Fourier transformation implementation to search for more oscillatory genes based on the recovered cell orders.

4.5 Discussion and future work

We have developed a statistical pipeline called Oscope for identifying oscillatory gene sets using single-cell RNA-seq data on an unsynchronized population. Oscope may be used to identify genes following the same oscillatory process and recover base cycle gene expression profile of these genes. Based on the recovered cell order, we are able to search for more (weaker) oscillatory genes using standard time series approaches, for example, the discrete Fourier transformation or spline fitting.

The paired-sine model implemented in our pipeline is only focused on the oscillators following a sinusoid signal. Therefore, oscillators with other profiles may be missed (for example, genes with constant expression that burst periodically). We would like to extend our pipeline with a more generalized model to accommodate other oscillating profiles.

In this chapter, we also detected the ordering effect in multiple independent single-cell RNA-seq data sets. We suspect it is due to an environmental effect on the Fluidigm C1 chip. As a future work, we would like to understand the cause of the effect more thoroughly. Also, we would like to develop a statistical method to detect and adjust for such effect. The method may adapt algorithms on spline fitting or piecewise regression. We believe such method will reduce the artificial signals in the single-cell RNA-seq data to ensure trustable downstream analysis results.

There are also a number of extensions we may consider for single-cell RNA-seq data analyses. In particular, we are interested in extending EBSeq and EBSeq-HMM to accommodate DE analysis and time (spatial) course analysis on single-cell populations, at the gene and isoform level. In a bulk RNA-seq data set, most of the methods assume that samples from the same biological condition are from one population with similar characteristics. Under this assumption, for a given gene, statisticians always model expression in the same condition using a unimodal distribution. However, at the single-cell level, such assumptions may be violated by the intrinsic biological variability between cells. Even within a biological condition, the existence of sub-populations of cells is expected and cells from different sub-

populations may have distinct expression profiles. Due to this reason, statistical methods designed for bulk RNA-seq data are not applicable on single-cell data. Therefore, we would like to extend EBSeq and EBSeq-HMM by accommodating the heterogeneity in single-cell data.

A APPENDIX OF "EBSEQ: AN EMPIRICAL BAYES MODEL FOR IDENTIFYING DE GENES AND ISOFORMS"

A.1 Data sets used in the main text

Thomson lab data; ESCs vs. iPSCs

We analyzed RNA-seq data from the James Thomson Lab at the Morgridge Institute for Research. Details on the samples are given in Phanstiel et al. (2011); the particular samples considered here as well as alignment and expression estimation vary from that reported in Phanstiel et al. (2011) as follows. We evaluate RNA-seq reads from ES cell lines H1, H7, H9 and H14 and iPS cell lines DF4.7, DF6.9, DF19.7 and DF19.11. We filter 42-base-pair reads to remove adapters in each lane. To obtain gene counts via HTSeq, reads were aligned to the human RefSeq Hg18 transcripts using Bowtie and TopHat, allowing for no multiple matches and two mismatches. HTSeq was then applied to obtain gene counts for 18,780 genes, in which 15,671 were expressed. To obtain estimates of gene and isoform expression via Cufflinks, Bowtie and Tophat were applied allowing for up to 20 multiple matches and two mismatches. Expression was then estimated using Cufflinks.

MicroArray Quality Control (MAQC) data

The raw read files (fasta format) were downloaded from SRA SRX016359 and SRX016367. As part of the MAQC project, RNA was extracted from one sample of human brain tissue (HBR) and one sample of mixtures of tissues (UHR); seven replicates from each sample are considered here. To obtain gene counts using HTSeq (Anders, 2012), reads were aligned to the human RefSeq Hg18 transcripts using Bowtie (Langmead et al., 2010) and TopHat (Trapnell et al., 2009), allowing for no multiple matches (HTSeq requires that multi-reads are discarded) and two mismatches. HTSeq was applied to obtain gene counts for 18,780 genes, in which 16,518 were expressed (with median expression greater than 0). To obtain estimates

of expression via Cufflinks (Trapnell et al., 2010), Bowtie and Tophat were applied allowing for up to 20 multiple matches and two mismatches. Expression was then estimated using Cufflinks for 18,780 genes and 30,802 isoforms, in which 17,152 genes and 26,210 isoforms were expressed.

Gould lab data

RNA-seq data was obtained from two groups of congenic rats (four samples in each condition) harboring the susceptible or resistance allele of the mammary carcinoma susceptibility locus (Mcs1a) (Haag et al., 2003). For these experiments, mammary glands are taken from 8 untreated, mammary cancer-free females per genotype. The tissue is disaggregated using physical shearing in a solution of Tri-reagent (Ambion). RNA is extracted using a total RNA extraction kit (Ambion). RNA integrity is monitored using a 2100 Bioanalyzer (Agilent). Equal RNA (approximately 5 μg) from 2 rats is pooled to obtain a single sample for one RNA-seq lane. A total of four samples per genotype (*Mcs1a* susceptible or resistant) were processed by the University of Wisconsin Biotechnology Gene Expression Center using the Illumina Genome Analyzer IIx. Reads are post-processed to a length of 30 basepairs and aligned to the rat Ensemble RGS3.4 transcripts using Bowtie (Langmead et al., 2010), allowing for up to 100 multiple matches and one mismatch with seed length 30. Expression is estimated using RSEM (Li et al., 2010; Li and Dewey, 2011).

Smith lab data

Tophat output files were downloaded from GEO GSM792454-61. Eight samples (4 in each of two conditions) are considered here. In short, RNA was extracted from atrial tissue samples and prepared using Illumina's mRNA protocol. The reads are single-end with read length 36-bp. Each sample was run on one lane of an Illumina Genome Analyzer IIx. Alignment was done using Bowtie and TopHat (without *de novo* transcript detection) with the hg19 RefSeq annotation. Isoform expression was estimated using Cufflinks (Trapnell et al., 2010).

A.2 Assessment of the I_g effect in multiple data sets

We evaluate differences among I_g groups in multiple single-end and paired-end data sets processed under different priming protocols, in different labs, using different isoform expression estimation methods, using different definitions of isoform complexity, and for a wide range of sample sizes (from four to sixty-nine).

Appendix Figure A.1 shows data from James Thomson's lab at the Morgridge Institute for Research at UW-Madison. The data sets are distinct from those shown in the manuscript. For these experiments, RNA was extracted from human embryonic stem cell line H1 and prepared using the *Illumina TrueSeq*, T7LA(Sengupta et al., 2010), and the MinAmp (Thomson Lab internal) protocols, respectively. For each protocol, three samples were considered. Each sample was run on one lane of an Illumina Genome Analyzer IIx; the reads are single-end with read length 42-bp. Alignment was done using Bowtie with the hg18 RefSeq annotation. Isoform expression was estimated using RSEM.

Appendix Figures A.2a and A.2e show data from Michael Gould's lab at UW-Madison.

Appendix Figures A.2b and A.2f show data from the Wold lab (Mortazavi et al., 2008). RNA was extracted from mouse brain tissue and two replicates were prepared using the Solexa protocol. For each replicate, random primers were used. The reads are single-end with read length 25-bp. Alignment was done using Bowtie and Tophat (without *de novo* transcript detection) with the UCSC mm9 annotation. Isoform expression was estimated using Cufflinks with multi-read correction.

Appendix Figures A.2c and A.2g show data from the MicroArray Quality Control (MAQC) experiment (Consortium, 2006) that is distinct from what is shown in the manuscript. For these figures, raw read files (fasta format) were downloaded from GEO GSM475204-09. RNA was extracted from human brain tissue and 3 replicates were considered. For each replicate, random primers were used. The reads are paired-end with read length 50-bp. Each sample was run on one lane of an Illumina Genome Analyzer IIx. Alignment was done using SeqMap (Jiang and Wing, 2008) with the hg18 RefSeq annotation. Isoform expression was estimated

using RSeq (Jiang and Wing, 2009).

Appendix Figures A.2d and A.2h show data from Pickrell et al. (2010). RNA was extracted from Yoruba Hapmap cell lines and 69 samples were prepared using the Illumina Genome Analyzer II. For each replicate, random primers were used. The reads are single-end with read length 35-bp. Raw read files (fasta format) were downloaded from http://eqtl.uchicago.edu. Only Yale data are used and for the subjects assayed twice only the first replicate is used. Alignment was done using Bowtie with the hg18 annotation. Isoform expression was estimated using RSEM with multi-read correction.

Appendix Figure A.3 shows results using an alternative method to define isoform complexity. Instead of I_g as defined in the manuscript, the unmappability score of each isoform is obtained from RSEM, and the unmappability scores are clustered to group isoforms. Panel (a) shows the results from K-means clustering with 3 centers; panel (b) shows the results from a Gaussian Mixture Model.

Recall that Figure 2.1 shows spline fits which are similar to the approaches used by DESeq and edgeR to estimate variance. Appendix Figure A.4 shows the exact estimators used in DESeq, and edgeR (both the common-dispersion model and the tag-wise-dispersion model) derived using the data from the ESCs vs iPSCs experiment that is shown in Figure 2.1.

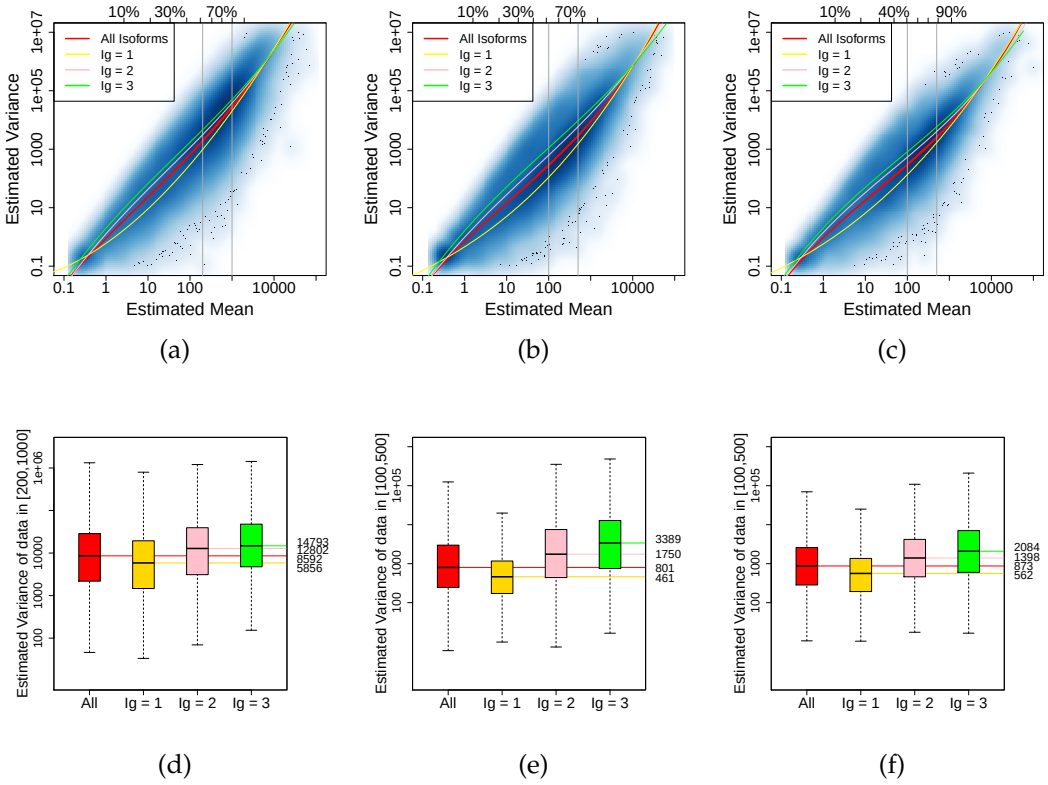


Figure A.1:

Panel (a) shows the empirical variance vs. mean for each isoform profiled in the experiment comparing ESCs with iPSCs (TrueSeq Protocol); details of this experiment are given earlier in this Appendix. A spline fit to all isoforms is shown in red with splines fit within the $I_g = 1$, $I_g = 2$, and $I_g = 3$ isoform groups shown in yellow, pink, and green, respectively. Panels (b) and (c) are similar to (a), but for data processed under the T7LA and MinAmp protocols, respectively. The estimated variance of isoforms with average expression in 50^{th} and 80^{th} percentiles of expressions are shown in (d), (e), (f).

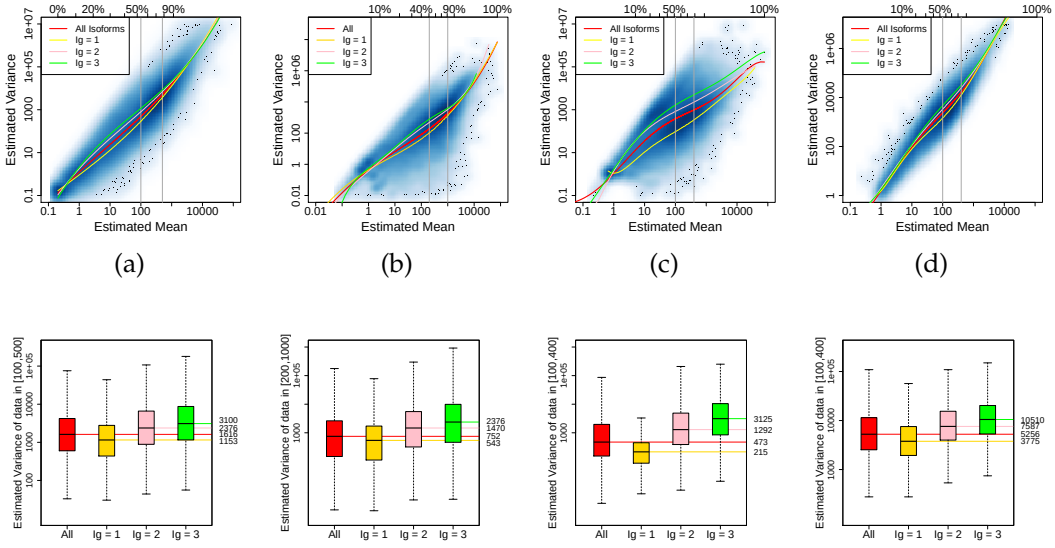
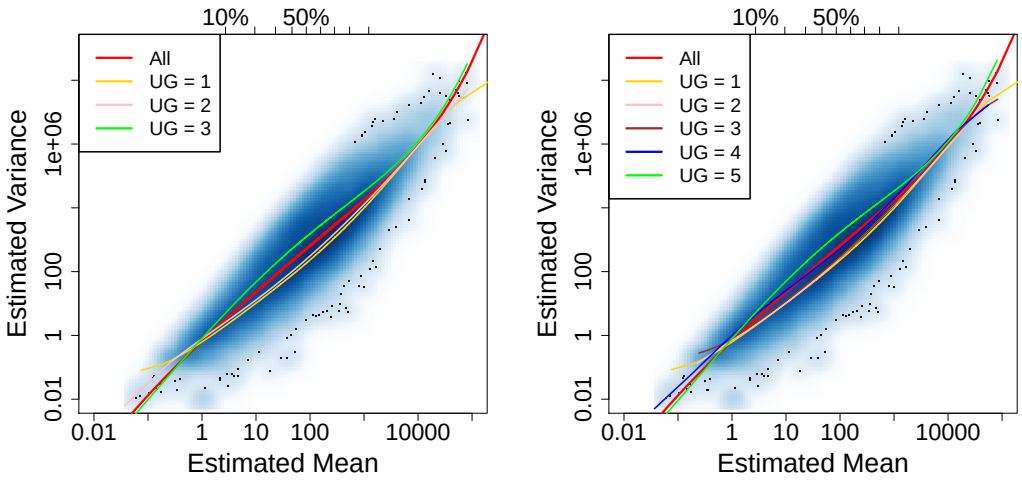


Figure A.2:

Shown are plots similar to Appendix Figure A.1 generated using data from the Gould lab (panel (a)) processed by RSEM, data from Wold lab (panel (b)) processed by Cufflinks, , MAQC data from Wong lab (panel(c)) processed by RSeq and data from Pickrell *et al.* (panel (d)); details of these experiments are given earlier in this Appendix. The estimated variance of isoforms with average expression in the 50^{th} to 80^{th} percentiles of expressions are shown in (e), (f), (g), (h).



(a) (b)

Figure A.3:

Shown are plots similar to Figure A.1(c), but with uncertainty groups obtained by K-means clustering of unmappability scores instead of I_g groups. The unmappability score of each isoforms as well as the isoform expected counts are obtained from RSEM. Panel (a) shows the results using K-means clustering with 3 centers. Panel (b) shows the results using a Gaussian Mixture Model.

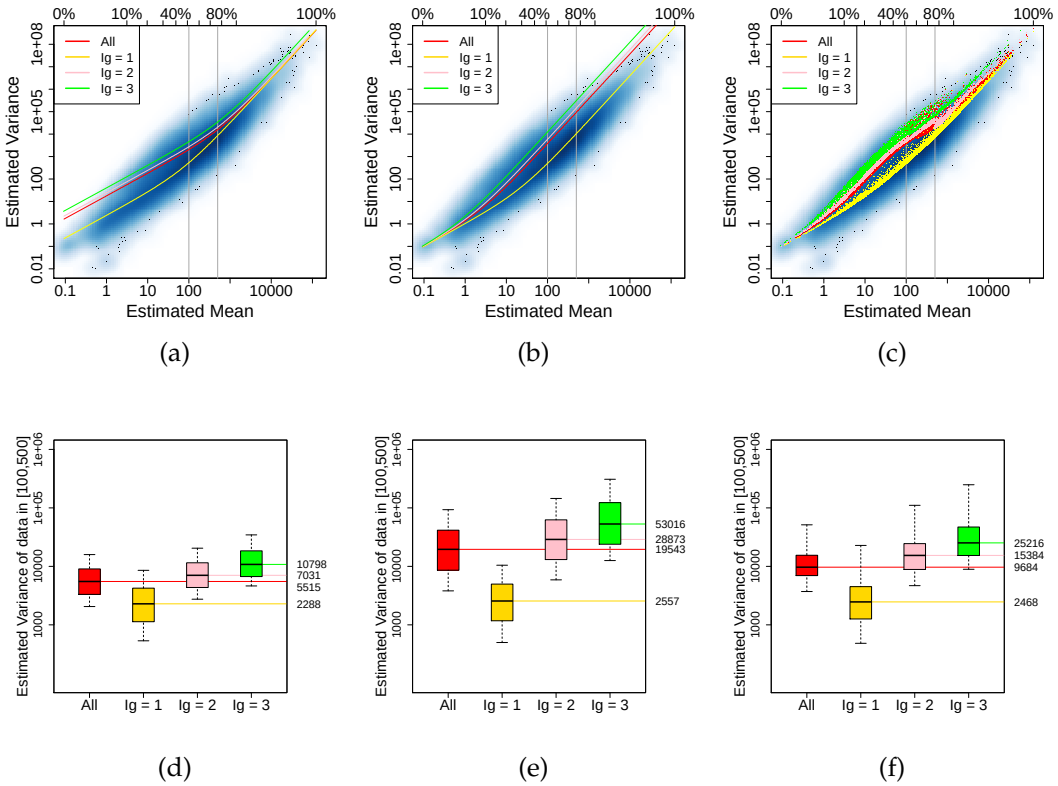


Figure A.4:

Recall that Figure 2.1(c) shows spline fits which are similar to the approaches used by DESeq and edgeR to estimate variance. This figure shows the exact estimators used in DESeq, and edgeR (both the common-dispersion model and the tag-wise-dispersion model) derived using the data from the ESCs vs iPSCs experiment that is shown in Figure 2.1(c). Specifically, panel (a) shows the fitted dispersion values provided by DESeq. The dispersion line is calculated across all isoforms (red) and within I_g group (shown in yellow, pink, and green, respectively). Panels (b) and (c) show similar plots from edgeR under their common dispersion (panel (b)) and tag-wise dispersion (panel (c)) models. Panels (d), (e) and (f) consider average expression in [100, 500]. The range was chosen as it approximates the 50^{th} and 80^{th} percentiles of expression across all isoforms. Shown are box-plots of the variances of these isoforms collectively, and within I_g group.

A.3 Comparison of features in simulated data vs. case study data

Appendix Figure A.5 demonstrates that characteristics observed in the case study data are reproduced in the simulated data sets.

Table 2.3 in the main text reports that count-based methods have well-controlled FDR. Appendix Figure A.6 shows that the likelihood of a false call increases in the presence of outliers, especially for edgeR. In particular, Appendix Figure A.6, panels (a) and (b) evaluate the operating characteristics shown in Table 2.3 within subsets of genes grouped by their Dixon's Q-statistic (Dixon, 1950). A gene harboring an outlier will have a Dixon's Q-statistic near one. Panel (b) of Appendix Figure A.6 shows that FDR is relatively constant for most methods when outliers are present, with the exception of edgeR, where FDR increases substantially with increases in Dixon's Q-statistic. Panel (c) of Appendix Figure A.6 shows that values of the Dixon's Q-statistic considered in Sim III are consistent with those observed in many data sets (the MAQC data set has fewer outliers given it is comprised of technical, not biological, replicates). Panel (d) provides an example of the types of genes identified by edgeR having high Dixon's Q-statistic. Specifically, shown are the nine genes with highest Dixon's Q-statistics in those exclusively identified by edgeR. Although FDR is well-controlled for edgeR overall (detailed in Table 2.3), these figures suggest that the majority of false discoveries that are identified by edgeR are likely in genes harboring outliers.

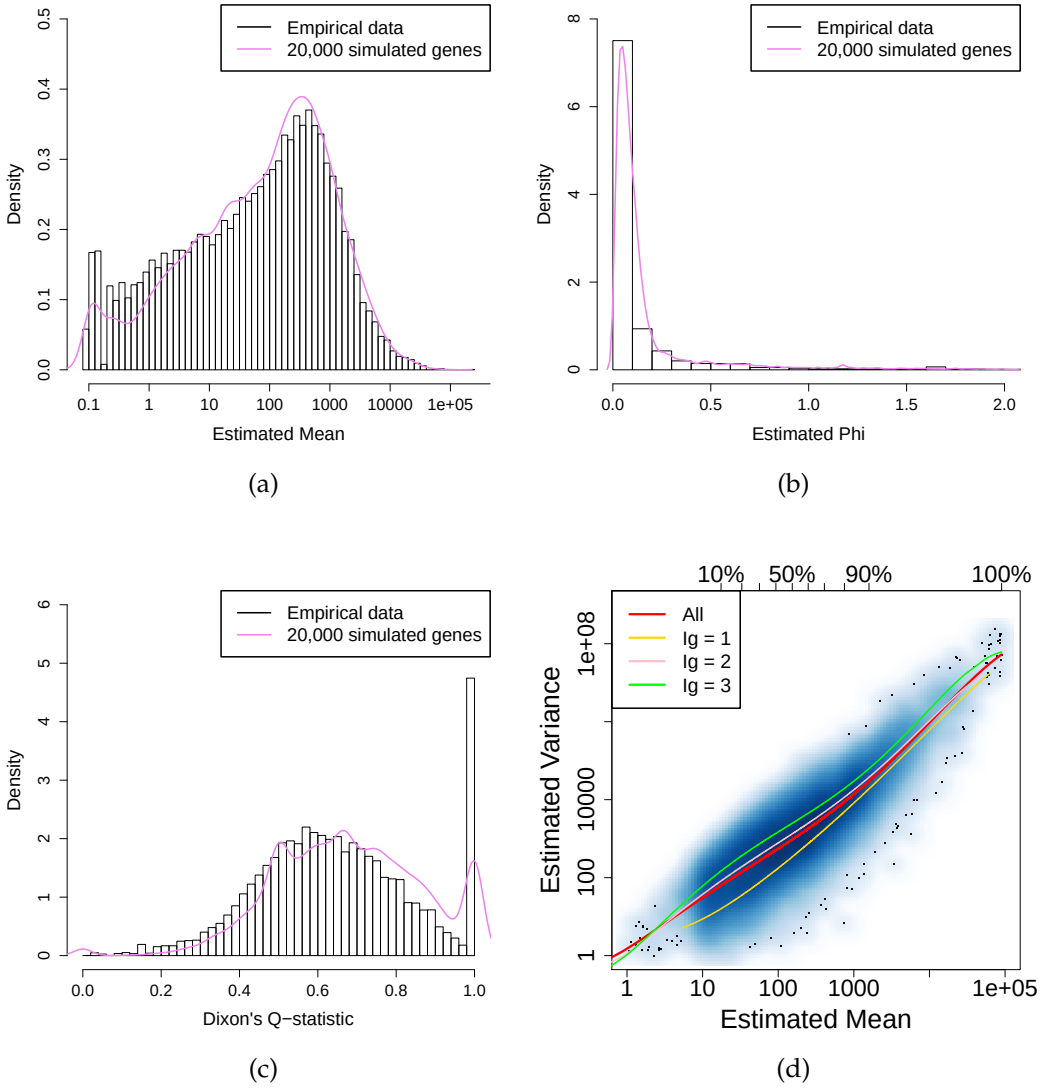


Figure A.5:

Panels (a)-(c) show the distribution of μ, ϕ and Dixon's Q-statistic, d_g, comparing one simulated data set from Sim III (histogram) with the empirical data from the experiment comparing ESCs with iPSCs (density, pink line). Panel (d) shows the scatter plot shown in Figure 2.1, but from one of the simulated data sets from Sim I.

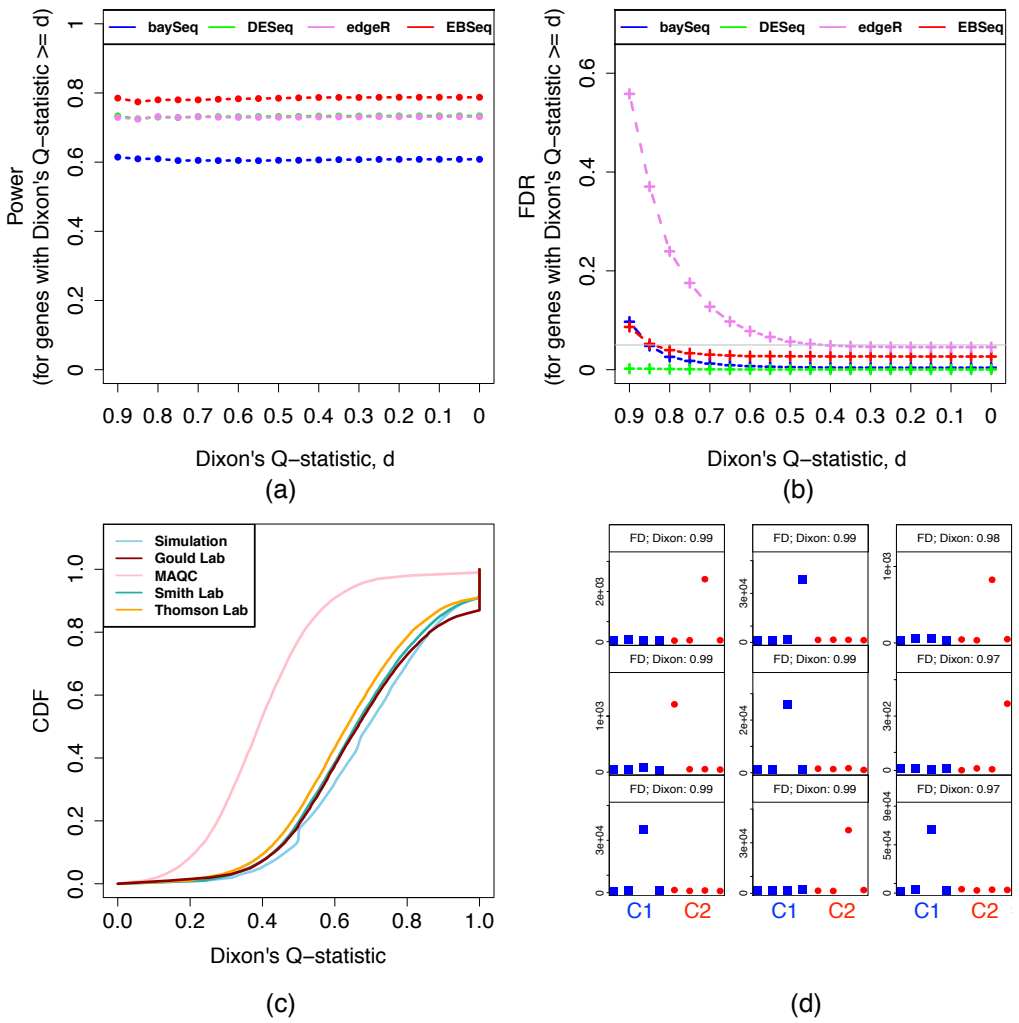


Figure A.6: Panel (a) and (b) show the operating characteristics of baySeq, DESeq, edgeR and EBSeq on subsets of genes averaged across 100 Sim III data sets for target FDR set at 5%. The subsets are defined as genes with Dixon's Q-statistic greater than the value given on the x-axis. Panel (c) shows the cumulative distribution function (CDF) of Dixon's Q-statistic in 4 empirical data sets as well as the CDF averaged across 100 simulations. Panel (d) shows 9 genes identified exclusively by edgeR having highest Dixon's Q-statistic for one simulated data set. The blue and red points correspond to two different conditions. The y-axis shows the normalized gene expression. The legend within each box shows whether the gene is a true positive (TP) or false discovery (FD) as well as the corresponding Dixon's Q-statistic value.

A.4 Additional case study results of experiment comparing ESCs vs. iPSCs

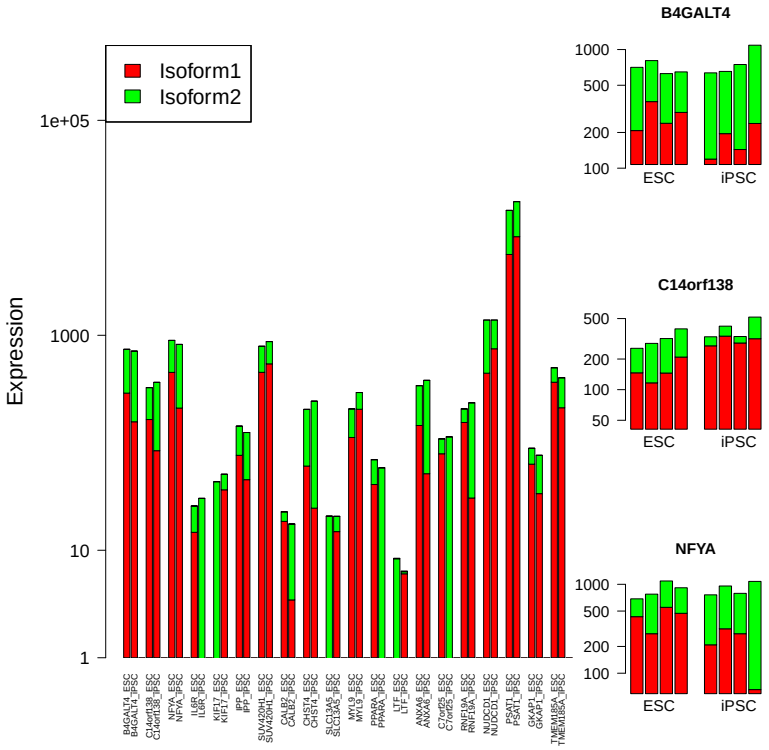
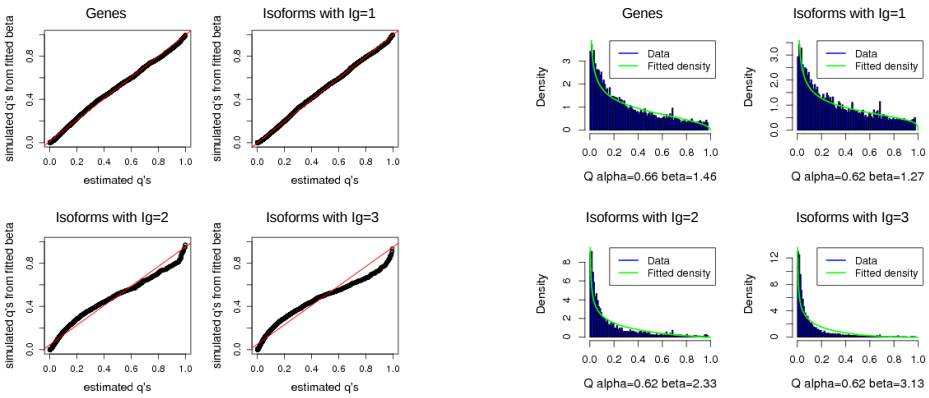


Figure A.7: The left panel shows 20 genes in the $I_g = 2$ group identified as EE by EBSeq (gene level posterior probability of EE > 0.95) with DE isoforms (isoform level posterior probability of DE > 0.95) in the experiment comparing ESCs with iPSCs. Each bar shows the isoform expression in each condition; expression of the constituent isoforms is shown in different colors within each gene. The right panel shows 3 example EE genes with DE isoforms. Each bar shows the isoform expression in each sample; expression of the constituent isoforms is shown in different colors within each gene.

A.5 Model diagnostics of EBSeq in experiment comparing ESCs vs. iPSCs

Appendix Figure A.8(a) shows the estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the same number of points simulated from the prior assumed in EBSeq, namely a Beta distribution with hyperparameters estimated as described in the main text using data from the experiment comparing ESCs with iPSCs. Appendix Figure A.8(b) shows the histogram of estimated $q_{g_i}^{C1}$'s (q_g^{C1}'s) and the fitted Beta density using that same data. These figures indicate that the prior assumed by EBSeq is reasonable for the experiments considered here.



(a) (b)

Figure A.8:

Panel (a) shows a QQ-plot comparing the estimated $q_{g_i}^{C1\prime}$s ($q_g^{C1\prime}$s) and the same number of points simulated from a Beta distribution with parameters estimated via EBSeq. The data is from the experiment comparing ESCs with iPSCs. Panel (b) shows a histogram of the estimated $q_{g_i}^{C1\prime}$s ($q_g^{C1\prime}$s) and the corresponding Beta densities.

A.6 EBSeq extension to accommodate dependence among isoforms from the same gene

In EBSeq model, we assume:

$X_{gi,s}|r_{gi,s}, q^C_{gi} \sim NB(r_{gi,s}, q^C_{gi})$

$q^C_{gi}|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$

The Negative Binomial distribution may also be viewed as a Gamma-Poisson mixture:

$X_{gi,s}|\lambda_{gi,s} \sim Poi(\lambda_{gi,s})$

$\lambda_{gi,s}|r_{gi,s}, q^C_{gi} \sim Gamma(r_{gi,s}, \frac{q^C_{gi}}{1-q^C_{gi}})$

$$P(X_{gi,s}|r_{gi,s}, q^C_{gi}) = \int P(X_{gi,s}|\lambda_{gi,s}) P(\lambda_{gi,s}|r_{gi,s}, q^C_{gi}) d\lambda_{gi,s} \tag{A.1}$$

To accommodate covariate structure among isoforms within gene g, we may assume the vector of isoform expressions $(X_{g1,s}, X_{g2,s}, ..., X_{gI,s})^T$ follows multivariate Poisson distribution with mean $(\lambda_{g1,s}, \lambda_{g2,s}, ..., \lambda_{gI,s})^T$ and variance-covariance matrix $\Sigma_{g,s}$. Then the prior predictive probability of a sample s can be written as:

$$\begin{aligned} &P\left(\begin{pmatrix} X_{g1,s} \\ X_{g2,s} \\ ... \\ X_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s} \\ r_{g2,s} \\ ... \\ r_{gI,s} \end{pmatrix}, \begin{pmatrix} q^C_{g1} \\ q^C_{g2} \\ ... \\ q^C_{gI} \end{pmatrix}\right) \\ &= \int ... \int P\left(\begin{pmatrix} X_{g1,s} \\ X_{g2,s} \\ ... \\ X_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix}, \Sigma_{g,s}\right) \prod_{i=1,...,I} P(\lambda_{gi,s}|r_{gi,s}, q^C_{gi}) d\lambda_{g1,s} ... d\lambda_{gI,s} \end{aligned} \tag{A.2}$$

The prior predictive distribution of the EE case $(q^{C1}_{g1}, q^{C1}_{g2}, ..., q^{C1}_{gI})^T = (q^{C2}_{g1}, q^{C2}_{g2}, ..., q^{C2}_{gI})^T$ can be written as:

$$g_0\begin{pmatrix} X_{g1}^{C1,C2} \\ X_{g2}^{C1,C2} \\ ... \\ X_{gI}^{C1,C2} \end{pmatrix} = \int \left[\prod_{s\in C1,C2} \int ... \int P\left(\begin{pmatrix} x_{g1,s} \\ x_{g2,s} \\ ... \\ x_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix}, \Sigma_{g,s} \right) \prod_i P(\lambda_{gi,s}|r_{gi,s}, q_{gi}) d\lambda_{g1,s} ... d\lambda_{gI,s} \right] P(q_{gi}|\alpha, \beta^{I_g}) dq_{gi} \tag{A.3}$$

The prior predictive distribution of the DE case $(q_{g1}^{C1}, q_{g2}^{C1}, ..., q_{gI}^{C1})^T \neq (q_{g1}^{C2}, q_{g2}^{C2}, ..., q_{gI}^{C2})^T$ can be written as:

$$g_1\begin{pmatrix} X_{g1}^{C1,C2} \\ X_{g2}^{C1,C2} \\ ... \\ X_{gI}^{C1,C2} \end{pmatrix} = g_0\begin{pmatrix} X_{g1}^{C1} \\ X_{g2}^{C1} \\ ... \\ X_{gI}^{C1} \end{pmatrix} g_0\begin{pmatrix} X_{g1}^{C2} \\ X_{g2}^{C2} \\ ... \\ X_{gI}^{C2} \end{pmatrix} \tag{A.4}$$

From the property of multivariate Poisson distribution, the marginal distribution of multivariate Poisson distribution is still a Poisson distribution (Kawamura et al., 1979). For any isoform j:

$$\begin{aligned} P(X_{gj,s}|\lambda_{gj,s}) &= \sum_{\substack{x_{gi,s}=0,...,\infty \\ i\neq j}} P\left(\begin{pmatrix} x_{g1,s} \\ ... \\ X_{gj,s} \\ ... \\ x_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ ... \\ \lambda_{gj,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix}, \Sigma_{g,s} \right) \\ &\sim Poi(\lambda_{gj,s}) \end{aligned} \tag{A.5}$$

So the marginal prior predictive distribution of isoform j in the case of $q_{gj}^{C1} = q_{gj}^{C2}$ can be written as:

$$
\begin{aligned}
g_0(X_{gj}^{C1,C2}) &= \sum_{\substack{x_{gi,s}=0,\dots,\infty \\ i\neq j \\ s\in C1,C2}} g_0 \begin{pmatrix} x_{g1}^{C1,C2} \\ \dots \\ X_{gj}^{C1,C2} \\ \dots \\ x_{gI}^{C1,C2} \end{pmatrix} \\
&= \int \left[\prod_{s\in C1,C2} \int \dots \int \sum_{\substack{x_{gi,s}=0,\dots,\infty \\ i\neq j}} P\left(\begin{pmatrix} x_{g1,s} \\ \dots \\ X_{gj,s} \\ \dots \\ x_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ \dots \\ \lambda_{gj,s} \\ \dots \\ \lambda_{gI,s} \end{pmatrix}, \Sigma_{g,s} \right) \prod_i P(\lambda_{gi,s}|r_{gi,s}, q_{gi}) d\lambda_{g1,s} \dots d\lambda_{gI,s} \right] * \\
&P(q_{gi}|\alpha, \beta^{I_g}) dq_{gi} \qquad \text{(A.6)} \\
&= \int \left[\prod_{s\in C1,C2} \int \dots \int P(X_{gj,s}|\lambda_{gj,s}) \prod_i P(\lambda_{gi,s}|r_{gi,s}, q_{gi}) d\lambda_{g1,s} \dots d\lambda_{gI,s} \right] * P(q_{gi}|\alpha, \beta^{I_g}) dq_{gi} \\
&= \int \left[\prod_{s\in C1,C2} \int P(X_{gj,s}|\lambda_{gj,s}) P(\lambda_{gj,s}|r_{gj,s}, q_{gj}) d\lambda_{gj,s} \right] \left[\int \dots \int_{i\neq j} \prod_i P(\lambda_{gi,s}|r_{gi,s}, q_{gi}) d\lambda_{g1,s} \dots d\lambda_{gI,s} \right] * \\
&P(q_{gi}|\alpha, \beta^{I_g}) dq_{gi} \\
&= \int \left[\prod_{s\in C1,C2} \int P(X_{gj,s}|\lambda_{gj,s}) P(\lambda_{gj,s}|r_{gj,s}, q_{gj}) d\lambda_{gj,s} \right] P(q_{gi}|\alpha, \beta^{I_g}) dq_{gi} \\
&= f_0(X_{gj}^{C1,C2})
\end{aligned}
$$

This is agreed with the prior predictive distribution defined in Chapter 2. In the case of $q_{gj}^{C1} \neq q_{gj}^{C2}$, the prior predictive probability can be written as:

$$g_1(X_{gj}^{C1,C2}) = g_0(X_{gj}^{C1}) g_0(X_{gj}^{C2}) = f_0(X_{gj}^{C1}) f_0(X_{gj}^{C2}) = f_1(X_{gj}^{C1,C2}) \qquad \text{(A.7)}$$

Assume

$$\Sigma_{g,s} = \begin{pmatrix} \lambda_{g1,s} & \lambda_{g12} & \dots & \lambda_{g1I} \\ \lambda_{g12} & \lambda_{g2,s} & \dots & \lambda_{g2I} \\ \dots & \dots & \dots & \dots \\ \lambda_{1I} & \lambda_{g2I} & \dots & \lambda_{gI,s} \end{pmatrix}$$

In which diagonal elements are sample specific and the off-diagonal elements are shared by all samples.

To estimate the variance-covariance matrix $\Sigma_{g,s}$, consider law of total variance:

$$Var\left[\begin{pmatrix} X_{g1,s} \\ X_{g2,s} \\ ... \\ X_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] = Var\left[E\left[\begin{pmatrix} X_{g1,s} \\ X_{g2,s} \\ ... \\ X_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix}\right] | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right]$$
$$+E\left[Var\left[\begin{pmatrix} X_{g1,s} \\ X_{g2,s} \\ ... \\ X_{gI,s} \end{pmatrix} | \begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix}\right] | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] \tag{A.8}$$

From our definition, equation A.8 equals to

$$Var\left[\begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] + E\left[\begin{pmatrix} \lambda_{g1,s} & \lambda_{g12} & ... & \lambda_{g1I} \\ \lambda_{g12} & \lambda_{g2,s} & ... & \lambda_{g2I} \\ ... & ... & ... & ... \\ \lambda_{1I} & \lambda_{g2I} & ... & \lambda_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] \tag{A.9}$$

Recall that we assume $\lambda_{gi,s}|r_{gi,s}, q^C_{gi} \sim \text{Gamma}(r_{gi,s}, \frac{q^C_{gi}}{1-q^C_{gi}})$ independently for each isoform, then:

$$Var\left[\begin{pmatrix} \lambda_{g1,s} \\ \lambda_{g2,s} \\ ... \\ \lambda_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] = \begin{pmatrix} \frac{r_{g1,s}(1-q^C_{g1})^2}{(q^C_{g1})^2} & 0 & ... & 0 \\ 0 & \frac{r_{g2,s}(1-q^C_{g2})^2}{(q^C_{g2})^2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{r_{gI,s}(1-q^C_{gI})^2}{(q^C_{gI})^2} \end{pmatrix} \tag{A.10}$$

$$E\left[\begin{pmatrix} \lambda_{g1,s} & \lambda_{g12} & ... & \lambda_{g1I} \\ \lambda_{g12} & \lambda_{g2,s} & ... & \lambda_{g2I} \\ ... & ... & ... & ... \\ \lambda_{1I} & \lambda_{g2I} & ... & \lambda_{gI,s} \end{pmatrix} | \begin{pmatrix} r_{g1,s}, q^C_{g1} \\ r_{g2,s}, q^C_{g2} \\ ... \\ r_{gI,s}, q^C_{gI} \end{pmatrix}\right] = \begin{pmatrix} \frac{r_{g1,s}(1-q^C_{g1})}{q^C_{g1}} & \lambda_{g12} & ... & \lambda_{g1I} \\ \lambda_{g12} & \frac{r_{g2,s}(1-q^C_{g2})}{q^C_{g2}} & ... & \lambda_{g2I} \\ ... & ... & ... & ... \\ \lambda_{1I} & \lambda_{g2I} & ... & \frac{r_{gI,s}(1-q^C_{gI})}{q^C_{gI}} \end{pmatrix} \tag{A.11}$$

Then the equation A.8 can be written as:

$$\begin{pmatrix} \frac{r_{g1,s}(1-q^C_{g1})}{q^C_{g1}} & \lambda_{g12} & ... & \lambda_{g1I} \\ \lambda_{g12} & \frac{r_{g2,s}(1-q^C_{g2})}{q^C_{g2}} & ... & \lambda_{g2I} \\ ... & ... & ... & ... \\ \lambda_{1I} & \lambda_{g2I} & ... & \frac{r_{gI,s}(1-q^C_{gI})}{q^C_{gI}} \end{pmatrix} + \begin{pmatrix} \frac{r_{g1,s}(1-q^C_{g1})^2}{(q^C_{g1})^2} & 0 & ... & 0 \\ 0 & \frac{r_{g2,s}(1-q^C_{g2})^2}{(q^C_{g2})^2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{r_{gI,s}(1-q^C_{gI})^2}{(q^C_{gI})^2} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{r_{g1,s}(1-q^C_{g1})}{(q^C_{g1})^2} & \lambda_{g12} & ... & \lambda_{g1I} \\ \lambda_{g12} & \frac{r_{g2,s}(1-q^C_{g2})}{(q^C_{g2})^2} & ... & \lambda_{g2I} \\ ... & ... & ... & ... \\ \lambda_{1I} & \lambda_{g2I} & ... & \frac{r_{gI,s}(1-q^C_{gI})}{(q^C_{gI})^2} \end{pmatrix} \tag{A.12}$$

Then the $r_{gi,s}, q^C_{gi}$'s may be estimated by the empirical mean and variance of the expression of each individual isoform. The λ_{gij}'s may be estimated from the empirical covariance matrix of $(X_{g1,s}, X_{g2,s}, ..., X_{gI,s})^T$.

B APPENDIX OF "EBSEQ-HMM: A BAYESIAN APPROACH FOR IDENTIFYING GENE-EXPRESSION CHANGES IN ORDERED RNA-SEQ EXPERIMENTS"

B.1 Parameter estimation

Of primary interest is the posterior probability of being in a certain pattern s, which can be written as

$$PP(S_g = s|X_g) = \frac{\gamma_{m2}P(X_g,S_g=s|m2)+\gamma_{m1}P(X_g,S_g=s|m1)}{\sum_k[\gamma_{m2}P(X_g,S_g=k|m2)+\gamma_{m1}P(X_g,S_g=k|m1)]}$$
$$=\frac{\prod_{t=2}^{T} P(X_{g,t}|X_{g,t-1},S_g^{\Delta_t}=s^{\Delta_t})*P(X_{g,1})*[\prod_{t=3}^{T} P(S_g^{\Delta_t}=s^{\Delta_t}|S_g^{\Delta_{t-1}}=s^{\Delta_{t-1}},m1)P(S_g^{\Delta_2}=s^{\Delta_2}|m1)+\prod_{t=2}^{T} P(S_g^{\Delta_t}=s^{\Delta_t}|m2)]}{\sum_k[\prod_{t=2}^{T} P(X_{g,t}|X_{g,t-1},S_g^{\Delta_t}=k^{\Delta_t})*P(X_{g,1})*[\prod_{t=3}^{T} P(S_g^{\Delta_t}=k^{\Delta_t}|S_g^{\Delta_{t-1}}=k^{\Delta_{t-1}},m1)P(S_g^{\Delta_2}=k^{\Delta_2}|m1)+\prod_{t=2}^{T} P(S_g^{\Delta_t}=k^{\Delta_t}|m2)]}$$

To estimate $PP(S_g = s|X_g)$, the Baum-Welch algorithm is used to estimate $\tilde{\pi}_j = P(S_g^{\Delta_2} = j|m_1)$ and $\tilde{a}_{dj}^{t,t+1} = P(S_g^{\Delta_{t+1}} = j|S_g^{\Delta_t} = d, m1)$ in Markov chain $m1$.

Denote estimated parameters in the current iteration by $\tilde{\pi}, \tilde{A}$. Then, in the next iteration, parameters are updated:

Forward Procedure:

$\alpha_{g,d}(2) = \tilde{\pi}_d b_d(X_{g,2})$. For $t > 2$, $\alpha_{g,j}(t+1) = [\sum_d \alpha_{g,d}(t)\tilde{a}_{dj}^{t,t+1}]b_j(X_{g,t+1})$

Backward Procedure:

$\beta_{g,d}(T) = 1$. For $t < T$, $\beta_{g,j}(t-1) = \sum_d [\beta_{g,d}(t)b_d(X_{g,t})\tilde{a}_{jd}^{t-1,t}]$

Then we have:

$P(X_g, S_g^{\Delta_t} = j|m1) \propto \alpha_{g,j}(t)\beta_{g,j}(t)$;

so $P(S_g^{\Delta_2} = J, M_g = m1|X_g) = \frac{\alpha_{g,J}(2)\beta_{g,J}(2)z_g^{m1}}{\sum_j \alpha_{g,j}(2)\beta_{g,j}(2)z_g^{m1}}$

$P(X_g, S_g^{\Delta_t} = d, S_g^{\Delta_{t+1}} = j|m1) \propto \alpha_{g,d}(t)\tilde{a}_{dj}^{t,t+1}b_j(X_{g,t+1})\beta_{g,j}(t+1)$;

so $P(S_g^{\Delta_t} = D, S_g^{\Delta_{t+1}} = J, M_g = m1|X_g) = \frac{\alpha_{g,D}(t)\tilde{a}_{D,J}^{t,t+1}b_{g,J}(X_{g,t+1})\beta_{g,J}(t+1))z_g^{m1}}{\sum_j \alpha_{g,D}(t)\tilde{a}_{D,j}^{t,t+1}b_{g,j}(X_{g,t+1})\beta_{g,j}(t+1))z_g^{m1}}$

The updated parameter estimates can then be obtained as defined in Chapter 3.

B.2 Comparison of features in simulated data vs. case study data

Appendix Figure B.1 shows comparison of features in mouse limb data and in data sets from two simulation studies. It indicates that the Mean-variance relationship and FC distribution in our simulated data set are consistent with those observed in case study data set.

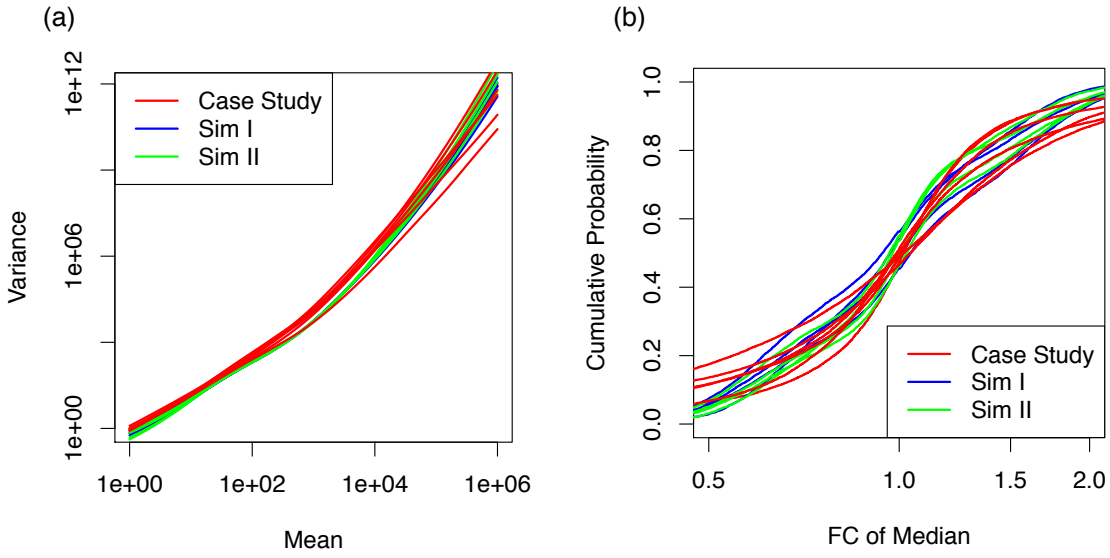


Figure B.1: (a) Shown are fitted mean-variance relationship within each condition. Each line shows relationship learnt from one position (time point). One random selected data set is used here for each simulation study. Therefore 7 lines are shown for mouse limb data and 5 lines are shown for each simulation study. (b) Shown are CDF of FCs comparing adjacent positions (time points). Each line shows distribution learnt from an adjacent pair of positions (time points). One random selected data set is used here for each simulation study. Therefore 6 lines are shown for mouse limb data and 4 lines are shown for each simulation study.

B.3 Evaluation of path classification using Sim II data sets

Appendix Figure B.2 shows evaluation of each method in clustering DE genes into expression path clusters. Results are based on 100 Sim II data sets.

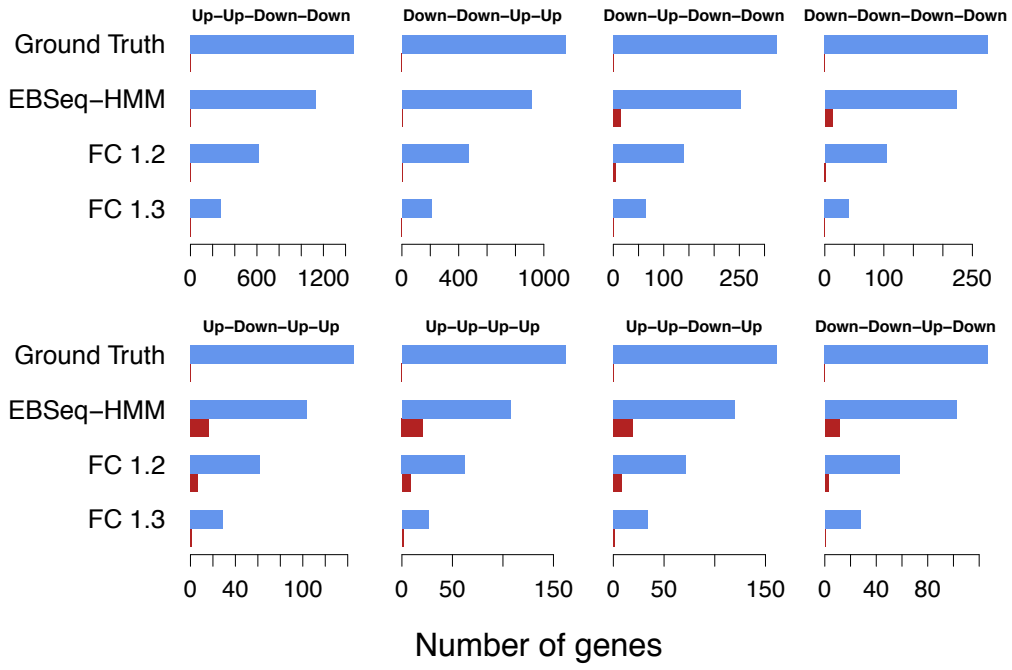


Figure B.2: Shown are the number of genes (ground truth) simulated in Sim II as being in each of 8 dynamic paths (these 8 are shown as they contain the most genes among all simulated paths). Also shown are the average number classified into each path by EBSeq-HMM and by FC analysis at thresholds 1.2 and 1.3 (averages are calculated over 100 Sim II datasets). Correct classifications are shown in blue; incorrect are shown in red.

B.4 Genes exclusively identified by EBSeq-HMM on mouse limb data

Appendix Figure B.3 shows 12 genes exclusively identified by EBSeq-HMM. EBSeq-HMM was able to identify genes showing relatively weak, but consistent, changes over ordered conditions.

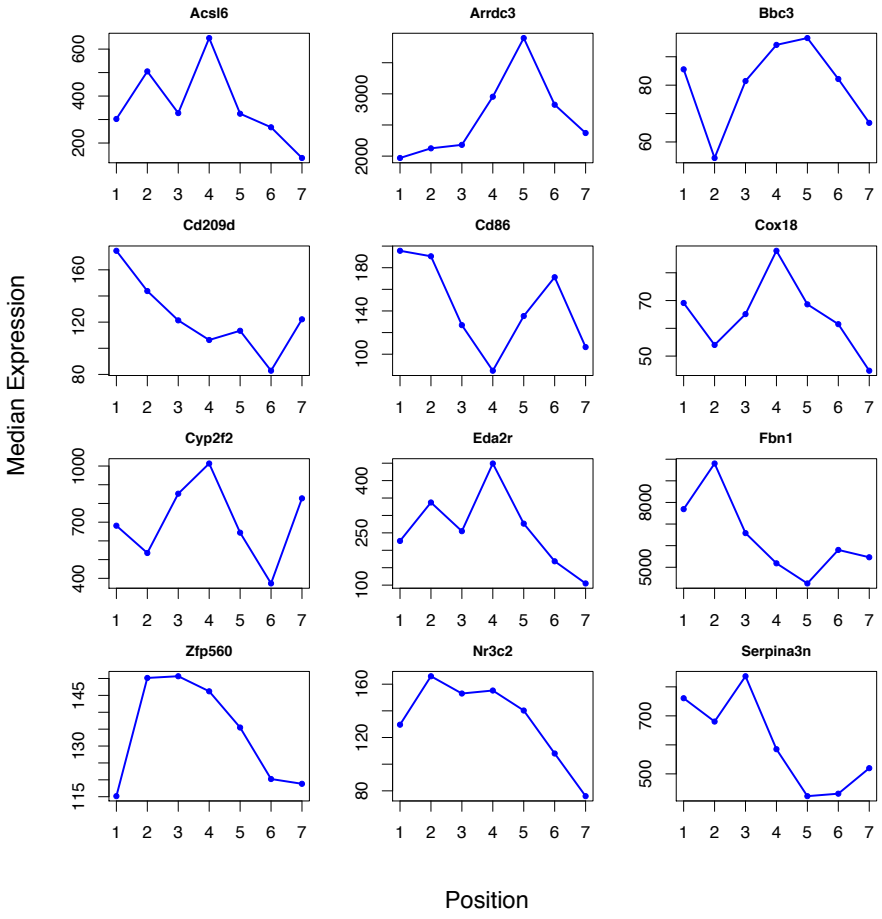


Figure B.3: Shown are genes identified by EBSeq-HMM but neither DESeq2 nor edgeR using case study data. The x-axis shows positions on mouse limb and the y-axis shows median gene expression adjusted for library sizes.

B.5 Model diagnostics of EBSeq-HMM on mouse limb data

Appendix Figure B.4 shows model diagnostic plots applying EBSeq-HMM on mouse limb case study data. Appendix Figure B.4 (a) shows the histogram of empirical q's estimated within each condition and the fitted Beta density using that same data. Appendix Figure B.4 (b) shows estimated q's and the same number of points simulated from the Beta prior. These figures indicate that the model assumed by EBSeq-HMM is reasonable for the experiments considered here.

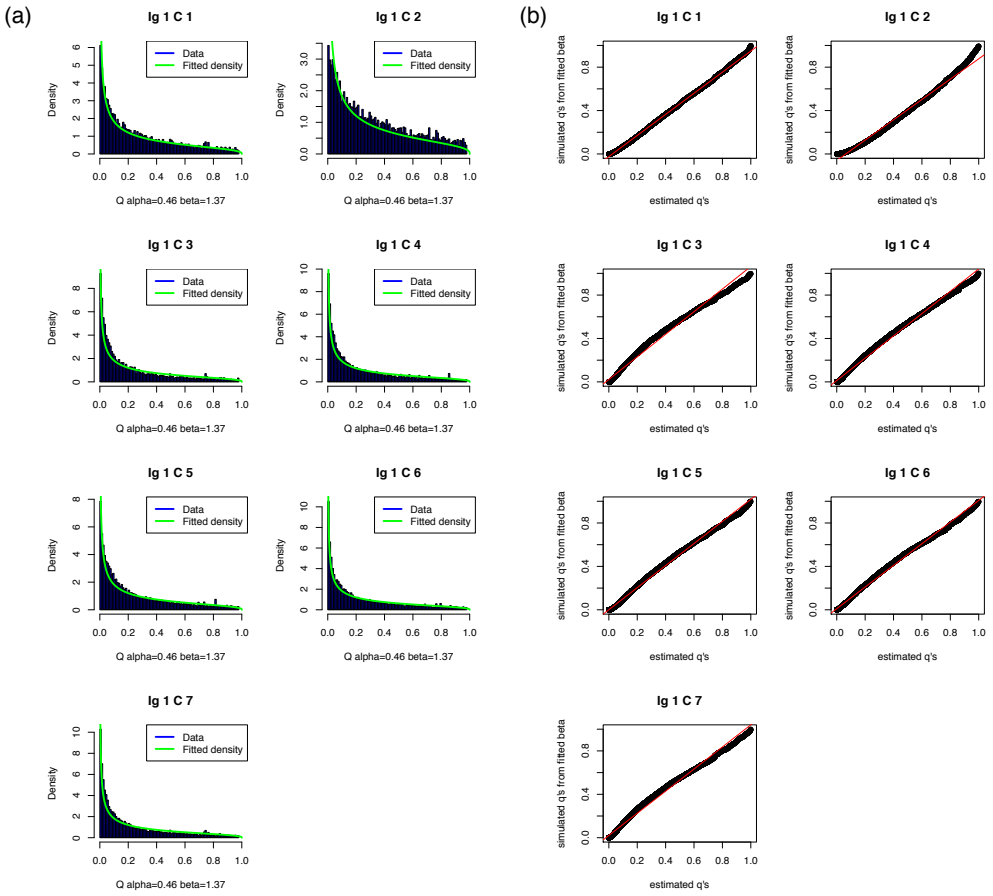


Figure B.4: (a) Shown are the histogram of empirical q's estimated within each condition and the fitted Beta density using that same data. (b) Shown are estimated q's and the same number of points simulated from the Beta prior.

C APPENDIX OF "OSCOPE: A STATISTICAL PIPELINE FOR IDENTIFYING OSCIALLATORY GENES USING UNSYNCHRONIZED SINGLE-CELL RNA-SEQ DATA"

C.1 Ordering effect in other data sets

We further examine the ordering effect in two other in-house data sets and two public available data sets generated by Fluidigm C1. Two in-house data sets include the H1-FUCCI data set and one data set provided by Fluidigm Corporation for demo purpose. The Fluidigm demo data set contains 72 H1 cells. Cells were sequenced via Illumina HiSeq 2000. Reads were mapped to human RefSeq reference via bowtie and quantified via RSEM. Trapnell et al. data set and Wu et al. data set were downloaded from supplements of previous papers (Trapnell et al., 2014; Wu et al., 2014). FPKM's of genes were provided in each of the papers.

Appendix Figure C.1 shows the same four example genes as in Figure 4.10 (e). These four genes have ordering effect consistently in all data sets.

Appendix Figure C.2 shows sample IDs and their corresponding physical locations on a Fluidigm C1 chip. Recall the cells with artificially high expression are likely to be those with small/large inlet ID (red). These cells come from top/bottom part of the capture sites.

C.2 Normalization and rescaling

In single-cell RNA-seq data, cells are normalized by their library sizes. Library sizes are obtained by median normalization in Anders and Huber (2010). Before applying paired-sine model, the normalized microarrays or RNA-seq data were further rescaled to values between [-1, 1]. The rescaling was applied within each gene. To minimize noises from outliers, values $>$ 95th ($<$ 5th) quantile of expressions are imputed as the 95th (5th) quantile. Then linear rescaling was applied within each gene.

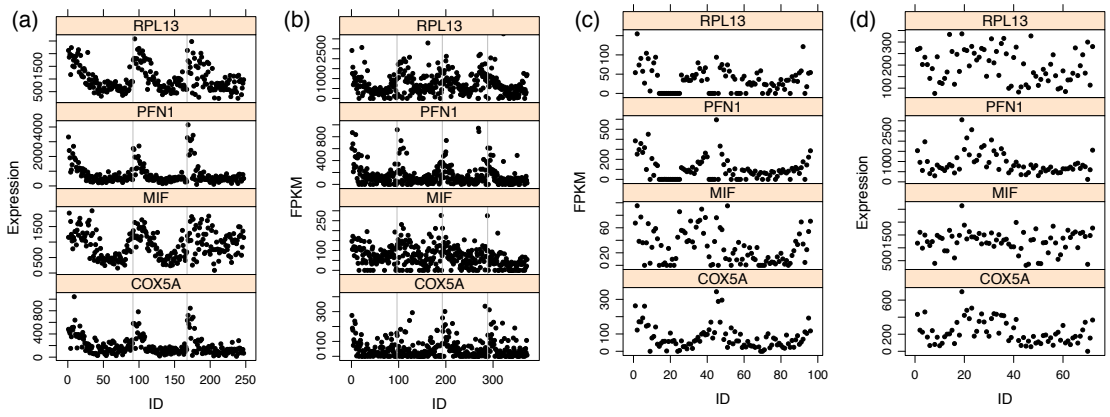


Figure C.1: Shown are 4 genes with ordering effect on 4 different data sets. The x axis is sample ID's and the y axis is normalized expression or FPKM. (a) H1-FUCCI data, three experiments are separated by gray lines (b) Data set from Trapnell et al. (2014). On the x axis, cells are shown with the original sample order in the supplementary data set of the original paper. Four experiments were defined in the supplementary data set as well. Here four experiments are separated by gray lines. The y axis shows the FPKM obtained from the supplementary data set. (c) Data set from Wu et al. (2014). On the x axis, cells are shown with the original sample order in the supplementary data set of the original paper. The y axis shows the FPKM obtained from the supplementary data set. (d) Data set provided by Fluidigm Corporation for demo purpose. On the x axis, cells are shown with the original sample order provided by Fluidigm Corporation. The y axis shows the expressions after adjusting for library sizes.

C.3 Case study results on Whitfield data

Microarrays gene expression data from human cancer cell (HeLa S3) was downloaded from `http://genome-www.stanford.edu/Human-CellCycle/HeLa/`. Total 5 experiments were conducted in Whitfield et al. (2002). We used experiment 3 which has largest sample size to evaluate Oscope. Double thymidine block was used to synchronize the cell population and 48 time points were measured. Total 9559 genes are included in this data set and we applied Oscope on all genes. After applying paired-sine model, top 10% genes were used as input for the K-Medoid

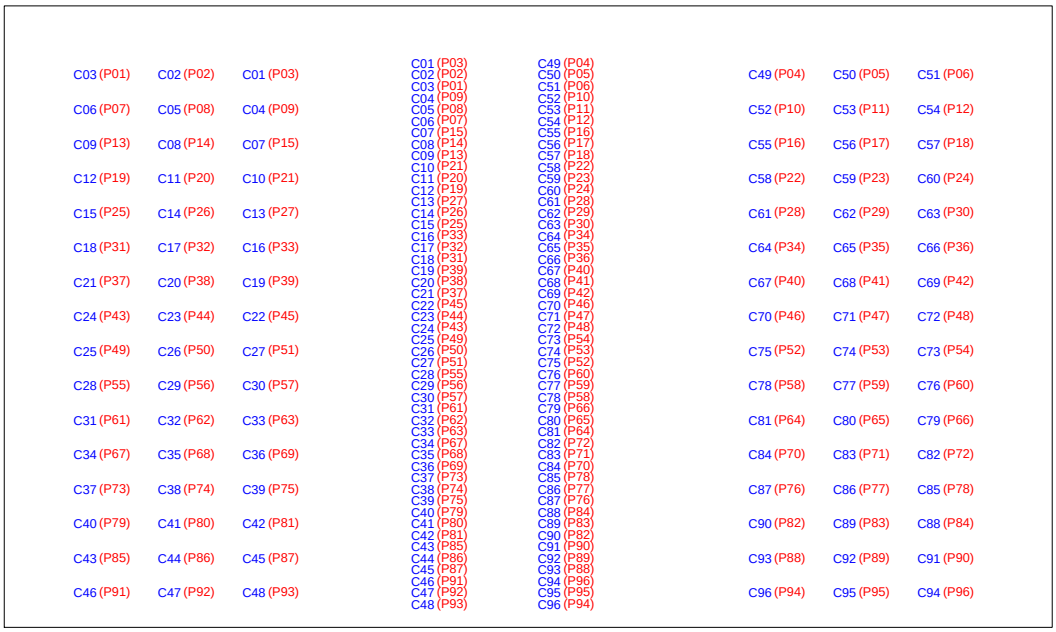


Figure C.2: Shown are sample IDs and their corresponding physical locations on a Fluidigm C1 chip. capture site IDs are shown in blue and its corresponding inlet IDs are shown in red. The inlet ID (red) is the one used in all down-stream experiments/analyses (sequencing, data analyses, etc.). All data sets generated in Thomson lab follow this order.

algorithm. ENI was applied with $m = 4$ and the degree of freedom of SPR was set as 3. 1134 genes are called as periodic genes by the auto-regression model in Whitfield et al. (2002). We used these 1134 genes as ground truth in our evaluation.

Appendix Figure C.3 shows 69 genes plotted using recovered order (a) or original order (b). Panel (b) shows that genes identified by Oscope have cyclic patterns in original time course experiments. Panel (a) shows that Oscope recovered smooth base cycle profiles for these genes. Comparing these panels, Oscope recovered order is able to reveal phase shifts among different genes.

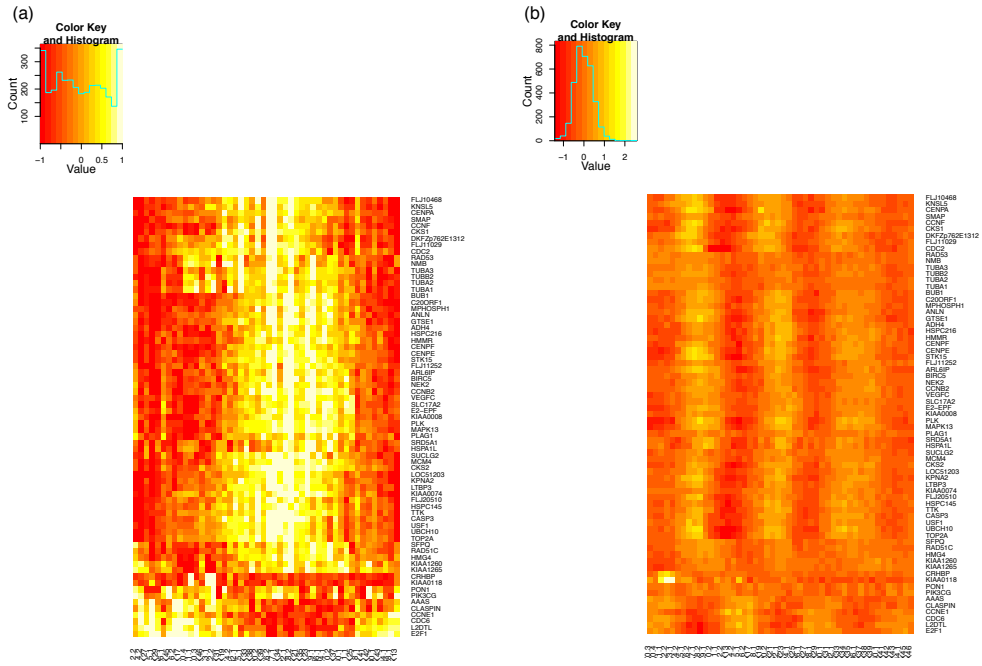


Figure C.3: (a) Shown are 69 genes identified by Oscope on Whitfield data. On the x axis, samples are shown following the order recovered by ENI using 69 genes identified by paired-sine model. (b) Shown are the same set of genes as in (a). Here samples are shown following the original order over time.

C.4 Case study results on H1 ESCs data

To reduce computational intensity and reduce affects from noise, we applied Oscope on 2993 genes with high mean and high variance. We first eliminated genes with mean expected counts less than 100. To define the high variance gene, we fitted a linear model on $\log(\text{variance}) \sim \log(\text{mean}) + c$. Genes with variance above this fitted line were defined as high variance genes. After applying paired-sine model, top 10% genes were used as input for the K-Medoid algorithm. ENI was applied with $m = 4$ and the degree of freedom of SPR was set as 3.

Appendix Figure C.4 shows mean vs. variance on H1 ESCs data. Genes with high mean and high variance (marked as green) were considered when applying Oscope pipeline.

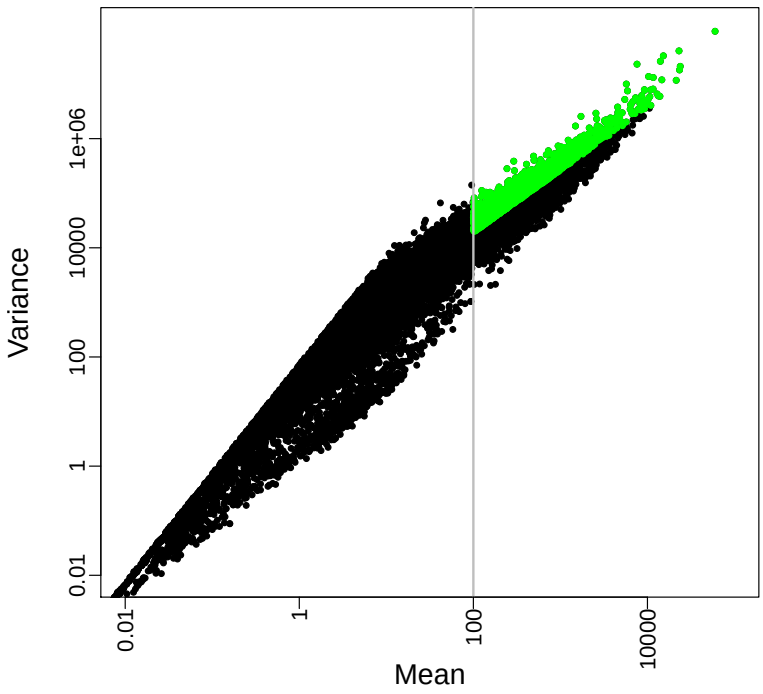


Figure C.4: Shown are mean vs. variance plot of H1 single-cell RNA-seq data. High mean genes are defined as the ones with normalized mean greater than 100. The gray line shows the fitted line of $\log(\text{var}) \sim \log(\text{mean}) + c$. High variance genes are defined as the ones with estimated variance above this line. Genes with high mean and high variance are marked as green. They are further used in downstream analyses.

Appendix Figure C.5 shows 32 CC genes identified by Oscope on ESCs data. Genes are plotted using recovered order on (a) H1 data set or (b) H1-FUCCI data set. Panel (b) shows the recovered order by Oscope successfully separated three CC phases on H1-FUCCI data set. Oscope recovered smooth base cycle profiles for these genes on both H1 data and H1-FUCCI data. And the recovered base cycle profiles are very consistent comparing two data sets.

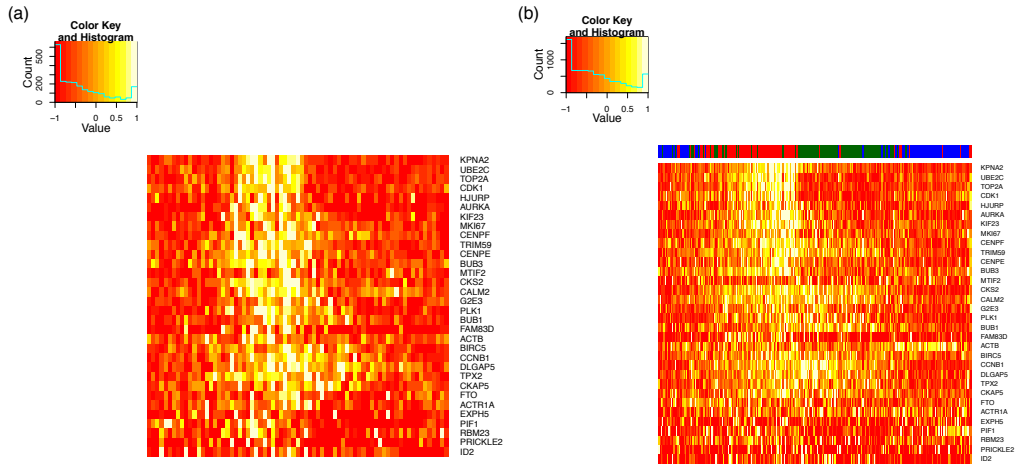


Figure C.5: (a) Shown are 32 genes identified by Oscope on H1 data. On the x axis, samples are shown following the order recovered by ENI using 32 genes identified by paired-sine model. (b) Shown are the same set of genes as in (a). Here samples are shown following the ENI recovered order on FUCCI data. Cells from S, G2 or G1 phase are marked as blue, red and green on the top sidebar of the heatmap.

REFERENCES

Ailliot, Pierre, and Valérie Monbet. 2012. Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software* 30:92–101.

Anders, S. 2012. *Htseq: Analysing high-throughput sequencing data with python*.

Anders, S, and W Huber. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11:R106.

Berger, James O. 1985. *Statistical decision theory and bayesian analysis*. Springer.

Bock, C, E Kiskinis, G Verstappen, H Gu, G Boulting, Z D Smith, M Ziller, G F Croft, M W Amoroso, D H Oakley, A Gnirke, K Eggan, and A Meissner. 2011. Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell.* 144(3):439–52.

Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. 2013. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*.

Bullard, J H, E A Purdom, K D Hansen, and S Dudoit. 2010. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics* 11:94.

Chang, H.Y. 2009. Anatomic demarcation of cells: genes to patterns. *Science* 326(5957):1206–1207.

Consortium, MAQC. 2006. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* 24:1151–1161.

Delhomme, N, and I Padioleau. 2012. *easyrnaseq, an overview*.

Derrien, Thomas, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. 2012. Fast computation and applications of genome mappability. *PloS one* 7(1):e30377.

Dixon, W J. 1950. Analysis of extreme values. *The Annals of Mathematical Statistics* 21:488.

Glaus, P., A. Honkela, and M. Rattray. 2012. Identifying differentially expressed transcripts from rna-seq data with biological variation. *Bioinformatics* 28(13):1721–1728.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, T Galaxy Team, et al. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11(8):R86.

Haag, J D, L A Shepel, B D Kolman, D M Monson, M E Benton, K T Watts, J L Waller, C C Lopez-Guajardo, D J Samuelson, and M N Gould. 2003. Congenic rats reveal three independent copenhagen alleles within the mcs1 quantitative trait locus that confer resistance to mammary cancer. *Cancer Res* 63:5808.

Hamilton, James D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society* 357–384.

Hardcastle, T J, and K A Kelly. 2010. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11: 422.

Jiang, H, and W H Wing. 2008. Seqmap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24(20):2395–2396.

———. 2009. Statistical inferences for isoform expression in rna-seq. *Bioinformatics* 25(8):1026–1032.

Katz, Y., E.T. Wang, E.M. Airoldi, and C.B. Burge. 2010. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nature Methods* 7: 1009–1015.

Kawamura, Kazutomo, et al. 1979. The structure of multivariate poisson distribution. *Kodai Mathematical Journal* 2(3):337–345.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. Bayesian approach to single-cell differential expression analysis. *Nature Methods*.

Koehler, Ryan, Hadar Issac, Nicole Cloonan, and Sean M Grimmond. 2011. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics* 27(2): 272–274.

Langmead, B, C Trapnell, M Pop, and S L Salzberg. 2010. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology* R25.

Leng, Ning, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. 2013. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics* 29(8):1035–1043.

Li, B, and C N Dewey. 2011. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Li, B, V Ruotti, R M Stewart, J A Thomson, and C N Dewey. 2010. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26(4):493–500.

Lionnet, Timothée, Kevin Czaplinski, Xavier Darzacq, Yaron Shav-Tal, Amber L Wells, Jeffrey A Chao, Hye Yoon Park, Valeria de Turris, Melissa Lopez-Jones, and Robert H Singer. 2011. A transgenic mouse for in vivo detection of endogenous labeled mrna. *nAture methods* 8(2):165–170.

Marcolino-Gomes, Juliana, Fabiana Aparecida Rodrigues, Renata Fuganti-Pagliarini, Claire Bendix, Thiago Jonas Nakayama, Brandon Celaya, Hugo Bruno Correa Molinari, Maria Cristina Neves de Oliveira, Frank G Harmon, and Alexandre Nepomuceno. 2014. Diurnal oscillations of soybean circadian clock and drought responsive genes. *PloS one* 9(1):e86402.

Mortazavi, A, B A Williams, K McCue, L Schaeffer, and B Wold. 2008. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods* 5(1):621–628.

Nicolae, M., S. Mangul, I. Măndoiu, and A. Zelikovsky. 2010. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms in Bioinformatics* 202–214.

Ohi, Y, H Qin, C Hong, L Blouin, J M Polo, T Guo, Z Qi, S L Downey, P D Manos, D J Rossi, J Yu, M Hebrok, K Hochedlinger, J F Costello, J S Song, and M Ramalho-Santos. 2011. Incomplete dna methylation underlies a transcriptional memory of somatic cells in human ips cells. *Nat Cell Biol.* 13(5):541–9.

Ozsolak, F, and P M Milos. 2011. Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12:87–98.

Phanstiel, H P, J Brumbaugh, C D Wenger, S Tian, M D Probasco, D J Bailey, D L Swaney, M A Tervo, J M Bolin, V Ruotti, R Stewart, J A Thomson, and J J Coon. 2011. Proteomic and phosphoproteomic comparison of human es and ips cells. *Nature Methods* 8:821–827.

Pickrell, J.K., J.C. Marioni, A.A. Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B. Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard. 2010. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature* 464(7289):768–772.

Rinn, J.L., C. Bondre, H.B. Gladstone, P.O. Brown, and H.Y. Chang. 2006. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS genetics* 2(7):e119.

Robinson, M., D. McCarthy, Y. Chen, and G.K. Smyth. 2014. *edger: differential expression analysis of digital gene expression data: UserâĽ™s guide*.

Robinson, M D, D J McCarthy, and G K Smyth. 2010. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–40.

Robinson, M D, and A Oshlack. 2010. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology* 11:R25.

Robinson, M D, and G K Smyth. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23(21):2881–2887.

———. 2008. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics* 9:321–332.

Rosenkrantz, Daniel J, Richard E Stearns, and Philip M Lewis, II. 1977. An analysis of several heuristics for the traveling salesman problem. *SIAM journal on computing* 6(3):563–581.

Sandmann, T., M.C. Vogg, S. Owlarn, M. Boutros, and K. Bartscherer. 2011. The head-regeneration transcriptome of the planarian schmidtea mediterranea. *Genome Biology* 12(8):R76.

Schlüter, Ralf, Thomas Scharrenbach, Volker Steinbiss, and Hermann Ney. 2005. Bayes risk minimization using metric loss functions. In *Interspeech*, 1449–1452.

Sengupta, S., V. Ruotti, J. Bolin, A. Elwell, A. Hernandez, J. Thomson, and R. Stewart. 2010. Highly consistent, fully representative mrna-seq libraries from ten nanograms of total rna. *Biotechniques* 49:898–904.

Shi, Yang, and Hui Jiang. 2013. rseqdiff: Detecting differential isoform expression from rna-seq data using hierarchical likelihood ratio test. *PloS one* 8(11):e79448.

Singh, D., C.F. Orellana, Y. Hu, C.D. Jones, Y. Liu, D.Y. Chiang, J. Liu, and JF. Prins. 2011. Fdm: a graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics* 27:2633–2640.

Smith, C W, J G Patton, and B Nadal-Ginard. 1989. Alternative splicing in the control of gene expression. *Annu Rev Genet.* 23:527–77.

Stamm, S, S Ben-Ari, I Rafalska, Y Tang, Z Zhang, D Toiber, T A Thanaraj, and Soreq H. 2005. Function of alternative splicing. *Gene* 344:1–20.

Trapnell, C., D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, and L. Pachter. 2012a. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature Biotechnology*.

Trapnell, C, L Pachter, and S L Salzberg. 2009. Tophat: discovering splice junctions with rna-seq. *Bioinformatics* 25(9):1105–1111.

Trapnell, C, A Roberts, L Goff, G Pertea, D Kim, D R Kelley, H Pimentel, S L Salzberg, J L Rinn, and L Pachter. 2012b. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature Protocols* 7(3): 562–578.

Trapnell, C, B A Williams, G Pertea, A Mortazavi, G Kwan, M J van Baren, S L Salzberg, B J Wold, and L Pachter. 2010. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):211–215.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*.

Wang, E T, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, S F Kingsmore, G P Schroth, and C B Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.

Wang, K.C., J.A. Helms, and H.Y. Chang. 2009a. Regeneration, repair and remembering identity: the three rs of hox gene expression. *Trends in cell biology* 19(6): 268–275.

Wang, Z, Gerstein M, and Snyder M. 2009b. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10:57–63.

Whitfield, Michael L, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* 13(6):1977–2000.

Wu, Angela R, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke, et al. 2014. Quantitative assessment of single-cell rna-sequencing methods. *Nature methods* 11(1):41–46.

Wu, J.Q., L. Habegger, P. Noisa, A. Szekely, C. Qiu, S. Hutchison, D. Raha, M. Egholm, H. Lin, S. Weissman, et al. 2010. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proceedings of the National Academy of Sciences* 107(11): 5254–5259.

Yuan, M., and C. Kendziorski. 2006. Hidden markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association* 101(476):1323–1332.

Zakany, Jozsef, and Denis Duboule. 2007. The role of hox genes during vertebrate limb development. *Current opinion in genetics & development* 17(4):359–366.